Econometric Theory

Seung Hyun Kim

July 17, 2026

1 Statistical Decision Theory **6**
1.1 The Elements of Decision Theory . . . 6
1.2 Admissibility of Decision Rules . . . 8
1.3 Completeness of Bayesian Decision Rules . . . 10
1.4 The Sufficiency Principle . . . 16

2 Hypothesis Testing and Set Estimation **31**
2.1 Hypothesis Testing . . . 31
2.1.1 Core Concepts of Hypothesis Testing . . . 32
2.1.2 Uniformly Most Powerful Tests . . . 35
2.1.3 Asymptotic Tests and Consistency . . . 43
2.2 Confidence Sets . . . 47
2.2.1 Evaluating Confidence Sets . . . 50

3 The Asymptotics of Extremum Estimators **52**
3.1 Existence of Extremum Estimators . . . 52
3.2 Uniform Convergence in Probability . . . 56
3.2.1 The Measurability Problem . . . 57
3.2.2 Characterization of Uniform Convergence in Probability . . . 59
3.2.3 Uniform Weak Laws of Large Numbers . . . 66
3.3 The Basic Consistency and Asymptotic Normality Theorems . . . 69
3.4 The Asymptotics of GMM Estimation . . . 78
3.4.1 A Motivating Example . . . 78
3.4.2 A General Framework and Regularity Assumptions . . . 79
3.4.3 Consistency of the GMM Estimator and Identification . . . 82
3.4.4 Asymptotic Distribution of the GMM Estimator . . . 85
3.4.5 Efficient GMM Estimation . . . 90
3.4.6 Inference and Weak Identification . . . 97
3.4.7 Testing for Overidentification . . . 101
3.5 The Asymptotics of Maximum Likelihood Methods . . . 106
3.5.1 The Gibbs Inequality and Motivating ML Estimation . . . 107

3.5.2 Identification, Existence and Consistency . . . 109
3.5.3 The Information Matrix and Cramer-Rao Lower Bound . . . 114
3.5.4 Asymptotic Distribution ofthe ML Estimator . . . 121
3.5.5 The Trinity: Wald, LR and LM . . . 133
3.5.6 A Special Case: I.I.D. Data . . . 144
3.6 Extremum Estimation with Convex Objective Functions . . . 167
3.6.1 Topological Properties of Sequence Spaces . . . 167
3.6.2 Measurability of Convex Maps . . . 174
3.6.3 The Basic Convergence Theorem for Convex Extremum Estimation . . . 188
3.7 Quantile Regression . . . 195
3.7.1 Quantiles and the Check Function . . . 195
3.7.2 Motivating the Quantile Regression Estimator . . . 198
3.7.3 Asymptotic Properties of the Quantile Regression Estimator . . . 200

4 The Asymptotics of Bayesian Estimation 217
4.1 Existence of the Posterior . . . 217
4.1.1 The Posterior Distribution under Domination . . . 218
4.1.2 The Posterior Distribution in the General Case . . . 221
4.2 Consistency of Bayesian Estimators and Posterior Consistency . . . 232
4.3 Berstein-von Mises Theorem . . . 235
4.3.1 Total Variation Distance . . . 235
4.3.2 Proof of the Theorem . . . 243
4.3.3 Primitive Identification Conditions . . . 262

5 Linear Regression Models 272
5.1 The Normal Univariate Linear Regression Model . . . 272
5.1.1 Model Assumptions . . . 272
5.1.2 MLEs of Model Parameters . . . 274
5.1.3 Matrices, Residual Makers and FWL . . . 274
5.1.4 Finite Sample Distributions of MLEs . . . 277
5.1.5 Testing Linear Hypotheses for β . . . 279
5.1.6 Equivalent Test Statistics for Zero Restrictions . . . 280
5.1.7 Testing for Single Zero Restrictions . . . 282
5.2 The Generalized Univariate Linear Regression Model . . . 284
5.2.1 Model Assumptions . . . 284
5.2.2 Asymptotic Properties of the OLS Estimator . . . 286
5.2.3 Heteroskedasticity Consistent Covariance Matrix Estimation . . . 288
5.2.4 Testing for General Non-linear Hypotheses for β . . . 292
5.3 A Bayesian Approach to Univariate Normal Linear Regression . . . 293
5.3.1 Model Assumptions . . . 293
5.3.2 Full Conditional Distributions of β and σ^2 . . . 294
5.3.3 Sampling from the Posterior Distribution . . . 296

5.3.4 The Posterior Density for a Special Case . . . 297

6 Endogeneity and GMM **300**

6.1 An Illustrative Example . . . 301
6.2 GMM Estimation of β . . . 302
6.2.1 Model Assumptions . . . 302
6.2.2 Setting up the GMM Problem . . . 304
6.2.3 The Existence and Asymptotic Properties of GMM Estimators . . . 305
6.2.4 The Optimal Weighting Matrix . . . 309
6.2.5 Feasible Efficient GMM . . . 314
6.2.6 The 2SLS Estimator . . . 317
6.3 A Test for Instrument Exogeneity: The J Test . . . 321
6.4 The Hausman Specification Test . . . 325
6.4.1 Model Assumptions . . . 325
6.4.2 Hausman's Result for Asymptotically Efficient Estimators . . . 326
6.4.3 Pseudoinversion of Positive Semidefinite Matrices . . . 328
6.4.4 Construction of Hausman Test Statistic . . . 332
6.4.5 The Hausman Test for Testing Endogeneity . . . 334
6.5 Heterogeneity and Local Average Treatment Effects . . . 339
6.5.1 A Model of Random Coefficients . . . 339
6.5.2 Average Treatment Effects under Exogenous Regressors . . . 341
6.5.3 Local Average Treatment Effects under Endogeneity . . . 342

7 Multivariate Linear Regression Models **346**

7.1 Vectorization and Kronecker Product . . . 346
7.2 Convergence Results with MDS Errors . . . 352
7.3 Weak Instruments . . . 360
7.3.1 The Problem with Weak Instruments . . . 361
7.3.2 Inference under Weak Instruments . . . 365
7.4 Efficient System GMM . . . 375
7.4.1 Model Assumptions . . . 375
7.4.2 Moment Conditions and Identification . . . 378
7.4.3 The Efficient System GMM Estimator . . . 379
7.4.4 Feasible Efficient System GMM Estimation . . . 381
7.4.5 Efficiency Gains of System GMM . . . 384
7.5 The Seemingly Unrelated Regression Equations (SURE) Model . . . 385
7.5.1 Model Assumptions . . . 385
7.5.2 The Efficient System GMM Estimator . . . 387
7.5.3 The Special Case $X_{it} = X_t$ . . . 389
7.6 The Simultaneous Equations Model . . . 391
7.6.1 Model Assumptions . . . 391
7.6.2 The Structural and Reduced Form Models . . . 393

7.6.3 Full Information Maximum Likelihood . . . 396
7.6.4 GMM Estimation . . . 400
7.6.5 Three Stage Least Squares . . . 403

8 Non-Parametric Methods 409
8.1 Non-parametric Estimation of the Distribution Function . . . 409
8.1.1 Concentration Inequalities . . . 409
8.1.2 A Consistent Estimator for the CDF: The Glivenko-Cantelli Theorem . . . 417
8.1.3 Rate of Convergence and the Kolmogorov-Smirnov Statistic . . . 428
8.2 Non-Parametric Estimation of Density Functions . . . 432
8.2.1 Motivation: Discrete Approximation to $F'(x)$ . . . 432
8.2.2 Consistency and Asymptotic Normality of Kernel Density Estimators . . . 433
8.2.3 Asymptotically Optimal Bandwidth . . . 442
8.3 Kernel Regression Models . . . 448
8.3.1 Motivating the Kernel Approach . . . 448
8.3.2 Model Assumptions and Preliminary Asymptotic Results . . . 450
8.3.3 Consistency and Asymptotic Normality . . . 456
8.3.4 The Optimal Bandwidth and the Cross Validation Criterion . . . 468
8.4 Local Linear Regression . . . 473
8.4.1 Motivation: The Boundary Bias Problem . . . 473
8.4.2 The Local Linear Estimator and Asymptotic Preliminaries . . . 474
8.4.3 Asymptotic Normality of the Local Linear Estimator . . . 483

9 Local Asymptotic Normality and Asymptotic Efficiency 490
9.1 Contiguity . . . 490
9.1.1 Absolute Continuity and Likelihood Ratios . . . 491
9.1.2 Contiguity and Le Cam's First Lemma . . . 496
9.1.3 Local Alternatives and Le Cam's Third Lemma . . . 512
9.2 Local Asymptotic Normality . . . 519
9.2.1 Motivation and the Gaussian Location Problem . . . 519
9.2.2 The Main Approximation Result . . . 521
9.2.3 Differentiability in Quadratic Mean . . . 528
9.3 Asymptotic Efficiency for Parametric Models . . . 545
9.3.1 A Criterion for Efficiency . . . 545
9.3.2 Efficiency in the Gaussian Location Problem . . . 549
9.3.3 The Convolution Theorem . . . 557
9.3.4 DQM and Asymptotically Efficient Estimators . . . 562
9.4 Asymptotic Efficiency in a Non-parametric Setting . . . 566
9.4.1 Differentiability and the Efficient Influence Function . . . 566
9.4.2 Revisiting the Convolution Theorem . . . 571
9.4.3 Application: The Parametric Model . . . 581
9.4.4 Application: A Completely Non-Parametric Model . . . 584

9.4.5 Application: Models Characterized by Moment Conditions 590
9.4.6 Application: Asymptotic Efficiency for Semiparametric Models 598

10 Empirical Process Theory **606**
10.1 Stochastic Convergence without Measurability 606
10.1.1 Outer Measures and Integrals 607
10.1.2 Stochastic Convergence without Measurability 612
10.2 Convergence of Stochastic Processes 621
10.2.1 Semimetric Spaces 621
10.2.2 The Main Theorem 623
10.2.3 Application: The Functional CLT 632
10.3 Gilvenko-Cantelli and Donsker Classes 638
10.3.1 Bracketing Numbers and the Glivenko-Cantelli Theorem Revisited 638
10.3.2 Maximal Inequalities 641
10.3.3 Donsker's Theorem 658

Chapter 1

Statistical Decision Theory

In this chapter we review some basic elements of statistical decision theory, pioneered by Abraham Wald, and prove a few fundamental results in the field. We will build up to the complete class theorem, which equates the class of admissible decision rules with the class of limiting Bayesian decision rules. Throughout, we let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space.

1.1 The Elements of Decision Theory

A statistical decision problem is a game played by the statistician against nature, characterized by the triple (Θ, A, L):

- Θ is the parameter space, a set from which nature chooses some value θ. This parameter θ is unobservable by the statistician. In most cases, we let ϑ be a σ-algebra on Θ and work with the measurable space (Θ, ϑ).
- A is the action space, a set from which the statistician must choose an action a. As above, we usually work with the measurable space $(A, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra on the set A.
- L is the loss, a function $L : A \times \Theta \to \mathbb{R}_+$ such that, $L(a, \theta)$ is the negative utility felt by the statistician if she chooses the action $a \in \mathcal{A}$ and nature chooses the parameter $\theta \in \Theta$. We usually require L to be a measurable function on $A \times \Theta$.

Where this setup differs from the usual game theoretic framework is in the fact that, when making statistical decisions, the statistician receives some information on nature's choice of θ by observing the data. The data is denoted by X, which is treated as a random variable taking values in some measurable space $(E, \mathcal{E})$.

Let the transition probability kernel $\mu : \Theta \times \mathcal{H} \to [0, 1]$ from (Θ, ϑ) into $(E, \mathcal{E})$ represent the possible distributions of the data X; specifically, let $\mu(\theta, \cdot)$ represent the distribution of X under the parameter $\theta \in \Theta$. This μ is referred to as a **statistical model**. If Θ is finite-dimensional, then we say that the model is parametric; if it is infinite dimensional, then it is non-parametric. In the special case that $\mu(\theta, \cdot)$ is absolutely continuous with respect to, say, the Lebesgue measure for any $\theta \in \Theta$, we can characterize the model via densities instead of distributions.

The statistician's goal is then to choose a decision rule, that is, a measurable function $\delta : E \to A$, such that his expected (negative) utility is minimized. That is, she wants to choose δ in order to minimize expected loss

$$R(\delta, \theta) = \mathbb{E}_\theta \left[L(\delta(X), \theta) \right] := \int_E L(\delta(x), \theta) \mu(\theta, dx).$$

The integral on the right is well-defined and $R(\delta, \cdot)$ is a non-negative ϑ-measurable function because the integrand is a non-negative measurable function on $\mathcal{E} \bigotimes \vartheta$ and μ is a transition probability kernel from (Θ, ϑ) into $(E, \mathcal{E})$.

Letting $\mathcal{D}$ be the set of all decision rules, the function $R : \mathcal{D} \times \Theta \to [0, +\infty]$ defined above is called the risk. Specifically, given a decision rule $\delta \in \mathcal{D}$ and a parameter $\theta \in \Theta$, $R(\delta, \theta)$ is the risk under δ given θ. The following are two types of statistical decision problems that we often deal with:

1. **Hypothesis Testing**
   In this case, the action space is $A = \{\text{accept}, \text{reject}\}$, with $\mathcal{A}$ being the discrete σ-algebra on A. Suppose we want to test the hypotheses

   $$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \notin \Theta_0.$$

   Then, the loss is given as

   $$L(\text{accept}, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \notin \Theta_0 \end{cases}$$

   and

   $$L(\text{reject}, \theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{if } \theta \notin \Theta_0 \end{cases}.$$

   The risk given a decision rule $\delta : E \to A$ is

   $$R(\delta, \theta) = \mathbb{E}_\theta \left[L(\delta(X), \theta) \right] = \begin{cases} \mathbb{P}_\theta \left[\delta(X) = \text{accept} \right] & \text{if } \theta \notin \Theta_0 \\ \mathbb{P}_\theta \left[\delta(X) = \text{reject} \right] & \text{if } \theta \in \Theta_0 \end{cases}.$$

   In other words, $R(\delta, \theta)$ is the probability of a type I error if $\theta \in \Theta_0$ and the probability of a type II error if $\theta \notin \Theta_0$.

2. **Point Estimation**
   In this case, the action space is $A = \Theta$, so that $\mathcal{A} = \vartheta$ as well. There are many loss functions that we use in the context of point estimation; if Θ is a subset of euclidean space, two of

the most popular are the squared loss

$$L_{ms}(\delta,\theta) = |\delta-\theta|^2$$

and the absolute loss

$$L_{abs}(\delta,\theta) = |\delta-\theta|.$$

The squared loss leads to risk being equal to the mean squared error:

$$R_{ms}(\delta,\theta) = \mathbb{E}_\theta|\delta(X)-\theta|^2 = \int_E |\delta(x)-\theta|^2\mu(\theta,dx).$$

1.2 Admissibility of Decision Rules

Given a statistical decision problem $((\Theta,\vartheta),(A,\mathcal{A}),L)$ with data X taking values in $(E,\mathcal{E})$ and statistical model μ, our goal is to choose an optimal decision rule $\delta\in\mathcal{D}$. The most straightforward criterion for optimality we can think of is to find a decision rule δ such that

$$R(\delta,\theta) < R(\delta',\theta)$$

for any $\delta'\in\mathcal{D}$ and $\theta\in\Theta$. It turns out to be very difficult to find such a decision rule, so we need to reformulate our criterion of optimality.

We could first start by ruling out some sub-optimal decision rules. One simple criterion we can think of is to rule out a decision rule δ if there exists another decision rule δ' such that

$$R(\delta',\theta) \leq R(\delta,\theta)$$

for any $\theta\in\Theta$, with the inequality being strict for some θ. Since this means that δ' always does at least as good as δ, there is no reason to continue using δ. We say that, in this case, the decision rule δ is dominated by the decision rule δ'. A decision rule $\delta\in\mathcal{D}$ is said to be **admissible** if it is not dominated by any other decision rule.

The criterion of admissibility seems very reasonable, but it can rule out some very intuitive estimators. The most notable example of this is with the James-Stein estimator, which renders the sample mean inadmissible as an estimator of the sample mean of a normal distribution. Let $X_1,\cdots,X_n$ be a random sample from the m-dimensional normal distribution $\mathcal{N}\left[\mu,\sigma^2\cdot I_m\right]$, where $\sigma^2>0$ is known and $m>2$. Usually, we use the sample mean $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$, to estimate μ; indeed, $\bar{X}_n$ is also the MLE of μ in this case, so it is expected to possess desirable properties.

Consider, however, the James-Stein estimator

$$\hat{\mu}_n = \left(1 - \frac{(m-2)\sigma^2/n}{\left|\bar{X}_n\right|^2}\right)\bar{X}_n.$$

In their 1961 paper, James and Stein prove that this estimator dominates the sample mean, and as such that $\bar{X}_n$ is inadmissible when $m > 2$. Note further how $\hat{\mu}_n$ is a Bayesian estimator of μ given a normal prior. Recall that, under the DGP

$$X_i \mid \mu \sim \mathcal{N}\left[\mu,\, \sigma^2 I_m\right]$$

and prior

$$\mu \sim \mathcal{N}\left[b_0, B_0\right],$$

the posterior distribution of μ given the data $X_1, \cdots, X_n$ is given as

$$\mu \mid X \sim \mathcal{N}\left[\left(\frac{n}{\sigma^2} I_m + B_0^{-1}\right)^{-1}\left(\frac{n}{\sigma^2}\bar{X}_n + B_0^{-1} b_0\right),\, \left(\frac{n}{\sigma^2} I_m + B_0^{-1}\right)^{-1}\right].$$

If $b_0 = \mathbf{0}$ and

$$B_0 = \left(\frac{m-2}{\left|\bar{X}_n\right|^2 - (m-2)\frac{\sigma^2}{n}}\right) I_m,$$

then the posterior mean of μ is precisely the James-Stein estimator $\hat{\mu}_n$; we say that $\hat{\mu}_n$ shrinks the estimator to 0 precisely because the prior mean of μ is equal to 0 in this specification. Note how the prior variance in this case depends on the data; this type of estimator, where the prior hyperparameters are not fixed but rather depend on the data, is called an empirical Bayes estimator. In fact, the discovery of the James-Stein estimator actually led to the development of this class of estimators.

The criterion of admissibility only helps us rule out the most unreasonable decision rules, and we may still be left with many decision rules even after narrowing our search to admissible decision rules. To zero in on a single (or a few) optimal decision rules, we can think of the following criteria:

1. **Bayesian Decision Rules**
   A probability measure π on the measurable space (Θ, ϑ) is referred to as a prior distribution. A Bayesian decision rule δ_π given a prior π is a decision rule that solves the following minimization problem:

$$\min_{\delta \in \mathcal{D}} \quad r(\delta, \pi) = \int_\Theta R(\delta, \theta) d\pi(\theta)$$

The integral on the right hand side is well-defiend for any decision rule $\delta \in \mathcal{D}$ because $R(\delta, \cdot)$ is a non-negative ϑ-measurable function. We refer to $r(\delta, \pi)$ as the Bayes risk given the decision rule δ under the prior π; it turns out that the Bayes risk is equivalent to the expected posterior loss.

2. **Minimax Rules**
   We may want to choose decision rules that minimize risk even in the worst-case scenario. This is formalized by choosing the decision rule that solves the following minimax problem:

$$\min_{\delta \in \mathcal{D}} \left(\sup_{\theta \in \Theta} R(\delta, \theta) \right).$$

The property of Bayesian decision rules, in particular, warrant attention.

1.3 Completeness of Bayesian Decision Rules

The bare minimum we require from any decision rule is that it be admissible. Fortunately, it can easily be shown that any Bayesian decision rule is admissible:

Theorem 1.1 (Admissibility of Bayesian Decision Rules I)
Consider a statistical decision problem $((\Theta, \vartheta), (A, \mathcal{A}), L)$ along with data X taking values in $(E, \mathcal{E})$ and a statistical model μ. Let π be a prior distribution on (Θ, ϑ). If there exists a unique Bayesian decision rule δ_π given the prior π, then δ_π is an admissible decision rule.

Proof) Suppose, for the sake of contradiction, that δ_π is inadmissible. Then, there exists a $\delta \in \mathcal{D}$ distinct from δ_π such that

$$R(\delta, \theta) \leq R(\delta_\pi, \theta)$$

for any $\theta \in \Theta$, where the inequality is strict for some θ. Integrating both sides with respect to π shows us that

$$\int_\Theta R(\delta, \theta) d\pi(\theta) \leq \int_\Theta R(\delta_\pi, \theta) d\pi(\theta).$$

Since $\int_\Theta R(\delta_\pi, \theta) d\pi(\theta)$ is the minimized Bayes risk by the definition of a Bayesian decision rule, it follows that

$$\int_\Theta R(\delta, \theta) d\pi(\theta) = \int_\Theta R(\delta_\pi, \theta) d\pi(\theta).$$

The uniquness of δ_π then tells us that $\delta_\pi = \delta$, which contradicts the fact that δ is distinct from δ_π. Therefore, δ_π must be admissible.

Q.E.D.

We can relax the uniqueness assumption by imposing a continuity assumption. This version of the above result is given as follows:

Theorem 1.2 (Admissibility of Bayesian Decision Rules II)
Consider a statistical decision problem $((\Theta, \vartheta), (A, \mathcal{A}), L)$ along with data X taking values in $(E, \mathcal{E})$ and a statistical model μ. Suppose Θ is a subset of some euclidean space $\mathbb{R}^m$, and let the risk funtion $R(\delta, \cdot)$ be continuous on Θ for any $\delta \in \mathcal{D}$. Finally, let π be a prior distribution on (Θ, ϑ) such that π is absolutely continuous with respect to the Lebesgue measure on Θ with a density that is positive everywhere on Θ.

If $\delta_\pi(X)$ is a Bayesian estimator of θ given the prior π such that $R(\delta_\pi, \theta) < +\infty$ for any $\theta \in \Theta$, then δ_π is an admissible decision rule.

Proof) Suppose, for the sake of contradiction, that δ_π is inadmissible. Then, there exists a $\delta \in \mathcal{D}$ distinct from δ_π such that

$$R(\delta, \theta) \leq R(\delta_\pi, \theta)$$

for any $\theta \in \Theta$, where the inequality is strict for some θ. $R(\delta_\pi, \cdot)$ is real-valued on Θ, so $R(\delta, \cdot)$ is also real-valued on Θ.

Integrating both sides with respect to π shows us that

$$\int_\Theta R(\delta, \theta) d\pi(\theta) \leq \int_\Theta R(\delta_\pi, \theta) d\pi(\theta).$$

Since $\int_\Theta R(\delta_\pi, \theta) d\pi(\theta)$ is the minimized Bayes risk by the definition of a Bayesian decision rule, it follows that

$$\int_\Theta R(\delta, \theta) d\pi(\theta) = \int_\Theta R(\delta_\pi, \theta) d\pi(\theta),$$

which implies

$$\int_\Theta \left(R(\delta_\pi, \theta) - R(\delta, \theta) \right) d\pi(\theta) = 0.$$

The integrand $\theta \mapsto R(\delta_\pi, \theta) - R(\delta, \theta)$ is non-negative by design, so by the vanishing property of non-negative functions, there exists a set $\Theta_0 \in \vartheta$ such that $\pi(\Theta_0) = 1$ and

$$R(\delta_\pi, \theta) = R(\delta, \theta)$$

for any $\theta \in \Theta_0$.

Choose any $\theta \in \Theta_0^c$, if any exists; by definition, $R(\delta_\pi, \theta) > R(\delta, \theta)$. Therefore, by the continuity of $R(\delta, \cdot)$ on Θ, there exists an $\eta > 0$ such that

$$\left|(R(\delta,\theta) - R(\delta_\pi,\theta)) - (R(\delta,\theta') - R(\delta_\pi,\theta'))\right| < R(\delta_\pi,\theta) - R(\delta,\theta)$$

for any $\theta' \in \Theta$ such that $|\theta' - \theta| < \eta$. By implication, if $|\theta' - \theta| < \eta$, then

$$R(\delta,\theta') - R(\delta_\pi,\theta') > 0,$$

that is, the open ball $B(\theta, \eta) \subset \Theta_0^c$. It follows that

$$0 < \pi(B(\theta,\eta)) \leq \pi(\Theta_0^c),$$

which contradicts the fact that $\pi(\Theta_0^c) = 0$. Therefore, δ_π must be admissible.

Q.E.D.

It is a remarkable result that we can actually assert that the converse holds true as well under certain assumptions; that is, under certain conditions any admissible decision rule must be a Bayesian decision rule given some prior distribution. For the sake of simplicity, we prove this result for finite parameter spaces only.

Theorem 1.3 (Complete Class Theorem I)
Consider a statistical decision problem $((\Theta, \vartheta), (A, \mathcal{A}), L)$ along with data X taking values in $(E, \mathcal{E})$ and a statistical model μ. Assume the following:

i) Θ is a finite set $\{\theta_1, \cdots, \theta_m\}$.

ii) A is a nonempty convex subset of some euclidean space $\mathbb{R}^d$, and $\mathcal{A}$ the corresponding Borel σ-algebra.

iii) $L(\cdot, \theta)$ is a convex function for any $\theta \in \Theta$.

iv) There exists some decision rule $\delta_0 \in \mathcal{D}$ with finite risk, that is, $R(\delta_0, \theta) < +\infty$ for any $\theta \in \Theta$.

Then, for any admissible decision rule δ^*, there exists a probability distribution $\pi = (\pi_1, \cdots, \pi_m) \in [0,1]^m$ on Θ such that δ^* is a Bayesian decision rule given prior equal to π.

Proof) First, define the set $S \subset \mathbb{R}^m$ as

$$S = \{r \in \mathbb{R}^m \mid \exists \delta \in \mathcal{D}, \quad R(\delta, \theta_j) \leq r_j \quad \forall 1 \leq j \leq m\},$$

as the set of vectors that dominate a certain set of risks associated with some decision rule $\delta \in \mathcal{D}$. S is non-empty, since by assumption there exists a decision rule $\delta_0 \in \mathcal{D}$ with finite risk. Furthermore, S is convex; for any $r, s \in S$ and $\lambda \in [0,1]$, there exist $\delta_r, \delta_s \in \mathcal{D}$ such that $R(\delta_r, \theta_j) = r_j$ and $R(\delta_s, \theta_j) = s_j$ for any $1 \leq j \leq n$. Defning $\tilde{\delta}$ as

$$\tilde{\delta}(x) = \lambda \cdot \delta_r(x) + (1-\lambda) \cdot \delta_s(x)$$

for any $x \in E$, $\tilde{\delta} \in \mathcal{D}$ because it is measurable with respect to $\mathcal{E}$ and $\mathcal{A}$. For any $1 \leq j \leq m$, by the convexity of $L(\cdot, \theta_j)$, we can see that

$$\begin{aligned} R(\tilde{\delta}, \theta_j) &= \int_E L(\tilde{\delta}(x), \theta_j) \mu(\theta, dx) \\ &\leq \lambda \cdot \int_E L(\delta_r(x), \theta_j) \mu(\theta, dx) + (1-\lambda) \cdot \int_E L(\delta_s(x), \theta_j) \mu(\theta, dx) \\ &= \lambda \cdot R(\delta_r, \theta_j) + (1-\lambda) \cdot R(\delta_s, \theta_j) \\ &\leq \lambda r_j + (1-\lambda) s_j. \end{aligned}$$

It follows that $\lambda r + (1-\lambda)s \in S$, with the dominated decision rule being the convex combination $\tilde{\delta}$ of δ_r and δ_s.

Let $\delta^* \in \mathcal{D}$ be an admissible decision rule, and define

$$r^* = \left(R(\delta^*, \theta_1), \cdots, R(\delta^*, \theta_m)\right).$$

Since $R(\delta^*, \theta_j) \leq R(\delta_0, \theta_j)$ for any $1 \leq j \leq m$ by the definition of admissibility, so it follows that $r^* \in \mathbb{R}^m$. Furthermore, $r^* \in S$ by definition.

Now define

$$Q^* = \{r \in \mathbb{R}^m \mid r < r^*\}.$$

This is clearly a non-empty and convex set. In addition, Q^* and S are disjoint. To see this, choose any $r \in Q^*$, so that $r_j < r_j^*$ for any $1 \leq j \leq m$. Suppose $r \in S$. Then there exists a $\delta \in \mathcal{D}$ such that $R(\delta, \theta_j) \leq r_j$ for any $1 \leq j \leq m$, so that

$$R(\delta, \theta_j) < r_j^* = R(\delta^*, \theta_j) \quad \text{for any } 1 \leq j \leq m$$

This contradicts the admissibiltiy of δ^*, so $r \notin S$.

So far, we have shown that Q^* and S are non-empty convex sets such that $Q^* \cap S = \emptyset$. Therefore, by the separating hyperplane theorem, there exists a non-zero $v \in \mathbb{R}^m$ such that

$$v'x \leq v'y$$

for any $x \in Q^*$ and $y \in S$. Letting $\{r_n\}_{n \in N_+}$ be defined as $r_n = r^* - \frac{1}{n} \cdot \iota \in Q^*$, where $\iota \in \mathbb{R}^m$ is a vector of ones, we can see that, for any $y \in S$,

$$v' r_n \leq v'$$

for any $n \in N_+$. Since $r_n \to r^*$ as $n \to \infty$, taking limits on both sides shows us that

$$v' r^* \leq v' y.$$

Furthermore, since $r^* \in S$, $v' x \leq v' r^*$ for any $x \in Q^*$. Putting these results together, we can see that

$$v' x \leq v' r^* \leq v' y$$

for any $x \in Q^*$ and $y \in S$.

Note that v must be non-negative valued; to see this, suppose $v_j < 0$ for some $1 \leq j \leq n$, and choose some $r \in Q^*$. Suppose $v' r = v' r^*$. Then, letting e_j be the jth standard basis vector of $\mathbb{R}^m$, $r - e_j \in Q^*$ as well, so that

$$v'(r - e_j) = v' r + |v_j| = v' r^* + a|v_j| > v' r^*,$$

a contradiction. On the other hand, if $v' r < v' r^*$, then defining $\epsilon = v' r^* - v' r > 0$, we can see that $r - \frac{2\epsilon}{|v_j|} e_j \in Q^*$ and therefore

$$v' \left(r - \frac{\epsilon}{|v_j|} e_j \right) = v' r + 2\epsilon = v' r^* + \epsilon > v' r^*,$$

which results in another contradiction. It follows that $v \in \mathbb{R}^m_+$, so we can define

$$\pi = \frac{1}{\sum_{j=1}^m v_j} v \in \mathbb{R}^m_+,$$

where the elements of π are non-negative and sum to 1.

Finally, we show that π is the prior distribution on Θ that we are looking for. This follows by noting that

$$\pi' r^* \leq \pi' r$$

for any $r \in S$. For any decision rule $\delta \in \mathcal{D}$, letting $r_j = R(\delta, \theta_j)$ for any $1 \leq j \leq m$, if $r_j = +\infty$ for any $1 \leq j \leq m$, then we trivially have

$$\sum_{j=1}^m \pi_j \cdot R(\delta^*, \theta_j) = \pi' r^* < \sum_{j=1}^m \pi_j \cdot R(\delta, \theta_j).$$

Suppose that $r \in \mathbb{R}^m$. We can also see that

$$\sum_{j=1}^{m} \pi_j \cdot R(\delta^*, \theta_j) = \pi' r^* \leq \pi' r = \sum_{j=1}^{m} \pi_j \cdot r_j = \sum_{j=1}^{m} \pi_j \cdot R(\delta, \theta_j)$$

in this case. In other words,

$$\sum_{j=1}^{m} \pi_j \cdot R(\delta^*, \theta_j) = \min_{\delta \in \mathcal{D}} \quad \sum_{j=1}^{m} \pi_j \cdot R(\delta, \theta_j),$$

so that δ^* is a Bayesian decision rule under the prior distribution π.

Q.E.D.

1.4 The Sufficiency Principle

The use of sufficient statistics is another way to narrow down our search for admissible decision rules. We begin by defining and discussing the properties of conditional expectations with respect to an arbitrary probability measure.

Let $(E, \mathcal{E}, \mu)$ be some probability space. Completely analogously to how we define the conditional expectation of random variables with respect to some σ-algebra, we define the conditional expectation of the non-negative measurable function $f \in \mathcal{E}_+$ with respect to the sub σ-algebra $\mathcal{F}$ of $\mathcal{E}$ as the non-negative function $\bar{f} \in \mathcal{F}_+$ such that

$$\int_A f d\mu = \int_A \bar{f} d\mu \quad \text{for any } A \in \mathcal{F}.$$

We already know that the following properties hold, from the results we derived for the conditional expectation of non-negative random variables:

- For any $f \in \mathcal{E}_+$, the set of all conditional expectations of f with respect to $\mathcal{F}$ is an equivalence class of functions that are equal μ-almost everywhere. The representative of this class, as well as the class itself, is denoted by $\mu(f \mid \mathcal{F})$.

- For any $f \in \mathcal{E}_+$, the set $\mu(f \mid \mathcal{F})$ of conditional expectations of f is non-empty.

- **(Projection Property)**
  For any $f \in \mathcal{E}_+$ and $h \in \mathcal{F}_+$,
  $$\int_E fh \, d\mu = \int_E \mu(f \mid \mathcal{F}) h \, d\mu.$$

- **(Monotonicity)**
  For any $f, g \in \mathcal{E}_+$ such that $f \leq g$ μ-a.e.,
  $$\mu(f \mid \mathcal{F}) \leq \mu(g \mid \mathcal{F}).$$

- **(Linearity)**
  For any functions $f, g \in \mathcal{E}_+$ and $a \in \mathbb{R}_+$,
  $$\mu(af + g \mid \mathcal{F}) = a \cdot \mu(f \mid \mathcal{F}) + \mu(g \mid \mathcal{F}).$$

- **(Conditional Determinism)**

For any $f \in \mathcal{E}_+$ and $h \in \mathcal{F}_+$,

$$\mu(fh \mid \mathcal{F}) = \mu(f \mid \mathcal{F})h,$$

where the equivalence is again μ-a.e.

- **(Tower Property)**
  Let $\mathcal{G}$ be a sub σ-algebra of $\mathcal{E}$ such that $\mathcal{F} \subset \mathcal{G}$. Then, for any $f \in \mathcal{E}_+$,

$$\mu(f \mid \mathcal{F}) = \mu(\mu(f \mid \mathcal{F}) \mid \mathcal{G}) = \mu(\mu(f \mid \mathcal{G}) \mid \mathcal{F}).$$

We say that a real-valued $\mathcal{E}$-measurable function $f : E \to \mathbb{R}$ is conditionally μ-integrable given $\mathcal{F}$ if $\mu(f^+ \mid \mathcal{F})$ and $\mu(f^- \mid \mathcal{F})$ are finite μ-almost everywhere. In this case, we define

$$\mu(f \mid \mathcal{F}) = \mu(f^+ \mid \mathcal{F}) - \mu(f^- \mid \mathcal{F}).$$

Note that, if f is μ-integrable, then f is conditionally μ-integrable given any sub σ-algebra of $\mathcal{E}$. The set of all functions that are conditionally μ-integrable given $\mathcal{F}$ is a linear subspace of the set of all functions from E to $\mathbb{R}$, and all the above properties generalize to conditional expectations involving conditionally μ-integrable functions.

Finally, a vector-valued $\mathcal{E}$-measurable function $f : E \to \mathbb{R}^k$ is said to be conditionally μ-integrable given $\mathcal{F}$ if each of its components $f_1, \cdots, f_k$ are conditionally μ-integrable given $\mathcal{F}$. The following theorem holds for conditionally integrable vector valued functions:

- **(Jensen's Inequality)**
  For any convex measurable function $g : \mathbb{R}^k \to \mathbb{R}$ and $f : E \to \mathbb{R}^k$ such that $g \circ f$ and f are conditionally μ-integrable given $\mathcal{F}$,

$$g\left(\mu(f \mid \mathcal{F})\right) \leq \mu(g(f) \mid \mathcal{F}) \quad \mu - \text{almost everywhere.}$$

Now let (Θ, ϑ) be the parameter space, X the data taking values in the measurable space $(E, \mathcal{E})$, and μ the corresponding statistical model. Let $(F, \mathcal{F})$ be another measurable space and $T : E \to F$ a measurable mapping. We call any such function T a statistic. Denote by σT the sub σ-algebra of $\mathcal{E}$ generated by the statistic T, that is,

$$\sigma T = \{T^{-1}(H) \mid H \in \mathcal{F}\}.$$

Fix any $\theta \in \Theta$; suppose this is the true parameter value so that $\mu(\theta, \cdot)$ becomes the distribution of X. For any $A \in \mathcal{E}$, $\mathbb{P}(X \in A \mid T(X))$ is measurable relative to the σ-algebra generated by $T(X)$. By the Doob-Dynkin lemma, it follows that there exists some $\mathcal{F}$-measurable function $f_{A,\theta}$ such that $\mathbb{P}(X \in A \mid T(X)) = f_{A,\theta} \circ T(X)$.

We let the conditional expectation of any function $f \in \mathcal{E}_+$ with respect to $\mu(\theta,\cdot)$ given σT be denoted by $\mu_\theta(f \mid T)$. Then, for any $H \in \mathcal{F}$,

$$\begin{aligned}\int_{T^{-1}(H)} \mu_\theta(I_A \mid T) d\mu(\theta,\cdot) &= \mu(\theta,\ A \cap T^{-1}(H)) \quad \text{(Definition of the Conditional Expectation)} \\ &= \mathbb{E}\left[I_{X^{-1}(A)} \cdot I_{\{T(X)\in H\}}\right] \\ &= \mathbb{E}\left[\mathbb{P}\left(X^{-1} \in A \mid T(X)\right) \cdot I_{\{T(X)\in H\}}\right] \\ &= \int_{T^{-1}(H)} (f_{A,\theta} \circ T) d\mu.\end{aligned}$$

Since both $\mu_\theta(I_A \mid T)$ and $f_{A,\theta} \circ T$ are σT-measurable non-negative functions, it follows that

$$\mu_\theta(I_A \mid T) = f_{A,\theta} \circ T \quad \mu - \text{a.e.}$$

In other words,

$$\mu_\theta(I_A \mid T) \circ X = f_{A,\theta} \circ T(X) = \mathbb{P}\left(X \in A \mid T(X)\right).$$

Note that, in general, the conditional probability of $\{X \in A\}$ given the statistic $T(X)$ depends on the parameter θ.

We say that $T \in \mathcal{E}/\mathcal{F}$ is a sufficient statistic for the model μ if, for any $A \in \mathcal{E}$, there exists some $\psi_A \in \mathcal{F}_+$ such that

$$\mu_\theta(I_A \mid T) = \psi_A \circ T$$

for any $\theta \in \Theta$. In other words, T is a sufficient statistic if the conditional probabilities of X given $T(X)$ do not depend on θ. The following theorem illustrates the importance of sufficient statistics for decision theory:

Theorem 1.4 (Rao-Blackwell Theorem)
Consider a statistical decision problem $((\Theta,\vartheta),(A,\mathcal{A}),L)$ along with data X taking values in $(E,\mathcal{E})$ and a statistical model μ. Let $(F,\mathcal{F})$ be a measurable space and $T \in \mathcal{E}/\mathcal{F}$ a sufficient statistic for μ. Suppose that $A = \mathbb{R}^k$ and $\mathcal{A} = \mathcal{B}(\mathbb{R}^k)$, and that $L(\cdot,\theta)$ is a convex function for any $\theta \in \Theta$. Then, for any decision rule $\delta \in \mathcal{D}$ such that $R(\delta,\theta) < +\infty$ and $\int_E |\delta(x)| \mu(\theta, dx) < +\infty$ for $1 \le i \le k$ and $\theta \in \Theta$, there exists a function $\delta^* : F \to A$ that is measurable relative to $\mathcal{F}$ and $\mathcal{A}$ such that:

i) For any $\theta \in \Theta$, we have

$$\delta^* \circ T = \mu_\theta(\delta \mid T)$$

ii) The risk under $\delta^* \circ T$ is never greater than that under δ:

$$R(\delta^* \circ T, \theta) \leq R(\delta, \theta) \quad \text{for any } \theta \in \Theta.$$

Proof) By the definition of sufficiency, for any $1 \leq i \leq k$ there exists some $\psi_i \in \mathcal{F}_+$ such that

$$\mu_\theta(\delta_i \mid T) = \psi_i \circ T.$$

Define

$$\delta^*(t) = (\psi_1(t), \cdots, \psi_k(t))'$$

for any $t \in F$. Then, δ^* is measurable relative to $\mathcal{F}$ and $\mathcal{B}(\mathbb{R}^k)$, and immediately satisfies

$$\delta^* \circ T = \mu_\theta(\delta \mid T)$$

for any $\theta \in \Theta$. It remains to show that $R(\delta^* \circ T, \cdot) \leq R(\delta, \cdot)$ on Θ.

To this end, choose any $\theta \in \Theta$ and note that the assumptions $R(\delta, \theta) < +\infty$ and $\int_E |\delta(x)| \mu(\theta, dx) < +\infty$ imply that $L(\delta, \theta)$ and δ are $\mu(\theta, \cdot)$-integrable. As such, $L(\delta, \theta)$ and δ are conditionally $\mu(\theta, \cdot)$-integrable given σT, and

$$L(\delta^* \circ T, \theta) = L(\mu_\theta(\delta \mid T), \theta) \leq \mu_\theta \left(L(\delta, \theta) \mid T \right)$$

$\mu(\theta, \cdot)$-a.e. by Jensen's inequality. It follows that

$$\begin{aligned} R(\delta^* \circ T, \theta) &= \int_E L(\delta^*(T(x)), \theta) \mu(\theta, dx) \\ &\leq \int_E \mu_\theta \left(L(\delta, \theta) \mid T \right) d\mu(\theta, \cdot) \\ &= \int_E L(\delta, \theta) d\mu(\theta, \cdot) \qquad \text{(Definition of Conditional Expectation)} \\ &= R(\delta, \theta). \end{aligned}$$

This holds for any $\theta \in \Theta$, so the proof is complete.

Q.E.D.

The Rao-Blackwell theorem shows us that, under convex loss functions, we can always improve upon an existing decision rule by taking its conditional expectation with respect to a sufficient statistic. In other words, we can reduce our search for admissible decision rules by searching

for functions of sufficient statistics. This is a powerful simplification if the statistical model of interest admits a particularly simple sufficient statistic.

The search for sufficient statistics can be facilitated by means of the factorization theorem. First, we present a preliminary lemma. We say that two collections Π and Ξ of probability measures on some measurable space $(E,\mathcal{E})$ are equivalent if, for any $A \in \mathcal{E}$, $v(A) = 0$ for any $v \in \Pi$ implies that $\lambda(A) = 0$ for any $\lambda \in \Xi$, and vice versa.

Lemma 1.5 Let $(E,\mathcal{E})$ be a measurable space and μ a σ-finite measure on that space. If Π is a collection of probability measures on $(E,\mathcal{E})$ that are absolutely continuous with respect to μ, then there exists an at most countable subset of Π_0 such that Π and Π_0 are equivalent.

Proof) We first define the following collection of probability measures on $(E,\mathcal{E})$:

$$\mathcal{M} = \Big\{ \sum_{i=1}^{\infty} c_i P_i \mid \{P_i\}_{i \in N_+} \subset \Pi, \quad c_i \geq 0 \; \forall i, \quad \sum_{i=1}^{\infty} c_i = 1 \Big\}.$$

Note that $\mathcal{M}$ contains the collection Π, and that every measure in $\mathcal{M}$ is also absolutely continuous with respect to μ. Furthermore, $\mathcal{M}$ is also clearly a convex set.

We also define the following collection of $\mathcal{E}$-measurable sets:

$$\mathcal{C} = \Big\{ C \in \mathcal{E} \mid \exists Q \in \mathcal{M} \quad \text{s.t.} \quad \mu\Big(A \cap C \cap \Big\{ \frac{\partial Q}{\partial \mu} > 0 \Big\} \Big) = \mu(A \cap C) \quad \forall A \in \mathcal{E}, \quad Q(C) > 0 \Big\}.$$

The sets in $\mathcal{C}$ are called equivalence sets for the class $\mathcal{M}$ and the σ-finte measure μ for the following reason. Choose any $C \in \mathcal{E}$ and $Q \in \mathcal{M}$ such that

$$\mu\Big(A \cap C \cap \Big\{ \frac{\partial Q}{\partial \mu} > 0 \Big\} \Big) = \mu(A \cap C)$$

for any $A \in \mathcal{E}$ and $Q(C) > 0$. Choose any $A \in \mathcal{E}$ such that $Q(A \cap C) = 0$. Then,

$$\begin{aligned} 0 = Q(A \cap C) &= \int_{A \cap C} \frac{\partial Q}{\partial \mu} d\mu \\ &= \int_{A \cap C \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\}} \frac{\partial Q}{\partial \mu} d\mu. \end{aligned}$$

By the vanishing property for non-negative functions, it follows that

$$\frac{\partial Q}{\partial \mu}(x) I_{A \cap C \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\}}(x) = 0$$

for any $x \in E_0$, where $E_0 \in \mathcal{E}$ satisfies $\mu(E_0^c) = 0$. Suppose that $E_0 \cap \Big(A \cap C \cap \Big\{ \frac{\partial Q}{\partial \mu} > 0 \Big\} \Big) \neq \emptyset$. Then, for any $x \in E_0 \cap \Big(A \cap C \cap \Big\{ \frac{\partial Q}{\partial \mu} > 0 \Big\} \Big)$, we must have $\frac{\partial Q}{\partial \mu}(x) = 0$, which is a contradiction. By implication,

$$A \cap C \cap \Big\{ \frac{\partial Q}{\partial \mu} > 0 \Big\} \subset E_0^c,$$

so that

$$\mu(A\cap C)=\mu\left(A\cap C\cap\left\{\frac{\partial Q}{\partial\mu}>0\right\}\right)=0.$$

In other words, if $Q(A\cap C)=0$ for any $A\in\mathcal{E}$, then $\mu(A\cap C)=0$; the trace of μ on C is absolutely continuous with respect to the trace of Q on C. The converse is automatically true because $Q<<\mu$, so it follows that the trace of Q on C and the corresponding trace of μ are equivalent. Therefore, for each $C\in\mathcal{C}$ there exists a measure $Q\in\mathcal{M}$ such that C is a non-null set on which Q and μ are equivalent.

We continue to investigate the properties of $\mathcal{C}$ by studying the types of sets that are contained in $\mathcal{C}$:

- **Support of any Measure in $\mathcal{M}$**
  For any $Q\in\mathcal{M}$ and $B\in\mathcal{E}$, any set $C=B\cap\left\{\frac{\partial Q}{\partial\mu}>0\right\}$ such that $Q(C)>0$ is contained in $\mathcal{C}$ with corresponding measure Q, since

  $$\mu\left(A\cap C\cap\left\{\frac{\partial Q}{\partial\mu}>0\right\}\right)=\mu(A\cap C)$$

  for any $A\in\mathcal{E}$ by the definition of C. In particular,

  $$\begin{aligned}1=Q(E)&=Q\left(\left\{\frac{\partial Q}{\partial\mu}>0\right\}\right)+Q\left(\left\{\frac{\partial Q}{\partial\mu}=0\right\}\right)\\&=Q\left(\left\{\frac{\partial Q}{\partial\mu}>0\right\}\right)>0\end{aligned}$$

  implies that $\left\{\frac{\partial Q}{\partial\mu}>0\right\}\in\mathcal{C}$ for any $Q\in\mathcal{M}$.

- **Closedness under Finite Unions of Disjoint Sets**
  Choose any disjoint $C_1,C_2\in\mathcal{C}$. Then, for each $i=1,2$ there exists some $Q_i\in\mathcal{M}$ such that

  $$\mu\left(A\cap C_i\cap\left\{\frac{\partial Q_i}{\partial\mu}>0\right\}\right)=\mu(A\cap C_i)\quad\forall A\in\mathcal{E}$$

  and $Q_i(C_i)>0$. Defining $Q=\frac{1}{2}Q_1+\frac{1}{2}Q_2\in\mathcal{M}$, the support of Q is exactly the union of the supports Q_1 and Q_2: since

  $$\frac{\partial Q}{\partial\mu}=\frac{1}{2}\frac{\partial Q_1}{\partial\mu}+\frac{1}{2}\frac{\partial Q_2}{\partial\mu}$$

  on some $E_0\in\mathcal{E}$ such that $\mu(E_0^c)=0$, $\frac{\partial Q}{\partial\mu}>0$ on E_0 if and only if either $\frac{\partial Q_1}{\partial\mu}>0$ or $\frac{\partial Q_2}{\partial\mu}>0$ on E_0. In other words,

  $$\left\{\frac{\partial Q}{\partial\mu}>0\right\}\cap E_0=\left(\left\{\frac{\partial Q_1}{\partial\mu}>0\right\}\cup\left\{\frac{\partial Q_2}{\partial\mu}>0\right\}\right)\cap E_0.$$

Therefore, for any $i = 1, 2$ and $A \in \mathcal{E}$,

$$\begin{aligned}\mu(A \cap C_i) &= \mu\left(A \cap C_i \cap \left\{\frac{\partial Q_i}{\partial \mu} > 0\right\}\right) \\ &\leq \mu\left(A \cap C_i \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) \leq \mu(A \cap C_i),\end{aligned}$$

so that

$$\mu(A \cap C_i) = \mu\left(A \cap C_i \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right).$$

This shows us that Q is equivalent to μ on C_1 and C_2. Letting $C = C_1 \cup C_2$, it follows that

$$\begin{aligned}\mu\left(A \cap C \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) &= \mu\left[\left(A \cap C_1 \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) \cup \left(A \cap C_2 \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right)\right] \\ &= \mu\left(A \cap C_1 \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) + \mu\left(A \cap C_2 \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) \\ &= \mu(A \cap C_1) + \mu(A \cap C_2) = \mu(A \cap C),\end{aligned}$$

where we used finite additivity alongside the fact that $C_1 \cap C_2 = \emptyset$. Finally,

$$\begin{aligned}Q(C_1 \cup C_2) &= \frac{1}{2} Q_1(C_1 \cup C_2) + \frac{1}{2} Q_2(C_1 \cup C_2) \\ &\geq \frac{1}{2} Q_1(C_1) + \frac{1}{2} Q_2(C_2) > 0,\end{aligned}$$

so that Q is equivalent to μ on C; $C \in \mathcal{C}$ by definition.

Suppose that Q^* is a measure in $\mathcal{M}$ that is maximally equivalent to μ, that is, Q^* and μ are equivalent on $C^* \in \mathcal{C}$, where $\mu(C^*) \geq \mu(C)$ for any $C \in \mathcal{C}$. Then, it must intuitively be the case that Q^* dominates any probability measure Q in $\mathcal{M}$, since any $Q \in \mathcal{M}$ will be equivalent to μ on a subset of E that is smaller than C^* in terms of μ-measure. Formally, choose any $Q \in \mathcal{M}$ and let $Q^*(A) = 0$ for some $A \in \mathcal{E}$. Then, $\mu(A \cap C^*) = 0$ because μ and Q^* are equivalent on C^*, and $Q(A \cap C^*) = 0$ follows from the fact that $Q << \mu$. Furthermore,

$$Q\left(A \cap (C^*)^c \cap \left\{\frac{\partial Q}{\partial \mu} = 0\right\}\right) = \int_{A \cap (C^*)^c \cap \left\{\frac{\partial Q}{\partial \mu} = 0\right\}} \frac{\partial Q}{\partial \mu} d\mu = 0.$$

Finally, if

$$Q\left(A \cap (C^*)^c \cap \left\{\frac{\partial Q}{\partial \mu} > 0\right\}\right) > 0,$$

then $A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \in \mathcal{C}$ and therefore so is

$$C^* \cup \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right),$$

since this is a union of two disjoint sets in $\mathcal{C}$. It follows from finite additivity that

$$\mu \left[C^* \cup \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) \right] = \mu(C^*) + \mu \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) > \mu(C^*),$$

since $Q << \mu$ and $Q \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) > 0$ implies that $\mu \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) > 0$ as well. This contradicts the fact that $\mu(C^*) \geq \mu(C)$ for any $C \in \mathcal{C}$, so it must be the case that

$$Q \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) = 0,$$

and we have

$$Q(A) = Q(A \cap C^*) + Q \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} = 0 \right\} \right) + Q \left(A \cap (C^*)^c \cap \left\{ \frac{\partial Q}{\partial \mu} > 0 \right\} \right) = 0.$$

This holds for any $A \in \mathcal{E}$ such that $Q^*(A) = 0$, so $Q << Q^*$.

It remains to construct this maximally equivalent measure Q^*. To this end, define

$$\alpha = \sup_{C \in \mathcal{C}} \mu(C) \in [0, 1],$$

and let $\{C_i\}_{i \in N_+}$ be a sequence in $\mathcal{C}$ such that $\mu(C_i) \nearrow \alpha$. Suppose that $\{Q_i\}_{i \in N_+}$ is a sequence of measures in $\mathcal{M}$ such that each C_i is an equivalence set for Q_i. Let $c_i^* = 2^{-i}$ for each $i \in N_+$, and define

$$Q^* = \sum_{i=1}^{\infty} c_i^* Q_i \in \mathcal{M}.$$

Defining

$$q^* = \sum_{i=1}^{\infty} c_i^* \frac{\partial Q_i}{\partial \mu} \in \mathcal{E}_+,$$

by the MCT for series we can see that

$$\begin{aligned} Q^*(A) = \sum_{i=1}^{\infty} c_i^* Q_i(A) &= \sum_{i=1}^{\infty} \int_E c_i^* \left(\frac{\partial Q_i}{\partial \mu} \cdot I_A \right) d\mu \\ &= \int_A \left(\sum_{i=1}^{\infty} c_i^* \frac{\partial Q_i}{\partial \mu} \right) d\mu = \int_A q^* d\mu \end{aligned}$$

for any $A \in \mathcal{E}$, so that

$$q^* = \frac{\partial Q}{\partial \mu} \quad \mu\text{-almost everywhere.}$$

Defining $C^* = \bigcup_{i=1}^{\infty} C_i \in \mathcal{E}$, we can see that $\mu(C_i) \leq \mu(C^*)$ for any $i \in N_+$, so that $\alpha \leq \mu(C^*)$. Furthermore, $q^*(x) > 0$ if and only if $\frac{\partial Q_i}{\partial \mu}(x) > 0$ for at least one $i \in N_+$, so that

$$\{q^* > 0\} = \bigcup_i \Big\{ \frac{\partial Q_i}{\partial \mu} > 0 \Big\}.$$

For any $A \in \mathcal{E}$, we now have

$$\begin{aligned} \mu\left(A \cap C^* \cap \{q^* = 0\}\right) &\leq \sum_{i=1}^{\infty} \mu\left(A \cap C_i \cap \{q^* = 0\}\right) \\ &\leq \sum_{i=1}^{\infty} \mu\left(A \cap C_i \cap \Big\{ \frac{\partial Q_i}{\partial \mu} = 0 \Big\}\right) \end{aligned}$$

by countable subadditivity. For any $i \in N_+$, the fact that $Q_i\left(A \cap C_i \cap \left\{ \frac{\partial Q_i}{\partial \mu} = 0 \right\}\right) = 0$ implies that

$$\mu\left(A \cap C_i \cap \Big\{ \frac{\partial Q_i}{\partial \mu} = 0 \Big\}\right) = 0$$

as well because Q_i and μ are equivalent on C_i. Therefore,

$$\mu\left(A \cap C^* \cap \{q^* = 0\}\right) = 0$$

and

$$\mu(A \cap C^*) = \mu\left(A \cap C^* \cap \{q^* = 0\}\right) + \mu\left(A \cap C^* \cap \{q^* > 0\}\right) = \mu\left(A \cap C^* \cap \{q^* > 0\}\right).$$

This holds for any $A \in \mathcal{E}$, and

$$Q(C^*) = \sum_{i=1}^{\infty} c_i^* Q_i(C^*) \geq \sum_{i=1}^{\infty} c_i^* Q_i(C_i) > 0,$$

so by definition $C^* \in \mathcal{C}$. This establishes that $\mu(C^*) \leq \alpha$, so that

$$\mu(C^*) = \alpha \geq \mu(C) \quad \text{for any } C \in \mathcal{C}.$$

Q^* is the desired dominating measure because Q^* and μ are equivalent on C^*.

Finally, because $Q^* \in \mathcal{M}$ there exists a sequence $\{P_i\}_{i \in N_+} \subset \Pi$ and a sequence $\{c_i\}_{i \in N_+}$

of non-negative reals with sum equal to 1 such that

$$Q^* = \sum_{i=1}^{\infty} c_i P_i.$$

Define $\Pi_0 = \{P_i\}_{i \in N_+}$. Clearly, if, for some $A \in \mathcal{E}$, $P(A) = 0$ for any $P \in \Pi$, then $P_i(A) = 0$ for any $i \in N_+$. Conversely, suppose, for some $A \in \mathcal{E}$, that $P_i(A) = 0$ for any $i \in N_+$. Then, $Q^*(A) = 0$, so that $Q(A) = 0$ for any $Q \in \mathcal{M}$. Since $\Pi \subset \mathcal{M}$, it follows that $P(A) = 0$ for any $P \in \Pi$ as well. This establishes that Π and Π_0 are equivalent.

Q.E.D.

Now we can prove the following factorization theorem, whose general form is due to Halmos and Savage (the above lemma is also part of their contribution, but the proof is taken from Lehmann). The proof of the factorization theorem presented below is due to Kagan.

Theorem 1.6 (Factorization Theorem)
Let (Θ, ϑ) be the parameter space, and X the data taking values in the measurable space $(E, \mathcal{E})$ with corresponding statistical model μ. Let $(F, \mathcal{F})$ be a measurable space and $T : E \to F$ a statistic for μ. Suppose there exist some σ-finite measure λ on $(E, \mathcal{E})$ such that $\mu(\theta, \cdot) << \lambda$ for any $\theta \in \Theta$. Then, T is sufficient for μ if and only if there exists a function $r \in \mathcal{E}_+$ and a collection $\{h_\theta\}_{\theta \in \Theta}$ of $\mathcal{F}$-measurable non-negative functions such that, for each $\theta \in \Theta$,

$$\frac{\partial \mu(\theta, \cdot)}{\partial \lambda}(x) = r(x) h_\theta(T(x))$$

for any λ-almost every $x \in E$.

Proof) Let $\Pi = \{\mu(\theta, \cdot) \mid \theta \in \Theta\}$. By the preceding lemma, there exists a sequence $\{\theta_i\}_{i \in N_+}$ of parameters such that the countable collection of probability measures $\Pi_0 = \{\mu(\theta_i, \cdot)\}_{i \in N_+}$ is equivalent to Π. Choose any sequence $\{c_i\}_{i \in N_+}$ of non-negative real numbers such that $\sum_{i=1}^{\infty} c_i = 1$, and define the probability measure

$$Q = \sum_{i=1}^{\infty} c_i \mu(\theta_i, \cdot).$$

Note that $Q << \lambda$, and that, if $Q(A) = 0$ for some $A \in \mathcal{E}$, then $\mu(\theta_i, A) = 0$ for any $i \in N_+$ and therefore $\mu(\theta, A) = 0$ for any $\theta \in \Theta$. Equivalently, we can say that $\mu(\theta, \cdot) << Q$ for any $\theta \in \Theta$. Because Q is a probability measure and thus finite, it follows from the Radon-Nikodym theorem that there exists a Radon-Nikodym derivative $\frac{\partial \mu(\theta, \cdot)}{dQ}$ that is Q-almost surely unique for each $\theta \in \Theta$.

Necessity

Suppose that T is a sufficient statistic for μ. Then, by definition, for any $A \in \mathcal{E}$ there exists a function $\psi_A \in \mathcal{F}_+$ such that

$$\mu_\theta(I_A \mid T) = \psi_A \circ T$$

for any $\theta \in \Theta$. Note that, since

$$\mu(\theta, A \cap T^{-1}(H)) = \int_E \left(I_A \cdot I_{T^{-1}(H)} \right) d\mu(\theta, \cdot) = \int_{T^{-1}(H)} \mu_\theta(I_A \mid T) d\mu(\theta, \cdot)$$

for any $\theta \in \Theta$ and $H \in \mathcal{F}$ by the definition of conditional expectations, it follows that

$$\begin{aligned} Q(A \cap T^{-1}(H)) = \sum_{i=1}^{\infty} c_i \mu(\theta_i, A \cap T^{-1}(H)) &= \sum_{i=1}^{\infty} c_i \left(\int_{T^{-1}(H)} \mu_\theta(I_A \mid T) d\mu(\theta_i, \cdot) \right) \\ &= \sum_{i=1}^{\infty} c_i \left(\int_{T^{-1}(H)} (\psi_A \circ T) d\mu(\theta_i, \cdot) \right) \\ &= \int_{T^{-1}(H)} (\psi_A \circ T) dQ \end{aligned}$$

for any $H \in \mathcal{F}$ as well. By definition, it follows that

$$Q(I_A \mid T) = \psi_A \circ T$$

as well. Now we can see that, for any $\theta \in \Theta$,

$$\begin{aligned} \int_A \frac{\partial \mu(\theta, \cdot)}{\partial Q} dQ &= \mu(\theta, A) && \text{(Definition of RN Derivative)} \\ &= \int_E \mu_\theta(I_A \mid T) d\mu(\theta, \cdot) && \text{(Law of Iterated Expectations)} \\ &= \int_E (\psi_A \circ T) d\mu(\theta, \cdot) && \text{(Definition of } \psi_A\text{)} \\ &= \int_E (\psi_A \circ T) \frac{\partial \mu(\theta, \cdot)}{\partial Q} dQ && \text{(Expression of Integral using RN Derivative)} \\ &= \int_E (\psi_A \circ T) \cdot Q\left(\frac{\partial \mu(\theta, \cdot)}{\partial Q} \mid T \right) dQ && \text{(Projection Property)} \\ &= \int_E Q(I_A \mid T) \cdot Q\left(\frac{\partial \mu(\theta, \cdot)}{\partial Q} \mid T \right) dQ \\ &= \int_E Q\left(I_A \cdot Q\left(\frac{\partial \mu(\theta, \cdot)}{\partial Q} \mid T \right) \mid T \right) dQ && \text{(Conditional Determinism)} \\ &= \int_E I_A \cdot Q\left(\frac{\partial \mu(\theta, \cdot)}{\partial Q} \mid T \right) dQ && \text{(Law of Iterated Expectations)} \end{aligned}$$

$$= \int_A Q\left(\frac{\partial \mu(\theta,\cdot)}{\partial Q} \mid T\right) dQ.$$

This equation holds for any $A \in \mathcal{E}$, so it follows that

$$\frac{\partial \mu(\theta,\cdot)}{\partial Q} = Q\left(\frac{\partial \mu(\theta,\cdot)}{\partial Q} \mid T\right)$$

Q-almost surely. Define $r : E \to [0, +\infty]$ as

$$r = \frac{\partial Q}{\partial \lambda},$$

and by the Doob-Dynkin lemma there exists a function $h_\theta \in \mathcal{F}_+$ such that

$$h_\theta \circ T = Q\left(\frac{\partial \mu(\theta,\cdot)}{\partial Q} \mid T\right).$$

Then, for any $A \in \mathcal{E}$,

$$\begin{aligned}
\int_A \frac{\partial \mu(\theta,\cdot)}{\partial \lambda} d\lambda &= \mu(\theta, A) \\
&= \int_A \frac{\partial \mu(\theta,\cdot)}{\partial Q} dQ \\
&= \int_A (h_\theta \circ T) dQ \\
&= \int_A (h_\theta \circ T) \frac{\partial Q}{\partial \lambda} d\lambda = \int_A (h_\theta \circ T) r \, d\lambda.
\end{aligned}$$

Therefore,

$$\frac{\partial \mu(\theta,\cdot)}{\partial \lambda}(x) = h_\theta(T(x)) r(x)$$

for λ-a.e. $x \in E$.

Sufficiency

Now suppose that there exists a function $r \in \mathcal{E}_+$ and a collection $\{h_\theta\}_{\theta \in \Theta}$ of $\mathcal{F}$-measurable non-negative functions such that, for each $\theta \in \Theta$,

$$\frac{\partial \mu(\theta,\cdot)}{\partial \lambda}(x) = r(x) h_\theta(T(x))$$

for any λ-almost every $x \in E$. We want to show that T is a sufficient statistic for μ.

We can immediately see that the factorization above implies

$$\begin{aligned}\int_A \frac{\partial Q}{\partial \lambda} d\lambda = Q(A) &= \sum_{i=1}^{\infty} c_i \mu(\theta_i, A) \\ &= \sum_{i=1}^{\infty} c_i \left(\int_A \frac{\partial \mu(\theta_i, \cdot)}{\partial \lambda} d\lambda \right) \\ &= \sum_{i=1}^{\infty} c_i \left(\int_A (h_{\theta_i} \circ T) r \; d\lambda \right) \\ &= \int_A \left[\left(\sum_{i=1}^{\infty} c_i h_{\theta_i} \right) \circ T \right] r \; d\lambda \qquad \text{(MCT for Series)}\end{aligned}$$

for any $A \in \mathcal{E}$. Therefore, defining the function $G = \sum_{i=1}^{\infty} c_i h_{\theta_i} \in \mathcal{F}_+$,

$$\frac{\partial Q}{\partial \lambda} = (G \circ T) r$$

λ-almost everywhere.

Now fix $\theta \in \Theta$. Since

$$(h_\theta \circ T) r = \frac{\partial \mu(\theta, \cdot)}{\partial \lambda} = \frac{\partial \mu(\theta, \cdot)}{\partial Q} \times \frac{\partial Q}{\partial \lambda} = \frac{\partial \mu(\theta, \cdot)}{\partial Q} \times (G \circ T) r$$

λ-almost everywhere, it is not unreasonable to suspect that the quotient $(h_\theta \circ T) r / (G \circ T) r$ will be Q-a.e. equal to the derivative $\frac{\partial \mu(\theta, \cdot)}{\partial Q}$. Formally, define $g_\theta : F \to [0, +\infty]$ as

$$g_\theta = \frac{h_\theta}{G} (I_{(0,+\infty)} \circ G).$$

Note that

$$Q\left(\left\{ \frac{\partial Q}{\partial \lambda} = 0 \right\} \right) = 0$$

trivially. Furthermore,

$$\int_E \frac{\partial Q}{d\lambda} d\lambda = Q(E) = 1 < +\infty$$

implies, by the finiteness property for non-negative functions, that

$$Q\left(\left\{ \frac{\partial Q}{d\lambda} = +\infty \right\} \right) = 0.$$

As such,

$$Q\left(\left\{ \frac{\partial Q}{d\lambda} \notin (0, +\infty) \right\} \right) = 0.$$

Let $E_0 \in \mathcal{E}$ be chosen so that $\lambda(E_0^c) = 0$ and

$$h_\theta(T(x))r(x) = \left(\frac{\partial \mu(\theta, \cdot)}{\partial Q}(x)\right) G(T(x))r(x)$$

for any $x \in E_0$. For any $x \in E_0 \cap \{(G \circ T)r \in (0, +\infty)\}$, we can see that

$$\frac{\partial \mu(\theta, \cdot)}{\partial Q}(x) = g_\theta(T(x)),$$

where the value on the right hand side is well-defined because $G(T(x)) \in (0, +\infty)$ if $G(T(x))r(x) \in (0, +\infty)$. Because

$$Q\left(\{(G \circ T)r \notin (0, +\infty)\} \cup E_0^c\right) \leq Q\left(\left\{\frac{\partial Q}{d\lambda} \notin (0, +\infty)\right\}\right) + Q(E_0^c) = 0,$$

we can see that

$$\frac{\partial \mu(\theta, \cdot)}{\partial Q} = g_\theta \circ T$$

Q-almost surely.

For any $A \in \mathcal{E}$ and $H \in \mathcal{F}$,

$$\begin{aligned}
\mu(\theta, A \cap T^{-1}(H)) &= \int_{A \cap T^{-1}(H)} (g_\theta \circ T) dQ \\
&= \int_{T^{-1}(H)} Q(I_A (g_\theta \circ T) \mid T) dQ \\
&\qquad\qquad \text{(Definition of Conditional Expectation)} \\
&= \int_{T^{-1}(H)} Q(I_A \mid T)(g_\theta \circ T) dQ \qquad \text{(Conditional Determinism)} \\
&= \int_{T^{-1}(H)} Q(I_A \mid T) d\mu(\theta, \cdot).
\end{aligned}$$

As such, $Q(I_A \mid T)$ is a version of the conditional probability of A given σT:

$$\mu_\theta(I_A \mid T) = Q(I_A \mid T)$$

$\mu(\theta, \cdot)$-a.e. By the Doob-Dynkin lemma, there exists a function $\psi_A \in \mathcal{F}_+$ such that $Q(I_A \mid T) = \psi_A \circ T$, so that

$$\mu_\theta(I_A \mid T) = \psi_A \circ T$$

$\mu(\theta, \cdot)$-a.e. This holds for any $\theta \in \Theta$, and we can find such a function ψ_A for any set $A \in \mathcal{E}$, so by definition T is sufficient for μ.

Q.E.D.

The factorization theorem has the following implication for statistical inference. Suppose we believe in the sufficiency principle, which tells us that we should draw the same conclusion for two statistical problems with the same sufficient statistic. Formally, consider two statistical decision making problems with the same parameter space, action space and loss function, but with possibly different statistical models. The sufficiency principle dictates that, if these two models admit sufficient statistics that have the same value given some realized data, the statistician must arrive at the same decision in both problems. In the case of convex loss functions, the Rao-Blackwell theorem tells us that we should always condition on sufficient statistics to avoid working with inadmissible decision rules, so the sufficiency principle is defensible on grounds of admissibility.

Consider now two statistical decision making problems with the same parameter space, action space and loss function $((\Theta,\vartheta),(A,\mathcal{A}),L)$ and statistical models μ_1,μ_2 on the data space $(E,\mathcal{E})$. Assume both models are dominated by the same σ-finite measure λ on $(E,\mathcal{E})$. Suppose now that the realized data in an experiment is given by the deterministic value $x \in E$. The likelihood principle dictates that, if the likelihoods for the two models at this x are proportional to one another, that is,

$$\frac{\partial \mu_1(\theta,\cdot)}{\partial \lambda}(x) = k\frac{\partial \mu_2(\theta,\cdot)}{\partial \lambda}(x)$$

for any $\theta \in \Theta$, then the two problems should arrive at the same decision.

Through the factorization theorem, we can easily see that the likelihood principle implies the sufficiency principle. Suppose the models μ_1,μ_2 admit the sufficient statistics T_1,T_2. By the factorization theorem, there exist non-negative measurable functions r_1,r_2 and collections $\{h_{\theta,1}\}_{\theta\in\Theta},\{h_{\theta,2}\}_{\theta\in\Theta}$ of non-negative measurable functions such that

$$\frac{\partial \mu_i(\theta,\cdot)}{\partial \lambda}(x) = r_i(x)h_{\theta,i}(T_i(x))$$

for $i=1,2$, $\theta\in\Theta$ and $x\in E$. Now let $x\in E$ be the realized data in an experiment, and suppose $T_1(x)=T_2(x)=t$. Then,

$$\frac{\partial \mu_i(\theta,\cdot)}{\partial \lambda}(x) = r_i(x)h_{\theta,i}(t)$$

for any $\theta\in\Theta$ and $i=1,2$, so that

Chapter 2

Hypothesis Testing and Set Estimation

In this chapter we delve into the theory on hypothesis testing and set estimation. Point estimation, which requires heavy asymptotic machinery, will be addressed in later chapters, with a separate focus on frequentist and Bayesian methods.

2.1 Hypothesis Testing

As we mentioned in the chapter on decision theory, hypothesis tests are statistical decision problems described by $((\Theta, \vartheta), A, L)$ where the action space is given by $A = \{accept, reject\}$. Let X be the data taking values in the measurable space $(E, \mathcal{E})$, and μ the corresponding statistical model, and consider testing the null hypothesis

$$H_0 : \theta \in \Theta_0$$

against the alternative

$$H_1 : \theta \notin \Theta_0,$$

where Θ_0 is some measurable subset of Θ. The loss function in this case is in general given by

$$L(accept, \theta) = \begin{cases} 0 & \text{if } \theta \in \Theta_0 \\ 1 & \text{if } \theta \notin \Theta_0 \end{cases}, \quad L(reject, \theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{if } \theta \notin \Theta_0 \end{cases}.$$

Then, the risk associated with a decision rule $\delta : E \to A$, which we call a test of H_0 against H_1, is

$$R(\delta, \theta) = \mathbb{E}_\theta \left[L(\delta(X), \theta) \right] = \begin{cases} \mathbb{P}_\theta \left[\delta(X) = reject \right] & \text{if } \theta \in \Theta_0 \\ \mathbb{P}_\theta \left[\delta(X) = accept \right] & \text{if } \theta \notin \Theta_0 \end{cases}.$$

2.1.1 Core Concepts of Hypothesis Testing

Below we introduce some important concepts related to hypothesis tests.

Rejection Region When we identify $accept = 0$ and $reject = 1$, any test $\delta : E \to A$ is a mapping from the data space E to $\{0,1\}$, and therefore δ can be written as an indicator function for some set $R \in \mathcal{E}$, that is,

$$\delta(x) = I_R(x) \quad \text{for any } x \in E.$$

In other words, under the test δ, we choose to reject the null if the data is contained in the set $R \subset E$. This is why we call R the rejection region associated with the test δ.

Type I and II Errors Given a test $\delta : E \to A$ with rejection region $R \in \mathcal{E}$, the risk associated with δ can be written as

$$R(\delta, \theta) = \begin{cases} \mathbb{P}_\theta [X \in R] & \text{if } \theta \in \Theta_0 \\ \mathbb{P}_\theta [X \notin R] & \text{if } \theta \notin \Theta_0. \end{cases}$$

When $\theta \in \Theta_0$, the probability $\mathbb{P}_\theta [X \in R]$ is the probability of a type I error under θ, that is, the probability of falsely rejecting a true null. On the other hand, if $\theta \notin \Theta_0$, then $\mathbb{P}_\theta [X \notin R]$ is the probability of a type II error under θ, that is, the probability of not rejecting a false null. The goal of the statistician is to find a test δ that minimizes the risk $R(\delta, \theta)$, that is, a test that minimizes the probability of the associated type I and II errors.

Power of a Test A related concept is the power of a test under some $\theta \notin \Theta_0$, which is defined as

$$\mathbb{P}_\theta [X \in R] = 1 - \mathbb{P}_\theta [X \notin R],$$

which is the probability of rejecting a false null; it is inversely proportional to the probability of a type II error under θ.

Size of a Test In an ideal world, we could find a test δ that minimizes the probability of both type I and type II errors. However, there exists a tradeoff between the probability of type I and II errors; the more likely we are to reject the null under $\theta \notin \Theta_0$, thus reducing the probability of type II errors, the more likely we are to reject it under $\theta \in \Theta_0$ as well, which raises the probability of a type I error.

Usually, type I errors are taken to be more serious than type II errors. Therefore, we would like to bound the probability of a type I error. A useful object in this case is the size of a test; the size of a test δ with rejection region R is defined as

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta [X \in R],$$

that is, the maximum probability of rejecting the null when the null is true.

Significance Level Returning to our original objective, we want to consider tests that bound the probability of type I errors at a certain level. A test δ with rejection region R is said to have significance level α if the size of δ does not exceed α, that is, if

$$\mathbb{P}_\theta [X \in R] \leq \alpha$$

for any $\theta \in \Theta_0$.

Power Function Among tests with the same significance level, that is, tests whose type I error probabilities are bounded at the same level, we want to search for the test under which the probability of type II errors are minimized. A useful object in this case is the power function; the power function $\beta_\delta : \Theta \to [0, 1]$ of a test δ with rejection region R is defined as

$$\beta_\delta(\theta) = \mathbb{P}_\theta [X \in R] = \mu(\theta, R)$$

for any $\theta \in \Theta$. We call β_δ the power function of the test δ because $\beta_\delta(\theta)$ is precisely the power of the test under any $\theta \notin \Theta_0$. Note also that

$$\sup_{\theta \in \Theta_0} \beta_\delta(\theta)$$

is the size of the test δ.

UMP Tests Among different tests with the same significance level α, we ideally want to find tests that maximize the power function on Θ_0^c; in other words, we want to find a test δ with significance level α such that, for any other test ϕ with significance level α,

$$\beta_\delta(\theta) \geq \beta_\phi(\theta)$$

for any $\theta \notin \Theta_0$. These types of tests are called uniformly most powerful (UMP) tests.

Unbiasedness In most cases, it is difficult to find a test that is uniformly most powerful across all values of Θ_0^c. It is therefore customary to impose additional restrictions on the kinds of tests that we are interested in, and try to find UMP tests among the tests that satisfy these additional restrictions. One reasonable criterion is to restrict our attention to unbiased tests. A test δ with significance level α is said to be unbiased if

$$\beta_\delta(\theta) \geq \alpha$$

for any $\theta \notin \Theta_0$. In other words, in an unbiased test the probability of rejection when the null is false is at least as high as the probability of rejection when the null is true.

P-value The p-value is way other than the power function to characterize the properties of a test. Consider a collection $\{R_\alpha\}_{\alpha\in(0,1)}$ of nested rejection regions, that is, a collection of $\mathcal{E}$-measurale sets such that $R_\alpha \subset R_{\alpha'}$ whenever $\alpha \leq \alpha'$. Let δ_α be the test with rejection region R_α. Then, the p-value associated with the sequence of tests $\{\delta_\alpha\}_{\alpha\in(0,1)}$ is the function $\phi : E \to [0,1]$ defined as

$$\phi(x) = \inf\{\alpha \in (0,1) \mid x \in R_\alpha\},$$

for any $x \in E$. Heuristically, we can interpret $\phi(X)$ as the smallest significance level under which we still reject the null hypothesis.

2.1.2 Uniformly Most Powerful Tests

Remarkably, for certain null and alternative hypotheses, there exist UMP tests; this is the content of the Neyman-Pearson lemma and its generalization, the Karlin-Rubin theorem. We consider the following setting. Let $((\Theta,\vartheta),A,L)$ be our statistical decision problem, where $A=\{0,1\}$, with 0 representing acceptance and 1 representing rejection. Let X be the data taking values in the measurable space $(E,\mathcal{E})$, and μ the corresponding statistical model, which as usual we take as a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$. Suppose in addition that there exists a σ-finite measure v on $(E,\mathcal{E})$ such that, for any $\theta\in\Theta$, the distribution $\mu(\theta,\cdot)$ of the data is absolutely continuous with respect to v:

$$\mu(\theta,\cdot) << v \quad \text{for any } \theta\in\Theta.$$

By the Radon-Nikodym theorem, this implies that there exists some non-negative function $f: E\times\Theta\to[0,+\infty]$ such that $f(\cdot\mid\theta)$ is $\mathcal{E}$-measurable and

$$\mu(\theta,A)=\int_A f(x\mid\theta)dv(x)$$

for any $\theta\in\Theta$ and $A\in\mathcal{E}$. We assume that f is real-valued and $\mathcal{E}\bigotimes\vartheta$-measurable.

This is the setting we work with in most applications, where E is assumed to be a subset of some euclidean space and v the Lebesgue or counting measure on that euclidean space. $f(x\mid\theta)$ is then interpreted as the density of the data given the parameter θ.

Suppose that $\Theta=\{\theta_0,\theta_1\}$, so that there are only two parameters of interest, and consider testing the null $H_0:\theta=\theta_0$ against the alternative $H_1:\theta=\theta_1$. In this case, the Neyman-Pearson lemma shows that there exists a simple UMP test based on the likelihood ratio:

Theorem 2.1 (Neyman-Pearson Lemma)
Let $\Theta=\{\theta_0,\theta_1\}$ and consider testing the null $H_0:\theta=\theta_0$ against the alternative $H_1:\theta=\theta_1$. Let $\alpha\in(0,1)$ be a predetermined significance level. Then, the test δ^* with rejection region

$$R^*=\{x\in E\mid f(x\mid\theta_1)>\lambda f(x\mid\theta_0)\},$$

where $\lambda\geq 0$ is chosen so that $\beta_{\delta^*}(\theta_0)=\alpha$, is a UMP test among tests with the same significance level.

Proof) Choose any test δ with rejection region $R\in\mathcal{E}$ and significance level α. By definition,

$$\beta_\delta(\theta_0)=\mu(\theta_0,R)\leq\alpha.$$

Now define the function $g: E\to\mathbb{R}$ as

$$g(x)=[\delta^*(x)-\delta(x)]\,[f(x\mid\theta_1)-\lambda f(x\mid\theta_0)]$$

for any $x \in E$. Since δ^*, δ are indicator functions of $\mathcal{E}$-measurable sets, and $f(\cdot \mid \theta_1), f(\cdot \mid \theta_0)$ are $\mathcal{E}$-measurable non-negative real functions, g is also $\mathcal{E}$-measurable. In addition, g is a non-negative valued function. To see this, we consider two cases. Suppose first that $f(x \mid \theta_1) > \lambda f(x \mid \theta_0)$ for some $x \in E$. Then, since $\delta^*(x) = 1 \geq \delta(x)$, by definition $g(x) \geq 0$. On the other hand, if $f(x \mid \theta_1) \leq \lambda f(x \mid \theta_0)$ for some $x \in E$, then $\delta^*(x) = 0 \leq \delta(x)$, so that $g(x)$, being the product of two non-positive terms, is non-negative. In any case, $g(x) \geq 0$ for any $x \in E$.

By the monotonicity and linearity of integration,

$$\begin{aligned}
0 \leq \int_E g dv &= \int_E [\delta^*(x) - \delta(x)] [f(x \mid \theta_1) - \lambda f(x \mid \theta_0)] \, dv(x) \\
&= \int_E \delta^*(x) f(x \mid \theta_1) dv(x) - \int_E \delta(x) f(x \mid \theta_1) dv(x) \\
&\quad - \lambda \left[\int_E \delta^*(x) f(x \mid \theta_0) dv(x) - \int_E \delta(x) f(x \mid \theta_0) dv(x) \right] \\
&= \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1) - \lambda [\beta_{\delta^*}(\theta_0) - \beta_\delta(\theta_1)] \\
&= \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1) - \lambda [\alpha - \beta_\delta(\theta_1)] .
\end{aligned}$$

Since the test δ has significance level α and $\lambda \geq 0$,

$$\lambda [\alpha - \beta_\delta(\theta_1)] \geq 0,$$

which implies that

$$0 \leq \int_E g dv \leq \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1),$$

or equivalently,

$$\beta_{\delta^*}(\theta_1) \geq \beta_\delta(\theta_1).$$

This shows us that δ^* is the uniformly most powerful test.

Q.E.D.

This lemma shows us that for simple hypotheses where Θ_0 is a singleton, there exists a UMP test formulated in terms of the likelihood ratio. It is even more remarkable that, under the right conditions, this is the only UMP test up to almost sure equivalence.

Theorem 2.2 (Unqiueness of the NP Test)
Let $\Theta = \{\theta_0, \theta_1\}$ and consider testing the null $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$. Let $\alpha \in (0,1)$ be a predetermined significance level, and suppose there exists a Neyman-Pearson test δ^{NP} with rejection region

$$R^{NP} = \{x \in E \mid f(x \mid \theta_1) > \lambda f(x \mid \theta_0)\},$$

where $\lambda \geq 0$ is chosen to satisfy $\beta_{\delta^{NP}}(\theta_0) = \alpha$. Suppose that:

i) The critical value λ is chosen so that $\lambda > 0$.

ii) The set

$$A = \{x \in E \mid f(x \mid \theta_1) = \lambda f(x \mid \theta_0)\}$$

is a negligible set under both $\mu(\theta_0, \cdot)$ and $\mu(\theta_1, \cdot)$

Then, for any UMP test δ^* of H_0 against H_1 with significance level α,

$$\mathbb{P}_\theta \left[\delta^*(X) = \delta^{NP}(X)\right] = 1$$

for $\theta = \theta_0, \theta_1$.

Proof) Let δ^* be a UMP test for H_0 against H_1 with significance level α. We first show that δ^* must have size equal to α. To see this, note that, since δ^{NP} is a UMP test for H_0 against H_1 with size α, we must have

$$\begin{aligned} \beta_{\delta^*}(\theta_0) &\leq \alpha = \beta_{\delta^{NP}}(\theta_0) \\ \beta_{\delta^*}(\theta_1) &= \beta_{\delta^{NP}}(\theta_1). \end{aligned}$$

Therefore, from what we showed in the last theorem,

$$\begin{aligned} 0 &\leq \beta_{\delta^{NP}}(\theta_1) - \beta_{\delta^*}(\theta_1) - \lambda\left[\beta_{\delta^{NP}}(\theta_0) - \beta_{\delta^*}(\theta_0)\right] \\ &= -\lambda\left[\alpha - \beta_{\delta^*}(\theta_0)\right], \end{aligned}$$

or equivalently,

$$\lambda\left[\alpha - \beta_{\delta^*}(\theta_0)\right] \leq 0.$$

Since $\lambda > 0$ by assumption, we must have $\beta_{\delta^*}(\theta_0) = \alpha$, that is, δ^* has size α.

Now we show that δ^* must equal δ^{NP} with probability 1. As in the previous theorem,

$$g(x) = \left[\delta^{NP}(x) - \delta^*(x)\right]\left[f(x \mid \theta_1) - \lambda f(x \mid \theta_0)\right] \geq 0,$$

for any $x \in E$, while the integral of g with respect to v is given as

$$\int_E g dv = \beta_{\delta^{NP}}(\theta_1) - \beta_{\delta^*}(\theta_1) - \lambda\left[\beta_{\delta^{NP}}(\theta_0) - \beta_{\delta^*}(\theta_0)\right] = 0.$$

Therefore, by the vanishing property for non-negative functions, there exists an $E_0 \in \mathcal{E}$ such that $v(E_0^c) = 0$ and

$$\left[\delta^{NP}(x) - \delta^*(x)\right]\left[f(x \mid \theta_1) - \lambda f(x \mid \theta_0)\right] = 0$$

for any $x \in E_0$. Define $E_1 = E_0 \cap A^c$. We consider two cases:

Case 1: $x \in R^{NP} \cap E_1$

If $x \in R^{NP}$, the $f(x \mid \theta_1) > \lambda f(x \mid \theta_0)$, it follows that

$$\delta^{NP}(x) - \delta^*(x) = 0$$

This establishes that

$$R^{NP} \cap E_1 \subset \{x \in E \mid \delta^{NP}(x) = \delta^*(x)\} \cap E_1.$$

Case 2: $x \in E_1 \cap (R^{NP})^c$

Since $x \notin R^{NP}$, we have $f(x \mid \theta_1) \leq \lambda f(x \mid \theta_0)$, and since $x \in E_1 \subset A^c$, it follows that $f(x \mid \theta_1) < \lambda f(x \mid \theta_0)$. Again, since $x \in E_0$, this implies that $\delta^{NP}(x) - \delta^*(x) = 0$, so that

$$(R^{NP})^c \cap E_1 \subset \{x \in E \mid \delta^{NP}(x) = \delta^*(x)\} \cap E_1.$$

We have thus shown that

$$E_1 = \{x \in E \mid \delta^{NP}(x) = \delta^*(x)\} \cap E_1.$$

For any $\theta \in \Theta$, both A and E_0^c are negligible sets under $\mu(\theta, \cdot)$, so $E_1^c = A \cup E_0^c$ is also negligible under $\mu(\theta, \cdot)$. It follows that, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{P}_\theta\left[\delta^*(X) = \delta^{NP}(X)\right] &= \mu(\theta, \{\delta^* = \delta^{NP}\} \cap E_1) + \mu(\theta, \{\delta^* = \delta^{NP}\} \cap E_1^c) \\
&= \mu(\theta, \{\delta^* = \delta^{NP}\} \cap E_1) \\
&= \mu(\theta, E_1) = 1.
\end{aligned}$$

Q.E.D.

While the Neyman-Pearson lemma is only stated for simple hypotheses, where the parameter space consists only of two points, it is possible to accomodate composite one-way hypotheses. Specifically, for $\Theta \subset \mathbb{R}$, it is possible to extend the above result for testing hypotheses of the form $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. In this case, we require that the densities satisfy an additional property:

MLR Property Let $\Theta \subset \mathbb{R}$. The densities $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ are said to satisfy the monotone likelihood ratio (MLR) property in the statistic $T(X)$ if there exists a measurable function $T : E \to \mathbb{R}$ such that, for any $\theta_0, \theta_1 \in \Theta$ such that $\theta_0 \leq \theta_1$, there exists an increasing function $h : \mathbb{R} \to \mathbb{R}_+$ such that

$$\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} = h(T(x))$$

for any $x \in E$.

If the statistical model μ admits a family of densities $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ with respect to the σ-finite measure v on $(E, \mathcal{E})$ that satisfy the MLR property, then we can formulate a generalization of the Neyman-Pearson lemma as follows:

Theorem 2.3 (Karlin-Rubin Theorem)
Let $\Theta = \mathbb{R}$ and let ϑ be the associated Borel σ-algebra on Θ. Suppose $\{f(\cdot \mid \theta) \mid \theta \in \Theta\}$ satisfies the MLR property in the statistic $T(X)$.

Consider testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Let δ^* be a test with rejection region R^* defined as

$$R^* = \{x \in E \mid T(x) > \lambda\}$$

for some $\lambda \in \mathbb{R}$ chosen so that $\beta_{\delta^*}(\theta_0) = \alpha \in (0, 1)$. Then, the power function β_{δ^*} is an increasing function on Θ, and δ^* is a uniformly powerful test with significance level α.

Proof) **δ^* has an Increasing Power Function**

We first show that β_{δ^*} is an increasing function on Θ. Choose any $\tilde{\theta}_0, \tilde{\theta}_1 \in \Theta$ such that $\tilde{\theta}_0 < \tilde{\theta}_1$. Then, by the MLR property, there exists an increasing function $h : \mathbb{R} \to \mathbb{R}_+$ such that

$$\frac{f(x \mid \tilde{\theta}_1)}{f(x \mid \tilde{\theta}_0)} = h(T(x))$$

for any $x \in E$. If $x \in R^*$, then $T(x) > \lambda$ and the monotonicity of h shows us that

$$f(x \mid \tilde{\theta}_1) \geq k \cdot f(x \mid \tilde{\theta}_0),$$

where $k = h(\lambda)$. On the other hand, if $x \notin R^*$, then $T(x) \leq \lambda$ and

$$f(x \mid \tilde{\theta}_1) \leq k \cdot f(x \mid \tilde{\theta}_0).$$

Define $\tilde{\alpha} = \beta_{\delta^*}(\tilde{\theta}_0)$ and $\phi : E \to [0,1]$ as $\phi(x) = \tilde{\alpha}$ for any $x \in E$. Then, we can easily see that

$$[\delta^*(x) - \phi(x)] \left[f(x \mid \tilde{\theta}_1) - k \cdot f(x \mid \tilde{\theta}_0) \right] \geq 0$$

for any $x \in E$; therefore,

$$\begin{aligned} 0 &\leq \int_E [\delta^*(x) - \phi(x)] \left[f(x \mid \tilde{\theta}_1) - k \cdot f(x \mid \tilde{\theta}_0) \right] dv(x) \\ &= \beta_{\delta^*}(\tilde{\theta}_1) - \tilde{\alpha} - k \left[\beta_{\delta^*}(\tilde{\theta}_0) - \tilde{\alpha} \right] = \beta_{\delta^*}(\tilde{\theta}_1) - \tilde{\alpha}. \end{aligned}$$

It follows that

$$\beta_{\delta^*}(\tilde{\theta}_0) = \tilde{\alpha} \leq \beta_{\delta^*}(\tilde{\theta}_1).$$

This shows us that β_{δ^*} is an increasing function on Θ. Consequently,

$$\beta_{\delta^*}(\theta) \leq \beta_{\delta^*}(\theta_0) = \alpha$$

for any $\theta \leq \theta_0$, so that δ^* is a test of level α.

δ^* is a UMP Test

Let δ be any test with rejection region $R \in \mathcal{E}$ and significance level α. Choose any $\theta_1 > \theta_0$. Then, by the MLR property, there exists an increasing function $h : \mathbb{R} \to \mathbb{R}_+$ such that

$$\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)} = h(T(x))$$

for any $x \in E$. If $x \in R^*$, then $T(x) > \lambda$, so that

$$f(x \mid \theta_1) = h(T(x)) f(x \mid \theta_0) > k \cdot f(x \mid \theta_0),$$

where $k = h(\lambda) \geq 0$. On the other hand, if $x \notin R^*$, then $T(x) \leq \lambda$ and similarly we get

$$f(x \mid \theta_1) \leq k \cdot f(x \mid \theta_0).$$

As in the proof of the Neyman-Pearson lemma, note that this implies

$$[\delta^*(x) - \delta(x)] \left[f(x \mid \theta_1) - k \cdot f(x \mid \theta_0) \right] \geq 0$$

for any $x \in E$. Integrating both sides with respect to v now yields

$$\begin{aligned} 0 &\leq \int_E [\delta^*(x) - \delta(x)] \, [f(x \mid \theta_1) - k \cdot f(x \mid \theta_0)] \, dv(x) \\ &= \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1) - k \, [\beta_{\delta^*}(\theta_0) - \beta_\delta(\theta_0)] \\ &= \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1) - k \, [\alpha - \beta_\delta(\theta_0)] \leq \beta_{\delta^*}(\theta_1) - \beta_\delta(\theta_1). \end{aligned}$$

Therefore,

$$\beta_\delta(\theta_1) \leq \beta_{\delta^*}(\theta_1),$$

and since this holds for any $\theta_1 \in \Theta_1$, δ^* is a uniformly most powerful test of H_0 against H_1 with significance level α.

Q.E.D.

As an application of the Karlin-Rubin theorem, consider a normally distributed random sample $X_1, \cdots, X_n \sim \mathcal{N}\left[\mu, \sigma^2\right]$, where μ is unknown but σ^2 is known. Suppose we want to test the null hypothesis

$$H_0 : \mu \leq \mu_0$$

against the alternative

$$H_1 : \mu > \mu_0.$$

The likelihood ratio for two different values $\mu_0, \mu_1 \in \mathbb{R}$ in this case is given as

$$\begin{aligned} \frac{f(x \mid \mu_1)}{f(x \mid \mu_0)} &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \Big((x_i - \mu_1)^2 - (x_i - \mu_0)^2 \Big) \right) \\ &= \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \Big(\mu_1^2 - \mu_0^2 - 2(\mu_1 - \mu_0) x_i \Big) \right) \\ &= \exp\left(-\frac{n}{2\sigma^2} (\mu_1^2 - \mu_0^2) + \frac{\mu_1 - \mu_0}{\sigma^2} \left(\sum_{i=1}^n x_i \right) \right). \end{aligned}$$

If $\mu_1 > \mu_0$, then the right hand side is an increasing function of the sample mean $T(X) = \frac{1}{n} \sum_{i=1}^n X_i$, so that $\{ f(\cdot \mid \mu) \mid \mu \in \mathbb{R} \}$ possesses the MLR property in $\bar{X}_n$. It follows that a UMP test of H_0 against H_1 is the test δ^* with rejection region

$$R^* = \{ x \in E \mid T(x) > \lambda \}$$

for some $\lambda \in \mathbb{R}$ such that

$$\mathbb{P}_{\mu_0} [X \in R^*] = \alpha.$$

Since

$$\mathbb{P}_{\mu_0}\left[X \in R^*\right] = \mathbb{P}_{\mu_0}\left[\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > \frac{\lambda - \mu_0}{\sigma/\sqrt{n}}\right] = \mathbb{P}\left(Z > \frac{\lambda - \mu_0}{\sigma/\sqrt{n}}\right),$$

where $Z \sim \mathcal{N}[0,1]$, it follows that we must choose $\lambda \in \mathbb{R}$ so that it satisfies

$$1 - \Phi\left(\frac{\lambda - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha,$$

where Φ is the standard normal distribution function. Letting $z_{1-\alpha}$ be the critical value satisfying

$$\Phi(z_{1-\alpha}) = 1 - \alpha,$$

it follows that the test that rejects the null hypothesis if

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha},$$

whcih is just the standard z-test, is a uniformly most powerful test of H_0 against H_1 with significance level α.

2.1.3 Asymptotic Tests and Consistency

In many cases, we are interested in asymptotic tests, since the finite sample properties of many test statistics are unknown. In this case, an additional criterion called consistency comes in handy when evaluating hypothesis tests. We first present a general framework for asymptotic statistical problems that will arise repeatedly throughout the remainder of the text, and then introduce some concepts related to the consistency of asymptotic tests.

Formally, consider a statistical decision problem $((\Theta,\vartheta),A,L)$, where $A=\{0,1\}$ and (Θ,ϑ) is our parameter space. The point of divergence from the finite-sample problem comes in the form of the data, which is not given as a random variable but rather as a stochastic process. Specifically, we let $\{X_t\}_{t\in N_+}$ be our data, where each X_t takes values in the measurable space $(E,\mathcal{E})$. The T-observation sample $X^T=(X_1,\cdots,X_T)$ is then a random vector taking values in the product space $(E,\mathcal{E})^T$.

For any $T\in N_+$, we assume that there exists a transition probability kernel μ_T from (Θ,ϑ) into $(E,\mathcal{E})^T$ that represents the statistical model for the T-observation sample. That is, for any $\theta\in\Theta$, $\mu_T(\theta,\cdot)$ is the distribution of X^T. Usually, we assume that there exists some sort of $\mathcal{E}^T\bigotimes\mathcal{F}$-measurable density f_T of μ_T with respect to some σ-finite measure on $(E,\mathcal{E})^T$, as in the finite-sample case above. This will come in handy when we study maximum likelihood methods below. As usual, we define

$$\mathbb{E}_\theta\Big[f(X^T,\theta)\Big]:=\int_{E^T}f(x,\theta)\mu_T(\theta,dx)$$

for any $T\in N_+$, $\theta\in\Theta$, and $f\in(\mathcal{E}^T\bigotimes\vartheta)_+$.

A test of the null hypothesis $H_0:\theta\in\Theta_0$ against $H_1:\theta\notin\Theta_0$ is now given as a sequence $\delta=\{\delta_T\}_{T\in N_+}$ of measurable functions, where each δ_T maps E^T into $\{0,1\}$ and has corresponding rejection regeion $R_T\in\mathcal{E}^T$. For any $\alpha\in(0,1)$, we say that δ is an asymptotic level α test if

$$\limsup_{T\to\infty}\sup_{\theta\in\Theta_0}\mathbb{P}_\theta\Big[X^T\in R_T\Big]\leq\alpha.$$

In other words, in asymptotic level α tests, the probability of incurring a type I error does not exceed α in the limit.

An asymptotic level α test $\delta=\{\delta_T\}_{T\in N_+}$ is said to be consistent under the alternative $H_1:\theta=\theta_1\notin\Theta_0$ if the probability of rejection under θ_1 approaches 1 as the sample size goes to infinity. This is what we mean when we say that a test has power against a specific alternative; if the true parameter equals the value θ_1 in the alternative hypothesis, the test can tell that the null is false as we gather more and more sample information.

Formally, δ is a consistent test under the alternative $H_1:\theta=\theta_1\notin\Theta_0$ if

$$\liminf_{T\to\infty}\mathbb{P}_{\theta_1}\Big[X^T\in R_T\Big]=1.$$

A closely related concept is that of divergence in probability. A sequence of real random variables $\{Y_T\}_{T\in N_+}$ is said to diverge in probability to $+\infty$ if, for any $M > 0$,

$$\liminf_{T\to\infty} \mathbb{P}(Y_T > M) = 1.$$

Likewise, $\{Y_T\}_{T\in N_+}$ diverges in probability to $-\infty$ if, for any $M < 0$,

$$\liminf_{T\to\infty} \mathbb{P}(Y_T < M) = 1.$$

The following is a sufficient condition for divergence in probability:

Lemma 2.4 (Conditions for Divergence in Probability)
Let $\{Y_T\}_{T\in N_+}$ be a sequence of real random variables. The following hold true:

i) Let $\{a_T\}_{T\in N_+}$ a sequence of real numbers that increases to $+\infty$. If there exists some $\epsilon > 0$ such that

$$\lim_{T\to\infty} \mathbb{P}(Y_T > \epsilon) = 1,$$

then the sequence $\{a_T Y_T\}_{T\in N_+}$ diverges in probability to $+\infty$.

ii) Suppose there exist sequences $\{X_T\}_{T\in N_+}$ and $\{Z_T\}_{T\in N_+}$ of real random variables such that $X_T = O_p(1)$ as $T \to \infty$, Z_T diverges in probability to $+\infty$, and $Y_T = X_T + Z_T$. Then, Y_T diverges in probability to $+\infty$.

Proof) i) Fix any $M > 0$ and $\delta > 0$. By assumption, there exists an $N \in N_+$ such that

$$1 - \mathbb{P}(Y_T > \epsilon) < \delta \quad \text{and} \quad a_T > \frac{M}{\epsilon}$$

for any $T \geq N$. Since

$$\{\omega \in \Omega \mid Y_T(\omega) > \epsilon\} \subset \{\omega \in \Omega \mid a_T Y_T(\omega) > M\}$$

for any $T \geq N$, it follows that

$$\mathbb{P}(a_T Y_T > M) \geq \mathbb{P}(Y_T > \epsilon) > 1 - \delta,$$

or equivalently,

$$1 - \mathbb{P}(a_T Y_T > M) < \delta,$$

for any $T \geq N$. Such an N exists for any $\delta > 0$, so by definition

$$\lim_{T \to \infty} \mathbb{P}(a_T Y_T > M) = 1.$$

This in turn holds for any $M > 0$, so we have our desired result.

ii) Choose any $\epsilon > 0$ and $M > 0$. By the definition of boundedness in probability, there exists an $N_0 \in N_+$ and a $K > 0$ such that

$$\mathbb{P}(X_T < -K) \leq \mathbb{P}(|X_T| > K) < \frac{\epsilon}{2} \quad \text{for any } T \geq N_0.$$

Likewise, by the definition of divergence in probability to $+\infty$, there exists an $N \in N_+$ such that $N \geq N_0$ and

$$\mathbb{P}(Z_T \leq M + K) < \frac{\epsilon}{2} \quad \text{for any } T \geq N.$$

For any $T \in N_+$, $Y_T = X_T + Z_T \leq M$ and $X_T \geq -K$ imply

$$Z_T \leq M - X_T \leq M + K,$$

so that, for any $T \geq N$,

$$\mathbb{P}(Y_T \leq M) = \mathbb{P}(\{Y_T \leq M\} \cap \{X_T \geq -K\}) + \mathbb{P}(\{Y_T \leq M\} \cap \{X_T < -K\})$$

$$\leq \mathbb{P}(Z_T \leq M + K) + \mathbb{P}(X_T < -K) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Such an $N \in N_+$ exists for any $\epsilon > 0$, so

$$\lim_{T \to \infty} \mathbb{P}(Y_T \leq M) = 0,$$

or equivalently,

$$\lim_{T \to \infty} \mathbb{P}(Y_T > M) = 1.$$

This in turn holds for any $M > 0$, so Y_T diverges in probability to $+\infty$.

Q.E.D.

To see how divergence in probability is useful for determining the consistency of asymptotic tests, consider an asymptotic level α test $\delta = \{\delta_T\}_{T \in N_+}$ that is based on the sequence of real-valued test statistics $\{W_T(X^T)\}_{T \in N_+}$, as is usually the case. Assume furthermore that the T-observation rejection region is given as

$$R_T = \{x \in E^T \mid W_T(x) > c\}$$

for some critical value $c > 0$ that is invariant to the sample size. Then,

$$\mathbb{P}_\theta \left[X^T \in R_T \right] = \mathbb{P}_\theta \left[W_T(X^T) > c \right]$$

for any $\theta \in \Theta$ and $T \in N_+$, so that the test is consistent against the alternative $H_1 : \theta = \theta_1 \notin \Theta_0$ if

$$\liminf_{T \to \infty} \mathbb{P}_\theta \left[W_T(X^T) > c \right] = 1.$$

Since $\{W_T(X^T)\}_{T \in N_+}$ is a sequence of real random variables, a sufficient condition for consistency is for $\{W_T(X^T)\}_{T \in N_+}$ to diverge in probability to $+\infty$ when the true parameter is θ_1, that is,

$$\liminf_{T \to \infty} \mu_T \left[\theta_1, \{x \in E^T \mid W_T(x) > c\} \right] = 1.$$

Indeed, for all but a select few cases, this is how we verify the consistency of asymptotic tests of interest.

2.2 Confidence Sets

A statistical decision problem closely related to hypothesis testing is described by $((\Theta, \vartheta), A, L)$, where the action space A is given as $A = 2^{\Theta}$, that is, as the collection of subsets of the parameter space. In this case, we are estimating the unknown parameter of interest not through a single point in the parameter space, as in point estimation, but as a set of values in the parameter space. This type of problem is called set estimation, and letting X be the data taking values in a measurable space $(E, \mathcal{E})$ with statistical model μ, the decision rules $\delta : E \to A$ are called confidence sets.

Confidence sets can equivalently be seen, not as functions from E to 2^{Θ}, but as correspondences from E to Θ. As in the case of point estimation, we must impose measurability conditions on confidence sets for the treatments below to be sensible. To this end, we let (Θ, ρ) be a complete and separable metric space, and define the collection of confidence sets $\mathcal{D}$ as the collection of all correspondences $C : E \to \Theta$ such that:

1. Are non-empty and closed valued.

2. For any $\theta \in \Theta$,

$$C^{-1}(\theta) := \{x \in E \mid \theta \in C(x)\}$$

   is a non-empty and closed set.

The second condition is analogous to the definition of measurability for functions, and is required for the coverage probability of confidence sets to be well-defined.

Given the set estimation problem, let $C : E \to \Theta$ be a confidence set. Let $g(\theta)$ be our parameter of interest. The probability

$$\mathbb{P}_\theta \left[g(\theta) \in C(X) \right] = \mu(X \in C^{-1}(g(\theta)), \theta)$$

is called the coverage probability of $C(X)$ at the parameter $\theta \in \Theta$; note that this value is well-defined by our definition of measurability.

We say that the confidence set C for the parameter $g(\theta)$ has confidence level $1 - \alpha$ if

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta \left[g(\theta) \in C(X) \right] \geq 1 - \alpha.$$

In this case, $C(X)$ is called a confidence set with level $1 - \alpha$.

There is a close relationship between confidence sets of level $1 - \alpha$ and hypothesis tests with significance level α. Indeed, this is how most confidence sets are recovered; as inversions of test statistics.

Theorem 2.5 (Test Inversion Theorem)
Let (Θ, ρ) and (E, d) be complete and separable metric spaces with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. For any confidence set $C : E \to \Theta$ of level $1-\alpha$, define the correspondence $N : \Theta \to E$ as

$$N(\theta) = C^{-1}(\theta) = \{x \in E \mid \theta \in C(x)\}$$

for any $\theta \in \Theta$. Then, N is a measurable correspondence, that is, it is a non-empty closed valued correspondence such that $N^{-1}(x) = \{\theta \in \Theta \mid x \in N(\theta)\}$ is a closed subset of Θ. Furthermore, for any $\theta_0 \in \Theta$, $N(\theta_0)^c$ is the rejection region for a level α test of $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$.

Conversely, let there be a measurable correspondence $N : \Theta \to E$ such that, for any $\theta_0 \in \Theta$, $N(\theta_0)^c$ is the rejection region of a level α test of the null $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. Then, the correspondence $C : E \to \Theta$ defined as

$$C(x) = N^{-1}(x) = \{\theta \in \Theta \mid x \in N(\theta)\} \in \vartheta$$

for any $x \in E$ is a level $1-\alpha$ confidence set.

Proof) **Confidence Set $\to$ Hypothesis Test**

First, suppose that $C : E \to \Theta$ is a confidence set with level $1-\alpha$, and define the correspondence $N : \Theta \to E$ as

$$N(\theta) = C^{-1}(\theta) := \{x \in E \mid \theta \in C(x)\}$$

for any $\theta \in \Theta$. Since $C^{-1}(\theta)$ is non-empty and closed for any $\theta \in \Theta$ by the measurability of C, N is a non-empty and closed valued correspondence. Now fix $x \in E$ and consider the inverse image

$$N^{-1}(x) = \{\theta \in \Theta \mid x \in N(\theta)\}.$$

It is easy to see that $N^{-1}(x) = C(x)$. If $\theta \in C(x)$, then $x \in N(\theta)$ by definition, so that $\theta \in N^{-1}(x)$; on the other hand, if $\theta \in N^{-1}(x)$, then $x \in N(\theta)$ and thus $\theta \in C(x)$. Since $C(x)$ is non-empty and closed, so is $N^{-1}(x)$, which makes N a measurable correspondence.

Now choose any $\theta_0 \in \Theta$, and let δ_0 be a test for H_0 against H_1 with rejection region $N(\theta_0)^c$. We now have

$$\begin{aligned}\mathbb{P}_{\theta_0}\left[X \in N(\theta_0)^c\right] &= \mathbb{P}_{\theta_0}\left[X \in C^{-1}(\theta_0)^c\right] \\ &= 1 - \mathbb{P}_{\theta_0}\left[X \in C^{-1}(\theta_0)\right] \leq 1 - (1-\alpha) = \alpha,\end{aligned}$$

where the inequality follows becauase C is a level $1-\alpha$ confidence set. This shows us that $N(\theta_0)^c$ is the rejection region for a level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

Hypothesis Test →Confidence Set

Conversely, suppose that $N : \Theta \to E$ is a measurable correspondence such that, for any $\theta_0 \in \Theta$, $N(\theta_0)^c$ is the rejection region of a level α test of the null $H_0 : g(\theta) = g(\theta_0)$ against the alternative $H_1 : g(\theta) \neq g(\theta_0)$. Define $C : E \to \Theta$ as

$$C(x) = N^{-1}(x) := \{\theta \in \Theta \mid x \in N(\theta)\}$$

for any $x \in E$. That C is a measurable correspondence follows in much the same way as in the previous direction by noting that $C^{-1}(\theta) = N(\theta)$ for any $\theta \in \Theta$. Thus, we only show that C is a level $1-\alpha$ confidence set.

To this end, note that

$$\mathbb{P}_{\theta_0}[X \in N(\theta_0)^c] \leq \alpha$$

for any $\theta_0 \in \Theta$, since $N(\theta_0)^c$ is the rejection region of a level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Since

$$C^{-1}(\theta_0) = N(\theta_0),$$

we have

$$\begin{aligned}\mathbb{P}_{\theta_0}[\theta_0 \in C(X)] &= \mathbb{P}_{\theta_0}\left[X \in C^{-1}(\theta_0)\right] \\ &= \mathbb{P}_{\theta_0}[X \in N(\theta_0)] = 1 - \mathbb{P}_{\theta_0}[X \in N(\theta_0)^c] \geq 1-\alpha.\end{aligned}$$

This holds for any $\theta_0 \in \Theta$, so

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta[\theta_0 \in C(X)] \geq 1-\alpha,$$

and C becomes a level $1-\alpha$ confidence set.

Q.E.D.

This theorem demonstrates how the problem of constructing a confidence set is in many ways the dual of hypothesis testing. Given a confidence set $C : E \to \Theta$, the correspondence $N : \Theta \to E$ defined as $N(\theta) = C^{-1}(\theta)$ for any $\theta \in \Theta$ is thus referred to as the dual acceptance regions of C; likewise, given the acceptance regions N, the corresponding confidence set C is called the dual confidence set of N.

2.2.1 Evaluating Confidence Sets

So far, we have not yet talked about the third component of a set estimation problem, the loss function. Consider a set estimation problem characterized by $((\Theta, \vartheta), 2^{\Theta}, L)$ with data X taking values in the measurable space $(E, \mathcal{E})$ with statistical model μ. When Θ is a subspace of some euclidean space $\mathbb{R}^d$ and ϑ is the corresponding Borel σ-algebra, the loss function L penalizes the length, or more generally, volume, of confidence sets under consideration. Formally, we restrict our attention to confidence sets $C : E \to \Theta$ such that the mapping

$$(\theta, x) \mapsto I_{C^{-1}(\theta)}(x)$$

is measurable with respect to the product σ-algebra on $\Theta \times E$; denote the collection of such confidence sets by $\tilde{\mathcal{D}}$. This is a technical assumption meant to facilitate the upcoming analysis, and is usually satisfied by confidence intervals and confidence sets in euclidean space.

Letting λ be the Lebesgue measure on $\mathbb{R}^d$, the loss function $L : \vartheta \to \mathbb{R}_+$ is defined as

$$L(C) = \lambda(C)$$

for any $C \in \vartheta$. That is, the loss given a measurable set C in the parameter space is simply the volume of that set with respect to the Lebesgue measure. The risk associated with a confidence set $C \in \tilde{\mathcal{D}}$ then becomes the expected volume

$$R(C, \theta) = \mathbb{E}_\theta \left[L(C(X)) \right]$$

of C for any parameter $\theta \in \Theta$. Letting $N : \Theta \to E$ be the dual acceptance region of C, for any $\theta_0 \in \Theta$ let δ_{θ_0} be the test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with rejection region $N(\theta_0)^c$. Then, since we can write

$$L(C(X)) = \int_\Theta I_{N(\theta)}(X) d\theta = \int_\Theta \left(1 - \delta_\theta(X)\right) d\theta,$$

where the integral form follows from the assumed measurability for confidence sets in $\tilde{\mathcal{D}}$, an applicaiton of Fubini's theorem shows us that

$$\begin{aligned} R(C, \theta_0) = \mathbb{E}_{\theta_0} \left[L(C(X)) \right] &= \int_E \left(\int_\Theta \left(1 - \delta_\theta(x)\right) d\theta \right) \mu(\theta_0, dx) \\ &= \int_\Theta \left(1 - \beta_\theta(\theta_0)\right) d\theta \end{aligned}$$

for any $\theta_0 \in \Theta$, where $\beta_\theta : \Theta \to [0, 1]$ is the power function corresponding to the test δ_θ. In other words, there is a close relationship between the risk associated with a confidence set and the power function of the hypothesis tests that are dual to it. We can exploit this relationship to find optimal confidence sets:

Theorem 2.6 (Pratt's Theorem)
Let (Θ, ρ) and (E, d) be complete and separable metric spaces with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. Assume in addition that Θ is a subset of some euclidean space $\mathbb{R}^d$. Let $C^* \in \tilde{\mathcal{D}}$ be a level $1-\alpha$ confidence set with dual acceptance regions $N^* : \Theta \to E$. Fix some $\theta_1 \in \Theta$.

Suppose that, for any $\theta_0 \neq \theta_1$, the test $\delta^*_{\theta_0}$ with rejection region $N^*(\theta_0)^c$ is a UMP level α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, and that the risk $R(C^*, \theta_1)$ is finite. Then, C^* achieves the minimum expected volume under θ_1 among confidence sets with level $1-\alpha$, that is, for any confidence set $C \in \tilde{\mathcal{D}}$ with level $1-\alpha$,

$$R(C^*, \theta_1) = \mathbb{E}_{\theta_1}\left[L(C^*(X))\right] \leq \mathbb{E}_{\theta_1}\left[L(C(X))\right] = R(C, \theta_1).$$

Proof) Let C be a level $1-\alpha$ confidence set in $\tilde{\mathcal{D}}$, and let $N : \Theta \to E$ be its dual acceptance region. For any $\theta_0 \in \Theta$, define δ_{θ_0} as the test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with rejection region $N(\theta_0)^c$. By the test inversion theorem, each δ_{θ_0} has level α. For any $\theta_0 \neq \theta_1$, by assumption,

$$\beta_{\theta_0}(\theta_1) \leq \beta^*_{\theta_0}(\theta_1),$$

where β_{θ_0} and $\beta^*_{\theta_0}$ are the power functions of the tests δ_{θ_0} and $\delta^*_{\theta_0}$. By the remark preceding the statement of the theorem,

$$R(C^*, \theta_1) = \int_\Theta \left(1 - \beta^*_\theta(\theta_1)\right) d\theta$$

and

$$R(C, \theta_1) = \int_\Theta \left(1 - \beta_\theta(\theta_1)\right) d\theta.$$

We showed above that $\beta^*_\theta(\theta_1) \geq \beta_\theta(\theta_1)$ almost for Lebesgue-almost every $\theta \in \Theta$.

If $R(C, \theta_1) = \infty$, then the theorem follows trivially. Suppose $R(C, \theta_1) < +\infty$. Since $R(C^*, \theta_1) < +\infty$ as well by assumption, we have

$$R(C, \theta_1) - R(C^*, \theta_1) = \int_\Theta \left(\beta^*_\theta(\theta_1) - \beta_\theta(\theta_1)\right) d\theta \geq 0,$$

which shows us that

$$R(C^*, \theta_1) \leq R(C, \theta_1).$$

Q.E.D.

Chapter 3

The Asymptotics of Extremum Estimators

3.1 Existence of Extremum Estimators

We consider the following basic setting. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be an n-dimensional time series that represents the data available to us; we understand Y_t to include both the dependent variable of interest alongside any exogenous variables or instruments. Let the parameter of interest be θ, which takes values in an arbitrary metric space (Θ, d). This allows us to not only take real valued vectors as our parameters, but also (if needed) functions and more abstract mathematical objects as parameters as well. Our goal is to estimate θ; one general way of doing so is to consider extremum estimators, or estimators that are found as solutions to optimization problems.

The idea is to find, as the estimator of θ, a value that minimizes some loss function. Generally, we consider objective functions of the form $Q : \Omega \times \Theta \to \mathbb{R}$, where $Q(\cdot, \theta)$ is dependent on the data in some arbitrary manner, with the only restriction being that $Q(\cdot, \theta)$ must be a real random variable. The extremum estimator is then found as the function $\hat{\theta} : \Omega \to \Theta$ such that

$$\hat{\theta}(\omega) = \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\omega, \theta)$$

for any $\omega \in \Omega$.

Two problems immediately present themselves. First, there may be $\omega \in \Omega$ for which the maximizer does not exist; in this case, we want $\hat{\theta}$ to exist almost surely at least asymptotically as the sample size $T \to \infty$. Second, even if $\hat{\theta}$ exists on Ω in a sufficiently tractable manner, it may not be measurable. There are theorems that show how requiring Q to be continuous in θ allows us to find a measurable extremum estimator, but here we take a different approach.

In most applications, we search for a sequence $\{\hat{\theta}_T\}_{T\in N_+}$ of estimators, where each $\hat{\theta}_T$ depends only on the data up to the Tth observation. In the case of time series, where using only the information available up to a certain period of time is important, this becomes an important consideration. Therefore, instead of specifying the objective function in an abstract manner as above, we could think of objective functions of the form $Q_T : \mathbb{R}^{nT} \times \Theta \to \mathbb{R}$. In this case, letting

$\hat{\theta}_T(y)$ be the solution to the problem

$$\max_{\theta \in \Theta} \quad Q_T(y, \theta)$$

for any $y \in \mathbb{R}^{nT}$, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ becomes the extremum estimator such that $\hat{\theta}_T(Y_1(\omega), \cdots, Y_T(\omega))$ solves the maximization problem

$$\max_{\theta \in \Theta} \quad Q_T(Y_1(\omega), \cdots, Y_T(\omega), \theta)$$

for any $\omega \in \Omega$. This formulation makes clear the dependence of the extremum estimator on the first T sample observations $Y_1, \cdots, Y_T$, and reduces the problem of the measurability of the extremum estimator to establishing the measurability of the function $\hat{\theta}_T$. The following theorem establishes sufficient conditions for this to be the case:

Theorem 3.1 (Existence and Measurability of Extremum Estimators)
Let $\{Y_t\}_{t \in \mathbb{Z}}$ be a sequence of random variables taking values in the measurable space $(E, \mathcal{E})$, where E is a topological space and $\mathcal{E}$ the corresponding Borel σ-algebra. Furthermore, let (Θ, d) be a compact metric space, and $Q_T : E^T \times \Theta \to \mathbb{R}$ a continuous function on $E^T \times \Theta$. Then, there exists a Borel measurable function $\hat{\theta}_T : E^T \to \Theta$ such that

$$Q_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) = \max_{\theta \in \Theta} Q_T(Y_1, \cdots, Y_T \mid \theta).$$

If, in addition, Θ is a non-empty convex subset of some euclidean space $\mathbb{R}^m$ and $Q_T(y_1, \cdots, y_T \mid \cdot)$ is a strictly concave function on Θ for any $(y_1, \cdots, y_T) \in E^T$, then $\hat{\theta}_T$ is a continuous function on E^T.

Proof) The result follows from a simple application of Berge's maximum theorem. Define the correspondence $\Gamma : E^T \to \Theta$ as $\Gamma(y) = \Theta$ for any $y \in E^T$. Then, Γ is a non-empty, compact valued and continuous correspondence, and we can now consider the maximization problem

$$\max_{\theta \in \Gamma(y)} \quad Q_T(y, \theta)$$

for any $y \in E^T$. Defining the value function $v : E^T \to [-\infty, +\infty]$ of this problem as

$$v(y) = \sup_{\theta \in \Gamma(y)} Q_T(y, \theta)$$

and the correspondence of maximizers $G : E^T \to \Theta$ as

$$G(y) = \{\theta \in \Gamma(y) \mid Q_T(y, \theta) = v(y)\}$$

for any $y \in E^T$, Berge's maximum theorem now tells us that v is real valued and

continuous, and that G is a non-empty compact valued and upper hemicontinuous correspondence. Since Θ is a complete and separable metric space (all compact spaces are complete and separable) and G is a non-empty closed valued and upper hemicontinuous correpsondence, the measurable selection theorem now shows us that there exists a Borel measurable function $\hat{\theta}_T : E^T \to \Theta$ such that $\hat{\theta}_T(y) \in G(y)$ for any $y \in E^T$. Therefore, for any $y \in E^T$,

$$Q_T(y, \hat{\theta}_T(y)) = \max_{\theta \in \Theta} Q_T(y, \theta),$$

and since the random vector $(Y_1, \cdots, Y_T)$ takes values in E^T,

$$Q_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) = \max_{\theta \in \Theta} Q_T(Y_1, \cdots, Y_T \mid \theta)$$

everywhere on the sample space Ω.

Suppose in addition that Θ is a non-empty convex subset of $\mathbb{R}^m$ and that the mapping $\theta \mapsto Q_T(y \mid \theta)$ is strictly concave for any $y \in E^T$. Then, the correspondence G of maximizers is single-valued; to see this, suppose $G(y)$ is not single-valued for some $y \in E^T$, so that there exist distinct $\theta_1, \theta_2 \in G(y)$ such that $\theta_1 \neq \theta_2$. Then, choosing $\lambda = \frac{1}{2} \in (0,1)$, by strict concavity

$$\lambda Q_T(y \mid \theta_1) + (1-\lambda) Q_T(y \mid \theta_2) < Q_T(y \mid \lambda\theta_1 + (1-\lambda)\theta_2),$$

where $\lambda\theta_1 + (1-\lambda)\theta_2 \in \Theta$ by the convexity of Θ. Furthermore, since $\theta_1, \theta_2 \in G(y)$,

$$Q_T(y \mid \theta_1) = Q_T(y \mid \theta_2) = v(y),$$

so we have

$$v(y) < Q_T(y \mid \lambda\theta_1 + (1-\lambda)\theta_2).$$

This contradicts the definition of the value function, so it must be the case that $G(y)$ contains at most one element; G is non-empty valued, so $G(y)$ is a singleton, and since this holds for any $y \in E^T$, G is single-valued. We saw above from Berge's theorem that G is upper hemicontinuous. As such, letting $\hat{\theta}_T : E^T \to \Theta$ be chosen so that $G(y) = \{\hat{\theta}_T(y)\}$ for any $y \in E^T$, $\hat{\theta}_T$ is a continuous function on E^T that satisfies

$$Q_T(y \mid \hat{\theta}_T(y)) = v(y)$$

for any $y \in E^T$.

Q.E.D.

The measurability of the extremum estimator follows because $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the composition of two measurable functions and measurability is preserved across compositions. The measurabiltiy of the extremum estimator is generally not a concern, since, as noted in Newey and Macfadden (1994), if the maximizer is non-measurable then we can usually take the underlying probability measure to be an outer measure, which can assign values to any subset of the sample space and therefore allows us to circumvent the measurability concern. However, in some cases the properties of the underyling probability space is very important, for instance when we are dealing with time series data and want to interpret σ-algebras on Ω as representing the flow of information over time. In this case, ensuring that the extremum estimator is measurable with respect to a certain σ-algebra proves essential, and it is in these contexts that the theorem above shines.

Note also that the above result ensures that the extremum estimator $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is a global maximizer of the objective function everywhere on the sample space, which is a very storng result that follows from the very strong assumptions that we imposed on the objective function (continuity) and the parameter space (Θ). It is possible to relax these assumptions and obtain more general results; for instance, as in the general case, we can take the objective function to be a measurable function on $\Omega \times \Theta$ instead of a continuous function on $E^T \times \Theta$, or we can impose conditions other than compactness on Θ that ensures the (almost sure) existence of a solution to the maximization problem. For the moment, however, we work under these generally innocuous assumptions.

3.2 Uniform Convergence in Probability

Newey and Macfadden (1994) furnish sufficient conditions for any extremum estimator to be consistent and asymptotic normality for the parameter of interest. One of the core conditions is uniform convergence in probability. The exposition in this section is based on Newey (1991).

Formally, let (Θ, ϑ) some measurable space. We consider sequences of stochastic processes $\{Q_T(\theta)\}_{\theta\in\Theta}$ with index set Θ and common state space $(E,\mathcal{E})$.

Note that we do not require Q to be a random function taking values in the space of real functions on Θ, since this is stronger than necessary. In general, there does not exist a random function corresponding to a stochastic process, and vice versa[1]. In the special case that Θ is a compact topological space, $(E,\mathcal{E})$ is a euclidean space and $\{Q_T(\theta)\}_{\theta\in\Theta}$ has continuous paths, there exists a unique random function Q^T, taking values in the continuous function space $C(\Theta, E)$ with the supremum norm, corresponding to $\{Q_T(\theta)\}_{\theta\in\Theta}$.

Let (E,d) be a metric space and $\mathcal{E}$ the associated Borel σ-algebra. In this case, a sequence $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ of real valued stochastic processes in $(E,\mathcal{E})$ is said to converge uniformly in probability to some stochastic process $\{Q(\theta)\}_{\theta\in\Theta}$, also in $(E,\mathcal{E})$, if

1) For any $T \in N_+$,

$$\sup_{\theta\in\Theta} d(Q_T(\theta), Q(\theta))$$

is a well-defined non-negative random variable, and

2) The supremum converges in probability to 0, that is,

$$\sup_{\theta\in\Theta} d(Q_T(\theta), Q(\theta)) \xrightarrow{p} 0.$$

We now study the two conditions above in turn.

[1] A stochastic process corresponding to a random function Q taking values in the space $\mathcal{F}_{\Theta\to E}$ is defined as the process $\{q(\theta)\}_{\theta\in\Theta}$ with state spce $(E,\mathcal{E})$ such that $q(\theta)$ is the mapping $\omega \mapsto [Q(\omega)](\theta)$.

Conversely, a random function corresponding to a stochastic process $\{q(\theta)\}_{\theta\in\Theta}$ with state space $(E,\mathcal{E})$ is the mapping $Q : \Omega \to \mathcal{F}_{\Theta\to E}$ such that $Q(\omega)$ is the mapping $\theta \mapsto [q(\theta)](\omega)$.

3.2.1 The Measurability Problem

The measurability condition is cumbersome to verify in every context, since the supremum can be taken over an uncountable set. Fortunately, there exist simple sufficient conditions that ensure the measurability of the supremum above.

Lemma 3.2 Let (Θ, τ) be a topological space and (E, d) a metric space, with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. Consider two stochastic processes $\{Q(\theta)\}_{\theta\in\Theta}$ and $\{S(\theta)\}_{\theta\in\Theta}$ with state space $(E, \mathcal{E})$. If Θ and E are separable, that is, if it they have countable dense subsets, and $\{Q(\theta)\}_{\theta\in\Theta}$ and $\{S(\theta)\}_{\theta\in\Theta}$ have continuous paths, then

$$\sup_{\theta\in\Theta} d(Q(\theta), S(\theta))$$

is a non-negative random variable.

Proof) First, the separability of (E, d) tells us that, for any $\theta \in \Theta$, the distance

$$X(\theta) = d(Q(\theta), S(\theta))$$

is a non-negative random variable. Furthermore, given any $\omega \in \Omega$, because the mappings $d : E^2 \to \mathbb{R}$ and $\theta \mapsto ([Q(\theta)](\omega), [S(\theta)](\omega))$ are continuous (the processes Q and S have continuous paths), the mapping $\theta \mapsto [X(\theta)](\omega)$ is continuous. In other words, $\{X(\theta)\}_{\theta\in\Theta}$ is a stochastic process taking values in $[0, +\infty]$ with continuous paths.

Now define the mapping $Y : \Omega \to [0, +\infty]$ as

$$Y = \sup_{\theta\in\Theta} d(Q(\theta), S(\theta)) = \sup_{\theta\in\Theta} X(\theta).$$

Let Θ_0 be a countable dense subset of Θ. By definition,

$$Y_0 = \sup_{\theta\in\Theta_0} X(\theta) \leq \sup_{\theta\in\Theta} X(\theta) = Y.$$

Suppose, for some $\omega \in \Omega$, that

$$Y_0(\omega) < Y(\omega).$$

Then, there exists a $\theta^* \in \Theta \setminus \Theta_0$ such that

$$Y_0(\omega) < [X(\theta^*)](\omega).$$

Let $\epsilon = [X(\theta^*)](\omega) - Y_0(\omega) > 0$. By the continuity of the paths of $\{X(\theta)\}_{\theta\in\Theta}$, there

exists a neighborhood U around θ^* such that

$$|[X(\theta^*)](\omega) - [X(\theta)](\omega)| < \epsilon,$$

and in particular

$$Y_0(\omega) < [X(\theta)](\omega)$$

for any $\theta \in U$. Since Θ_0 is dense in Θ, there must exist a $\theta' \in U \cap E$; by implication,

$$Y_0(\omega) = \sup_{\theta \in \Theta_0} [X(\theta)](\omega) < [X(\theta')](\omega).$$

This contradicts the definition of the supremum as an upper bound, so it must be the case that $Y_0 = Y$ everywhere on Ω. Finally, since Θ_0 is countable, and measurability is preserved across countable supremums, Y_0 and as such Y is a random variable.

Q.E.D.

In light of the preceding lemma, we can now reformulate uniform convergence in probability in terms of stochastic processes on separable metric spaces with continuous paths on separable topological spaces. Formally, letting (Θ, τ) be a separable topological space and (E, d) a separable metric space, we say that a sequence $\{\{Q_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ of stochastic processes taking values in E with continuous paths converges uniformly in probability to some stochastic process $\{Q(\theta)\}_{\theta \in \Theta}$, also in E with continuous paths, if

$$\sup_{\theta \in \Theta} d(Q_T(\theta), Q(\theta)) \xrightarrow{p} 0.$$

3.2.2 Characterization of Uniform Convergence in Probability

Now that we have dealt with the measurability problem, we want to search for necessary and sufficient conditions for uniform convergence in probability. Let (Θ, ϑ) be a measurable space and (E, d) a metric space. We abstract from measurability issues at this time.

Recall that, given a sequence $\{f_n\}_{n\in N_+}$ on the continuous function space $\mathcal{C}(E, F)$, where (E, ρ) is a compact metric space and $F = \mathbb{R}^n$ or $\mathbb{C}$, $\{f_n\}_{n\in N_+}$ has a uniformly convergent subsequence if and only if $\{f_n\}_{n\in N_+}$ is pointwise bounded and uniformly equicontinuous on E; this is the content of the Arzela-Ascoli theorem. Here, we say $\{f_n\}_{n\in N_+}$ is uniformly equicontinuous if, for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\sup_{n\in N_+} \sup_{\rho(x,y)<\delta} |f_n(x) - f_n(y)| < \epsilon.$$

Similarly, it turns out that a sequence $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ of stochastic processes to convergence uniformly in probability if and only if it converges pointwise and is stochastically equicontinuous, given that Θ is compact. We say $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinous if:

> For any $\theta \in \Theta$ and $\epsilon, \eta > 0$, there exists a constant $N(\epsilon, \eta)$, a sequence of non-negative random variables $\{\Delta_T(\epsilon, \eta)\}_{T\in N_+}$, and a neighborhood $U(\theta, \epsilon, \eta)$ of θ such that, for any $T \geq N(\epsilon, \eta)$
>
> $$\mathbb{P}\left(\Delta_T(\epsilon, \eta) > \epsilon\right) < \eta$$
>
> and
>
> $$\sup_{\theta'\in U(\theta,\epsilon,\eta)} d(Q_T(\theta), Q_T(\theta')) < \Delta_T(\epsilon, \eta).$$

The formal characterization of uniform convergence in probability is given below. It makes clever use of the assumed compactness of the parameter space and reduces the problem of uniform convergence over a (possibly) uncountable parameter space to that of pointwise convergence over a finite number of parameters.

Theorem 3.3 (Characterization of Uniform Convergence in Probability) Let (Θ, τ) be a topological space and (E, d) a metric space, with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. Let $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ be a sequence of stochastic processes taking values in E, and $\{\bar{Q}_T\}_{T\in N_+}$ a sequence of (non-random) functions from Θ to E. Assume the following:

i) For any $T \in N_+$,

$$\sup_{\theta\in\Theta} d(Q_T(\theta), \bar{Q}_T(\theta))$$

is a non-negative random variable.

ii) Θ is compact.

iii) $\{\bar{Q}_T\}_{T\in N_+}$ is equicontinuous on Θ.

Then,

$$\sup_{\theta\in\Theta} d(Q_T(\theta), \bar{Q}_T(\theta)) \xrightarrow{p} 0$$

if and only if $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous and

$$d(Q_T(\theta), \bar{Q}_T(\theta)) \xrightarrow{p} 0$$

pointwise for any $\theta \in \Theta$.

Proof) Define $R_T(\theta) = d(Q_T(\theta), \bar{Q}_T(\theta))$ for any $T \in N_+$ and $\theta \in \Theta$. We first prove a preliminary result, namely that

> $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous if and only if $\{\{R_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous.

Proof of Preliminary Result

First, suppose that $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous. Then, by definition, for any $\epsilon, \eta > 0$ and $\theta \in \Theta$ there exists a constant $N(\epsilon/2, \eta)$, a sequence of non-negative random variable $\{\Delta_T(\epsilon/2, \eta)\}_{T\in N_+}$, and a neighborhood $U(\theta, \epsilon/2, \eta)$ of θ such that, for any $T \geq N(\epsilon, \eta)$

$$\mathbb{P}\left(\Delta_T(\epsilon/2, \eta) > \frac{\epsilon}{2}\right) < \eta$$

and

$$\sup_{\theta'\in U(\theta,\epsilon/2,\eta)} d(Q_T(\theta), Q_T(\theta')) < \Delta_T(\epsilon/2, \eta).$$

By the equicontinuity of $\{\bar{Q}_T\}_{T\in N_+}$, for any $\theta\in\Theta$ there exists a neighborhood $V(\theta,\epsilon/2,\eta)\subset U(\theta,\epsilon/2,\eta)$ of θ such that

$$\sup_{T\in N_+}\sup_{\theta'\in V(\theta,\epsilon/2,\eta)} d(\bar{Q}_T(\theta),\bar{Q}_T(\theta'))<\frac{\epsilon}{2}.$$

Now fix $T\geq N(\epsilon/2,\eta)$. Since

$$\begin{aligned}\left|R_T(\theta)-R_T(\theta')\right| &= \left|d(Q_T(\theta),\bar{Q}_T(\theta))-d(Q_T(\theta'),\bar{Q}_T(\theta'))\right| \\ &\leq d(Q_T(\theta),Q_T(\theta'))+d(\bar{Q}_T(\theta),\bar{Q}_T(\theta'))\end{aligned}$$

for any $\theta,\theta'\in\Theta$, we can see that

$$\begin{aligned}\sup_{\theta'\in V(\theta,\epsilon/2,\eta)}\left|R_T(\theta)-R_T(\theta')\right| &\leq \sup_{\theta'\in V(\theta,\epsilon/2,\eta)} d(\bar{Q}_T(\theta),\bar{Q}_T(\theta'))+\sup_{\theta'\in U(\theta,\epsilon/2,\eta)} d(Q_T(\theta),Q_T(\theta')) \\ &< \Delta_T(\epsilon/2,\eta)+\frac{\epsilon}{2}.\end{aligned}$$

In addition,

$$\mathbb{P}\left(\Delta_T(\epsilon/2,\eta)+\frac{\epsilon}{2}>\epsilon\right)=\mathbb{P}\left(\Delta_T(\epsilon/2,\eta)>\epsilon/2\right)<\eta.$$

As such, $\{\{R_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ can be seen to be stochastically equicontinuous by setting

$$\begin{aligned}\Delta'_T(\epsilon,\eta) &= \Delta_T(\epsilon/2,\eta)+\frac{\epsilon}{2}, \\ N'(\epsilon,\eta) &= N(\epsilon/2,\eta) \\ U'(\theta,\epsilon,\eta) &= V(\theta,\epsilon,\eta)\end{aligned}$$

for any $\epsilon,\eta>0$ and $\theta\in\Theta$.

Conversely, suppose $\{\{R_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous. Then, for any $\epsilon,\eta>0$ and $\theta\in\Theta$, there exists a constant $N(\epsilon/2,\eta)\in N_+$, a sequence of non-negative random variables $\{\Delta_T(\epsilon/2,\eta)\}_{T\in N_+}$, and a neighborhood $U(\theta,\epsilon/2,\eta)$ around θ such that, for any $T\geq N(\epsilon/2,\eta)$,

$$\mathbb{P}\left(\Delta_T(\epsilon/2,\eta)>\frac{\epsilon}{2}\right)<\eta$$

and

$$\sup_{\theta'\in U(\theta,\epsilon/2,\eta)}\left|R_T(\theta)-R_T(\theta')\right|<\Delta_T(\epsilon/2,\eta).$$

The equicontinuity of $\{\bar{Q}_T\}_{T\in N_+}$ again tells us that there exists a neighborhood $V(\theta,\epsilon,\eta)\subset$

$U(\theta, \epsilon/2, \eta)$ of θ such that

$$\sup_{T \in N_+} \sup_{\theta' \in V(\theta,\epsilon,\eta)} d(\bar{Q}_T(\theta), \bar{Q}_T(\theta')) < \frac{\epsilon}{2}.$$

As above, fix $T \geq N(\epsilon/2, \eta)$. Then,

$$\begin{aligned} d(Q_T(\theta), Q_T(\theta')) &\leq \left| d(Q_T(\theta), \bar{Q}_T(\theta)) - d(\bar{Q}_T(\theta'), Q_T(\theta')) \right| + d(\bar{Q}_T(\theta), \bar{Q}_T(\theta')) \\ &= \left| R_T(\theta) - R_T(\theta') \right| + d(\bar{Q}_T(\theta), \bar{Q}_T(\theta')) \end{aligned}$$

for any $\theta, \theta' \in \Theta$, so that

$$\begin{aligned} \sup_{\theta' \in V(\theta,\epsilon,\eta)} d(Q_T(\theta), Q_T(\theta')) &\leq \sup_{\theta' \in U(\theta,\epsilon/2,\eta)} \left| R_T(\theta) - R_T(\theta') \right| + \sup_{\theta' \in V(\theta,\epsilon,\eta)} d(\bar{Q}_T(\theta), \bar{Q}_T(\theta')) \\ &\leq \Delta_T(\epsilon/2, \eta) + \frac{\epsilon}{2}. \end{aligned}$$

Since

$$\mathbb{P}\left(\Delta_T(\epsilon/2, \eta) + \frac{\epsilon}{2} > \epsilon \right) = \mathbb{P}\left(\Delta_T(\epsilon/2, \eta) > \epsilon/2 \right) < \eta,$$

as above $\{\{Q_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ is stochastically equicontinuous under the values

$$\begin{aligned} \Delta'_T(\epsilon, \eta) &= \Delta_T(\epsilon/2, \eta) + \frac{\epsilon}{2}, \\ N'(\epsilon, \eta) &= N(\epsilon/2, \eta) \\ U'(\theta, \epsilon, \eta) &= V(\theta, \epsilon, \eta) \end{aligned}$$

for any $\epsilon, \eta > 0$ and $\theta \in \Theta$.

Equicontinuity & Pointwise Convergence $\to$ Uniform Convergence

Suppose $\{\{Q_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ is stochastically equicontinuous and

$$d(Q_T(\theta), \bar{Q}_T(\theta)) \xrightarrow{p} 0$$

as $T \to \infty$ for any $\theta \in \Theta$. Choose any $\delta > 0$; we want to show that

$$\lim_{T \to \infty} \mathbb{P}\left(\sup_{\theta \in \Theta} d(Q_T(\theta), \bar{Q}_T(\theta)) > \delta \right) = 0.$$

By what we just proved above, $\{\{R_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ is also stohastically equicontinuous, so that, for any $\eta > 0$ and $\theta \in \Theta$, there exists a constant $N(\delta/2, \eta) \in N_+$, a sequence of

non-negative random variables $\{\Delta_T(\delta/2,\eta)\}_{T\in N_+}$ and a neighborhood $U(\theta,\delta/2,\eta)$ of θ such that, for any $T \geq N(\delta/2,\eta)$,

$$\mathbb{P}\left(\Delta_T(\delta/2,\eta) > \frac{\delta}{2}\right) < \eta$$

and

$$\sup_{\theta'\in U(\theta,\delta/2,\eta)} \left|R_T(\theta) - R_T(\theta')\right| < \Delta_T(\delta/2,\eta).$$

Consider the open cover $\{U(\theta,\delta/2,\eta)\}_{\theta\in\Theta}$ of Θ. By the compactness of Θ, there exists a finite subcover

$$\{U(\theta_1,\delta/2,\eta),\cdots,U(\theta_J,\delta/2,\eta)\}$$

of Θ. If

$$R_T(\theta_j) \leq \frac{\delta}{2}$$

for any $1 \leq j \leq J$, then for any $\theta \in \Theta$, since $\theta \in U(\theta_j,\delta/2,\eta)$ for some $1 \leq j \leq J$ by the fact that the above finite collection covers Θ, and we have

$$\begin{aligned} R_T(\theta) &\leq R_T(\theta_j) + |R_T(\theta) - R_T(\theta_j)| \\ &\leq R_T(\theta_j) + \sup_{\theta'\in U(\theta,\delta/2,\eta)} \left|R_T(\theta) - R_T(\theta')\right| \\ &\leq \frac{\delta}{2} + \Delta_T(\delta/2,\eta). \end{aligned}$$

It follows that

$$\sup_{\theta\in\Theta} R_T(\theta) \leq \frac{\delta}{2} + \Delta_T(\delta/2,\eta).$$

If, in addition, $\Delta_T(\delta/2,\eta) \leq \frac{\delta}{2}$, we can see that

$$\sup_{\theta\in\Theta} R_T(\theta) \leq \delta.$$

Therefore,

$$\left[\bigcap_{j=1}^{J}\left\{R_T(\theta_j) \leq \frac{\delta}{2}\right\}\right] \cap \left\{\Delta_T(\delta/2,\eta) \leq \frac{\delta}{2}\right\} \subset \{\sup_{\theta\in\Theta} R_T(\theta) \leq \delta\}.$$

By implication,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} R_T(\theta) > \delta\right) \leq \sum_{j=1}^{J} \mathbb{P}\left(R_T(\theta_j) > \frac{\delta}{2}\right) + \mathbb{P}\left(\Delta_T(\delta/2,\eta) > \frac{\delta}{2}\right).$$

For any $T \geq N(\delta/2, \eta)$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} R_T(\theta) > \delta\right) \leq \sum_{j=1}^{J} \mathbb{P}\left(R_T(\theta_j) > \frac{\delta}{2}\right) + \eta,$$

and $R_T(\theta_j) \xrightarrow{p} 0$ for any $1 \leq j \leq J$, so taking $T \to \infty$ on both sides shows us that

$$\limsup_{T\to\infty} \mathbb{P}\left(\sup_{\theta\in\Theta} R_T(\theta) > \delta\right) \leq \eta.$$

Finally, this holds for any $\eta > 0$, so it follows that

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{\theta\in\Theta} R_T(\theta) > \delta\right) = 0,$$

which is exactly the desired result.

Uniform Convergence → Equicontinuity & Pointwise Convergence

Now suppose that

$$\sup_{\theta\in\Theta} R_T(\theta) \xrightarrow{p} 0.$$

Pointwise convergence follows trivially, since $R_T(\theta') \leq \sup_{\theta\in\Theta} R_T(\theta)$ for any $\theta' \in \Theta$. To show equicontinuity, it suffices, in light of the preliminary result, to show that $\{\{R_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous. Choose any $\epsilon, \eta > 0$ and $\theta \in \Theta$. By the definition of convergence in probability, there exists an $N(\epsilon, \eta) \in N_+$ such that

$$\mathbb{P}\left(\sup_{\theta\in\Theta} R_T(\theta) > 2\epsilon\right) < \eta.$$

for any $T \geq N(\epsilon, \eta)$. Define

$$\Delta_T(\epsilon, \eta) = 2 \cdot \sup_{\theta\in\Theta} R_T(\theta)$$

$$U(\theta, \epsilon, \eta) = \Theta.$$

Then, for any $T \geq N(\epsilon, \eta)$, we have

$$\mathbb{P}\left(\Delta_T(\epsilon, \eta) > \epsilon\right) < \eta,$$

and

$$\sup_{\theta' \in U(\theta,\epsilon,\eta)} \left| R_T(\theta) - R_T(\theta') \right| \leq 2 \cdot \sup_{\theta' \in \Theta} R_T(\theta') \leq \Delta_T(\epsilon, \eta).$$

This holds for any $\epsilon, \eta > 0$ and $\theta \in \Theta$, so by definition $\{\{R_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ is stochastically equicontinuous.

Q.E.D.

In light of the lemma on measurability, we immediately have the following corollary:

Corollary to Theorem 3.3 Let (Θ, τ) be a separable topological space and (E, d) a separable metric space, with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. Let $\{\{Q_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ be a sequence of stochastic processes with continuous paths taking values in E, and $\{\bar{Q}_T\}_{T \in N_+}$ a sequence of (non-random) functions from Θ to E. Assume the following:

i) Θ is compact.

ii) $\{\bar{Q}_T\}_{T \in N_+}$ is equicontinuous on Θ.

Then,

$$\sup_{\theta \in \Theta} d(Q_T(\theta), \bar{Q}_T(\theta)) \xrightarrow{p} 0$$

if and only if $\{\{Q_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ is stochastically equicontinuous and

$$d(Q_T(\theta), \bar{Q}_T(\theta)) \xrightarrow{p} 0$$

pointwise for any $\theta \in \Theta$.

Proof) Since $\{\bar{Q}_T\}_{T \in N_+}$ is equicontinuous on Θ, it follows trivially that each $\bar{Q}_T$ is continuous on Θ, so that we can treat $\{\{\bar{Q}_T(\theta)\}_{\theta \in \Theta}\}_{T \in N_+}$ as a sequence of degenerate stochastic processes with continuous paths. Under the separability of Θ and E, along with the continuity of the paths of the stochastic processes in question, we noted in the lemma above that

$$\sup_{\theta \in \Theta} d(Q_T(\theta), \bar{Q}_T(\theta))$$

is a non-negative random variable for any $T \in N_+$. The rest of the corollary then follows from the preceding characterization theorem.

Q.E.D.

3.2.3 Uniform Weak Laws of Large Numbers

We can apply the characterization theorem above to obtain sufficient conditions for uniform weak law of large numbers. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a sequence of random variables taking values in $(E,\mathcal{E})$ and $\{g_t\}_{t\in N_+}$ a sequence of real matrix-valued functions on $E\times\Theta$. We say that $\{g_t\}_{t\in N_+}$ satisfies a uniform WLLN if

$$\sup_{\theta\in\Theta}\left\|\frac{1}{T}\sum_{t=1}^{T}g_t(Y_t,\theta)-\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[g_t(Y_t,\theta)\right]\right\|\xrightarrow{p}0,$$

where the expectations are well-defined and the term on the left hand side is a non-negative random variable. We furnish sufficient conditions for a more general version of the uniform WLLN to hold below, one that is more suited for non-i.i.d. data that are often encountered in the time series literature.

Theorem 3.4 (Uniform WLLN)
Let (Θ,ρ) and (E,d) be separable metric spaces, with corresponding Borel σ-algebras ϑ and $\mathcal{E}$. Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a sequence of random variables taking values in E, and $\{g_t\}_{t\in N_+}$ a sequence of functions from $E\times\Theta$ to $\mathbb{R}^{m\times n}$. Assume the following:

i) Θ is compact.

ii) $g:\Theta\to\mathbb{R}^{m\times n}$ is a continuous function on Θ.

iii) There exists a sequence $\{b_t\}_{t\in N_+}$ of real non-negative measurable functions on E such that

$$\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)=O_p(1)\quad\text{as } T\to\infty.$$

iv) g_t is Lipschitz in the following sense: there exists a measurable function $h:\mathbb{R}_+\to\mathbb{R}_+$ that is continuous at 0 such that $h(0)=0$ and, for any $t\in N_+$,

$$\left\|g_t(y,\theta)-g_t(y,\theta')\right\|\le b_t(y)\cdot h(\rho(\theta,\theta'))$$

for any $(y,\theta),(y,\theta')\in E\times\Theta$.

v) We have pointwise convergence:

$$\frac{1}{T}\sum_{t=1}^{T}g_t(Y_t,\theta)\xrightarrow{p}g(\theta)$$

as $T\to\infty$ pointwise for any $\theta\in\Theta$.

Then, we have the uniform WLLN

$$\sup_{\theta\in\Theta}\left\|\frac{1}{T}\sum_{t=1}^{T}g_t(Y_t,\theta)-g(\theta)\right\|\xrightarrow{p}0.$$

Proof) We define the sequence $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ of stochastic processes taking values in $\mathbb{R}^{n\times n}$ as

$$Q_T(\theta)=\frac{1}{T}\sum_{t=1}^{T}g_t(Y_t,\theta)$$

for any $T\in N_+$ and $\theta\in\Theta$. The Lipschitz condition on each g_t shows us that each $\{Q_T(\theta)\}_{\theta\in\Theta}$ has continuous paths. Similarly, define the sequence $\{B_T\}_{T\in N_+}$ of non-negative random variables as

$$B_T=\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t).$$

Finally, we define the sequence $\{\bar{Q}_T\}_{T\in N_+}$ of non-random functions on Θ as

$$\bar{Q}_T(\theta)=g(\theta)$$

for any $T\in N_+$ and $\theta\in\Theta$. Since g is a continuous function on the compcat set Θ, it is uniformly continuous on Θ; and it follows that $\{\bar{Q}_T\}_{T\in N_+}$ is an equicontinuous sequence of functions on Θ.

In light of the corollary to the characterization theorem, the desired result will follow if we can show that $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous. We first rely on the Lipschitz property to see that, for any $T\in N_+$ and $\theta,\theta'\in\Theta$,

$$\begin{aligned}\|Q_T(\theta)-Q_T(\theta')\|&\leq\frac{1}{T}\sum_{t=1}^{T}\|g_t(Y_t,\theta)-g_t(Y_t,\theta')\|\\&\leq\left(\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)\right)\cdot h(\rho(\theta,\theta'))=B_T\cdot h(\rho(\theta,\theta')).\end{aligned}$$

By assumption, $\{B_T\}_{T\in N_+}$ is $O_p(1)$, so that, for any $\epsilon,\eta>0$, there exists an $N(\epsilon,\eta)\in N_+$ and an $m(\epsilon,\eta)>0$ such that

$$\mathbb{P}\left(B_T>m(\epsilon,\eta)\cdot\epsilon\right)\leq\eta$$

for any $T \geq N(\epsilon, \eta)$ by the definition of boundedness in probability. Furthermore, by the continuity of h at 0, there exists a $\delta(\epsilon, \eta) > 0$ such that $h(d) < \frac{1}{2m(\epsilon,\eta)}$ for any $|d| < \delta(\epsilon, \eta)$. Setting

$$\Delta_T(\epsilon, \eta) = \frac{B_T}{m(\epsilon, \eta)}$$

and

$$U(\theta, \epsilon, \eta) = B_\rho(\theta, \delta(\epsilon, \eta)),$$

we can now see that, for any $T \geq N(\epsilon, \eta)$,

$$\mathbb{P}\left(\Delta_T(\epsilon, \eta) > \epsilon\right) < \eta$$

and

$$\sup_{\theta' \in U(\theta,\epsilon,\eta)} \left\| Q_T(\theta) - Q_T(\theta') \right\| \leq \frac{B_T}{2m(\epsilon, \eta)} < \Delta_T(\epsilon, \eta).$$

This holds for any $\epsilon, \eta > 0$ and $\theta \in \Theta$, so by definition $\{\{Q_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ is stochastically equicontinuous. We can see from the Lipschitz property that each $\{Q_T(\theta)\}_{\theta\in\Theta}$ also has continuous paths, so by the corollary to the characterization theorem,

$$\sup_{\theta\in\Theta} \left\| Q_T(\theta) - \bar{Q}_T(\theta) \right\| \xrightarrow{p} 0,$$

which is equivalent to the claim that

$$\sup_{\theta\in\Theta} \left\| \frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \theta) - g(\theta) \right\| \xrightarrow{p} 0.$$

Q.E.D.

3.3 The Basic Consistency and Asymptotic Normality Theorems

Based on the concept of uniform convergence in probability, we can now present the general conditions for the consistency of extremum estimators:

Theorem 3.5 (The Basic Consistency Theorem)
Let (Θ, d) be a separable metric space, and $\{\{\hat{Q}_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ a sequence of real stochastic processes with index set Θ. Let $\{Q_0(\theta)\}_{\theta\in\Theta}$ be a real valued but non-stochastic process. Suppose we have a sequence of extremum estimators $\{\hat{\theta}_T\}_{T\in N_+}$, where each $\hat{\theta}_T$ is found as a solution to the maximization problem

$$\max_{\theta\in\Theta} \hat{Q}_T(\theta).$$

Assume that the following conditions hold:

i) Θ is a compact space.

ii) Q_0 is uniquely maximized at some θ_0 in the interior of Θ.

iii) Q_0 is continuous on Θ.

iv) $\{\{\hat{Q}_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ converges uniformly in probability to Q_0, that is,

$$\sup_{\theta\in\Theta}\left|\hat{Q}_T(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0.$$

Then, $\{\hat{\theta}_T\}_{T\in N_+}$ is consistent for θ_0, that is,

$$\hat{\theta}_T \xrightarrow{p} \theta_0.$$

Proof Choose any $\epsilon > 0$. For any $T \in N_+$, since $\hat{\theta}_T$ is the maximizer of $\hat{Q}_T$ on θ,

$$\hat{Q}_T(\hat{\theta}_T) > \hat{Q}_T(\hat{\theta}_T) - \frac{\epsilon}{3} \geq \hat{Q}_T(\theta_0) - \frac{\epsilon}{3}.$$

everywhere on Ω. Meanwhile, the uniform consistency result tells us that

$$\lim_{T\to\infty} \mathbb{P}\left(\left|\hat{Q}_T(\hat{\theta}_T) - Q_0(\hat{\theta}_T)\right| > \frac{\epsilon}{3}\right) = 0$$

and

$$\lim_{T\to\infty} \mathbb{P}\left(\left|\hat{Q}_T(\theta_0) - Q_0(\theta_0)\right| > \frac{\epsilon}{3}\right) = 0.$$

Consider the event

$$H_T = \{Q_0(\hat{\theta}_T) > Q_0(\theta_0) - \epsilon\}.$$

Since

$$\left\{\hat{Q}_T(\hat{\theta}_T) > \hat{Q}_T(\theta_0) - \frac{\epsilon}{3}\right\} \cap \left\{\hat{Q}_T(\theta_0) > Q_0(\theta_0) - \frac{\epsilon}{3}\right\} \cap \left\{Q_0(\hat{\theta}_T) > \hat{Q}_T(\hat{\theta}_T) - \frac{\epsilon}{3}\right\}$$

is contained in H, and the probability of this intersection increases to 1 as $T \to \infty$, we have

$$\mathbb{P}(H_T) = \mathbb{P}\left(Q_0(\hat{\theta}_T) > Q_0(\theta_0) - \epsilon\right) \to 1$$

as $T \to \infty$.

Now choose any open ball $U = B_d(\theta_0, \delta)$ around θ_0; we can find such a $\delta > 0$ because θ_0 is in the interior of Θ. Define $\Theta^* = \Theta \cap U^c$; being a closed subset of a compact set, Θ^* is also compact. By the continuity of Q_0, the extreme value theorem tells us that there exists some $\theta^* \in \Theta^*$ such that

$$Q_0(\theta^*) = \max_{\theta \in \Theta^*} Q_0(\theta).$$

Since $\theta^* \notin U$, we have $\theta^* \neq \theta_0$. Q_0 is uniquely maximized at θ_0, so we must have $Q_0(\theta_0) > Q_0(\theta^*)$. Defining $\epsilon = Q_0(\theta_0) - Q_0(\theta^*) > 0$, it follows that, if $\theta \in \Theta$ satisfies

$$Q_0(\theta) > Q_0(\theta_0) - \epsilon,$$

or equiavlently, $Q_0(\theta) > Q(\theta^*)$, then we must have $\theta \in U$. In other words, for any $\delta > 0$ small enough,

$$\{Q_0(\hat{\theta}_T) > Q_0(\theta_0) - \epsilon\} \subset \{d(\hat{\theta}_T, \theta_0) < \delta\},$$

so that

$$\mathbb{P}\left(Q_0(\hat{\theta}_T) > Q_0(\theta_0) - \epsilon\right) \leq \mathbb{P}\left(d(\hat{\theta}_T, \theta_0) < \delta\right)$$

for any $T \in N_+$. Taking $T \to \infty$ on both sides now shows us that

$$\lim_{T \to \infty} \mathbb{P}\left(d(\hat{\theta}_T, \theta_0) < \delta\right) = 1.$$

It follows by definition that

$$\hat{\theta}_T \xrightarrow{p} \theta_0.$$

Q.E.D.

Like the basic consistency theorem, we first furnish very general sufficient conditions for asymptotic normality, and then apply it to obtain the asymptotic distribution of GMM estimators and other extremum estimators. First, we require a preliminary lemma:

Lemma 3.6 (Inversion at the Limit)
Let $\{X_T\}_{T\in N_+}$ be a sequence of n-dimensional random vectors, and $\{W_T\}_{T\in N_+}$ a sequence of $n\times n$ matrices. Suppose that there exists an n-dimensional random vector Y and a nonrandom, nonsingular $n\times n$ matrix W such that

$$\begin{aligned} Y_T := W_T \cdot X_T &\xrightarrow{d} Y \\ W_T &\xrightarrow{p} W. \end{aligned}$$

Then,

$$X_T \xrightarrow{d} W^{-1}Y.$$

Proof) Note initially that, because $\{Y_T\}_{T\in N_+}$ converges in distribution, it is bounded in probability, or $O_p(1)$. The claim above will follow from Slutsky's theorem if

$$Y_T - W\cdot X_T \xrightarrow{p} \mathbf{0},$$

since in this case the fact that $Y_T \xrightarrow{d} Y$ implies that $W\cdot X_T \xrightarrow{d} Y$ by the first version of Slutsky's theorem, and by the second version of Slutsky's theorem, we will have

$$X_T \xrightarrow{d} W^{-1}Y.$$

Choose any $\epsilon > 0$. Because the set of invertible $n\times n$ matrices $\overset{o}{\mathbb{R}}^{n\times n}$ is an open set with respect to the metric induced by the trace norm, there exists a $\delta_1 > 0$ such that $A \in \overset{o}{\mathbb{R}}^{n\times n}$ for any $A\in\mathbb{R}^{n\times n}$ such that

$$\|A - W\| \leq \delta_1.$$

Furthermore, since matrix inversion is a continuous operation on $\overset{o}{\mathbb{R}}^{n\times n}$, there exists a

$\delta_2 > 0$ such that, for any $A \in \overset{o}{\mathbb{R}}^{n\times n}$ such that $\|A - W\| \leq \delta_2$,

$$\left\|A^{-1} - W^{-1}\right\| < \epsilon.$$

It follows that, putting $\delta = \min(\delta_1, \delta_2) > 0$, for any $A \in \mathbb{R}^{n\times n}$ such that $\|A - W\| \leq \delta$, A^{-1} exists and

$$\left\|A^{-1} - W^{-1}\right\| < \epsilon.$$

In particular, for such A,

$$\left\|A^{-1}\right\| \leq \left\|A^{-1} - W^{-1}\right\| + \left\|W^{-1}\right\| < \epsilon + \left\|W^{-1}\right\|.$$

On the other hand, the fact that $Y_T = O_p(1)$ implies that there exists an $M > 0$ and $N \in N_+$ such that, for any $T \geq N$,

$$\mathbb{P}(|Y_T| > M) < \epsilon.$$

Fix any $\eta > 0$. Then, for $T \geq N$,

$$\begin{aligned}
\mathbb{P}(|Y_T - W \cdot X_T| > \eta) &= \mathbb{P}(\{|Y_T - W \cdot X_T| > \eta\} \cap \{\|W_T - W\| \leq \delta\}) \\
&\quad + \mathbb{P}(\{|W_T \cdot X_T - W \cdot X_T| > \eta\} \cap \{\|W_T - W\| > \delta\}) \\
&\leq \mathbb{P}\left(\left\{\left\|I_n - W \cdot W_T^{-1}\right\| |Y_T| > \eta\right\} \cap \{\|W_T - W\| \leq \delta\}\right) \\
&\quad + \mathbb{P}(\|W_T - W\| > \delta) \\
&\leq \mathbb{P}\left(\left\{\|W_T - W\| \cdot \left\|W_T^{-1}\right\| > \frac{\eta}{M}\right\} \cap \{\|W_T - W\| \leq \delta\}\right) \\
&\quad + \mathbb{P}(|Y_T| > M) + \mathbb{P}(\|W_T - W\| > \delta) \\
&\leq \mathbb{P}\left(\left\{\|W_T - W\| > \frac{\eta}{M(\epsilon + \|W^{-1}\|)}\right\} \cap \{\|W_T - W\| \leq \delta\}\right) \\
&\quad + \mathbb{P}(|Y_T| > M) + \mathbb{P}(\|W_T - W\| > \delta) \\
&\leq \mathbb{P}\left(\|W_T - W\| > \frac{\eta}{M(\epsilon + \|W^{-1}\|)}\right) + \epsilon + \mathbb{P}(\|W_T - W\| > \delta).
\end{aligned}$$

Since $W_T - W = o_p(1)$, taking $T \to \infty$ on both sides yields

$$\limsup_{T\to\infty} \mathbb{P}(|Y_T - W \cdot X_T| > \eta) \leq \epsilon.$$

This holds for any $\epsilon > 0$, so we can see that

$$\lim_{T\to\infty} \mathbb{P}(|Y_T - W \cdot X_T| > \eta) = 0,$$

and because $\eta > 0$ was chosen arbitrarily, $Y_T - W \cdot X_T \xrightarrow{p} \mathbf{0}$.

Q.E.D.

Now we are ready to prove the basic asymptotic normality result. It relies on the observation that, if the objective function is differentiable, then any extremum estimator is actually the zero of the gradient of the objective function. A mean value expansion of the gradient of the objective function then allows us to recover the asymptotic distribution of the extremum estimator. Below we state the general result for zero estimators, that is, the class of estimators that are obtained as the solution to some system of equations.

Theorem 3.7 (Basic Asymptotic Normality Theorem)
Let Θ be a non-empty convex subset of $\mathbb{R}^m$, and $\{Y_t\}_{t\in\mathbb{Z}}$ a sequence of random variables taking values in the separable metric space (E, d) with corresponding Borel σ-algebra $\mathcal{E}$. Let $\{\hat{G}_T\}_{T\in N_+}$ be a sequence of m-dimensional continuous functions where each $\hat{G}_T$ is defined on $E^T \times \Theta$. Suppose we have a sequence of zero estimators $\{\hat{\theta}_T(Y_1, \cdots, Y_T)\}_{T\in N_+}$, where each $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the unique solution to the equation

$$\hat{G}_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) = O_{m\times 1}.$$

Suppose that the following conditions hold:

i) The zero estimator is consistent for the true parameter:

$$\hat{\theta}_T(Y_1, \cdots, Y_T) \xrightarrow{p} \theta_0.$$

ii) $\theta_0 \in \Theta^o$, that is, the true parameter value lies in the interior of the parameter space.

iii) $\hat{G}_T$ is continuously differentiable with respect to θ on Θ^o.

iv) For some positive definite matrix $\Sigma(\theta_0) \in \mathbb{R}^{m\times m}$,

$$\sqrt{T} \cdot \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) \xrightarrow{d} \mathcal{N}[\mathbf{0}, \Sigma(\theta_0)].$$

v) There exists a matrix-valued function $H : \Theta \to \mathbb{R}^{m\times m}$ that is continuous and non-singular at θ_0, and is uniformly consistent for the Jacobian of $\hat{G}_T$ with respect to θ:

$$\sup_{\theta\in\Theta^o} \left\| \frac{\partial \hat{G}_T(Y_1, \cdots, Y_T \mid \theta)}{\partial \theta} - H(\theta) \right\| \xrightarrow{p} 0.$$

Then, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is asymptotically normal with rate of convergence $T^{1/2}$:

$$\sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0},\ H(\theta_0)^{-1}\Sigma(\theta_0)H(\theta_0)'^{-1}\right].$$

Furthermore, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ has the asymptotic expansion

$$\sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] = -H(\theta_0)^{-1}\sqrt{T} \cdot \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + o_p(1).$$

Proof) **Interiority of the Zero Estimator**

θ_0 lies in the interior of Θ, so there exists an $\epsilon > 0$ such that $B(\theta_0, \epsilon) \subset \Theta$. Since $\hat{\theta}_T$ is consistent for θ_0, by definition

$$\mathbb{P}\left(\left|\hat{\theta}_T - \theta\right| < \epsilon\right) \to 1$$

as $T \to \infty$; denoting

$$A_T = \{\omega \in \Omega \mid \hat{\theta}_T(\omega) \in \Theta^o\},$$

since $\{\omega \in \Omega \mid \left|\hat{\theta}_T(\omega) - \theta\right| < \epsilon\} \subset A_T$, we can claim that the probability of A_T goes to 1 as $T \to \infty$. This is also equivalent to saying that the indicator I_{A_T} converges to 1 in probability. Define

$$\theta_T^* = \hat{\theta}_T \cdot I_{A_T} + \theta_0 \cdot I_{A_T^c}$$

for any $T \in N_+$; $\{\theta_T^*\}_{T \in N_+}$ is a sequence of random variables taking values in Θ^o such that $\theta_T^* \xrightarrow{p} \theta_0$.

Mean Value Expansion of the Zero Function

Note that the ith component $\hat{G}_{T,i}$ of $\hat{G}_T$ is a continuously differentiable real valued function whose gradient we denote by $\boldsymbol{\nabla}_\theta \hat{G}_{T,i}(y, \theta)$ for any $(y, \theta) \in E^T \times \Theta^o$. By the stochastic mean value theorem,

$$\hat{G}_{T,i}(Y_1, \cdots, Y_T \mid \theta_T^*) = \hat{G}_{T,i}(Y_1, \cdots, Y_T \mid \theta_0) + \boldsymbol{\nabla}_\theta \hat{G}_{T,i}(Y_1, \cdots, Y_T \mid \tilde{\theta}_T^{(i)})'\,(\theta_T^* - \theta_0)\,,$$

where

$$\left|\tilde{\theta}_T^{(i)} - \theta_0\right| \leq |\theta_T^* - \theta_0|.$$

Taking $T \to \infty$ on both sides of the inequality above shows us that

$$\tilde{\theta}_T^{(i)} \xrightarrow{p} \theta_0.$$

This holds for any $1 \leq i \leq m$, so we can combine the results we have obtained so far to see that

$$\hat{G}_T(Y_1, \cdots, Y_T \mid \theta_T^*) = \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + \underbrace{\begin{pmatrix} \nabla_\theta \hat{G}_{T,1}(Y_1, \cdots, Y_T \mid \tilde{\theta}_T^{(1)})' \\ \vdots \\ \nabla_\theta \hat{G}_{T,m}(Y_1, \cdots, Y_T \mid \tilde{\theta}_T^{(m)})' \end{pmatrix}}_{\hat{H}_T} (\theta_T^* - \theta_0)$$

$$= \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + I_{A_T} \hat{H}_T \cdot (\hat{\theta}_T - \theta_0).$$

Rearragning terms and pre-multiplying by $\sqrt{T}$ shows us that

$$I_{A_T} \hat{H}_T \cdot \sqrt{T} \left[\hat{\theta}_T - \theta_0 \right] = -\sqrt{T} \cdot \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + \sqrt{T} \cdot \hat{G}_T(Y_1, \cdots, Y_T \mid \theta_T^*).$$

Asymptotic Behavior of Components of Mean Value Expansion

We study each term in the mean value expansion in turn:

1. $I_{A_T} \hat{H}_T$
   For any $1 \leq i \leq m$,

   $$\left| \nabla_\theta \hat{G}_{T,i}(Y_1, \cdots, Y_T \mid \tilde{\theta}_T^{(i)}) - H_i(\tilde{\theta}_T^{(i)}) \right| \leq \sup_{\theta \in \Theta^o} \left\| \frac{\partial \hat{G}_T(Y_1, \cdots, Y_T \mid \theta)}{\partial \theta} - H(\theta) \right\|,$$

   where $H_i(\theta)$ is the ith row of $H(\theta)$ for any $\theta \in \Theta^o$. The right hand side goes to 0 in probability as $T \to \infty$, while

   $$H_i(\tilde{\theta}_T^{(i)}) \xrightarrow{p} H_i(\theta_0)$$

   by the fact that $\tilde{\theta}_T^{(i)} \xrightarrow{p} \theta_0$ and the continuity of H at θ_0, so putting these results together, we have

   $$\nabla_\theta \hat{G}_{T,i}(Y_1, \cdots, Y_T \mid \tilde{\theta}_T^{(i)}) \xrightarrow{p} H_i(\theta_0).$$

   This holds for any $1 \leq i \leq m$, so

   $$\hat{H}_T \xrightarrow{p} H(\theta_0).$$

   Finally, since $I_{A_T} \xrightarrow{p} 1$ and convergence in probability is preserved under arithmetic

operations,

$$I_{A_T}\hat{H}_T \xrightarrow{p} H(\theta_0).$$

2. $\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_T^*)$
   Observe that

$$\begin{aligned}\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_T^*) &= I_{A_T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T) + I_{A_T^c}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0)\\ &= (1-I_{A_T})\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0)\end{aligned}$$

   by the assumption that

$$\hat{G}_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T) = O_{m\times 1}.$$

   By the asymptotic normality assumption,

$$\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0) = O_p(1)$$

   while $1-I_{A_T} = o_p(1)$, which implies that

$$\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_T^*) = (1-I_{A_T})\cdot\sqrt{T}\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0) = o_p(1).$$

3. $\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0)$
   This term is equal to

$$\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_t \mid \theta_0)$$

   by the definition of $\hat{G}_T$.

Inverting at the Limit

Therefore, we can see that

$$\begin{aligned}I_{A_T}\hat{H}_T\cdot\sqrt{T}\left[\hat{\theta}_T-\theta_0\right] &= -\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0) + o_p(1)\\ &\xrightarrow{d}\mathcal{N}\left[\mathbf{0},\Sigma(\theta_0)\right].\end{aligned}$$

by Slutsky's theorem. Finally, by the inversion lemma, we have

$$\sqrt{T}\left[\hat{\theta}_T-\theta_0\right]\xrightarrow{d}\mathcal{N}\left[\mathbf{0},\ H(\theta_0)^{-1}\Sigma(\theta_0)H(\theta_0)'^{-1}\right].$$

Asymptotic Expansion of $\hat{\theta}_T$

We just proved that $\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = O_p(1)$, and because

$$\left\| I_{A_T}\hat{H}_T - H(\theta_0) \right\| \xrightarrow{p} 0,$$

it follows that

$$H(\theta_0)\cdot\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] - I_{A_T}\hat{H}_T\cdot\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = o_p(1).$$

By implication,

$$H(\theta_0)\cdot\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = -\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0) + o_p(1),$$

and premultiplying both sides by $H(\theta_0)^{-1}$ yields

$$\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = -H(\theta_0)^{-1}\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T \mid \theta_0) + o_p(1).$$

Q.E.D.

3.4 The Asymptotics of GMM Estimation

Generalized method of moments (GMM) is a semi-parametric estimation method that extends classical method of moments estimation. It is a method that estimates a finite number of parameters by minimizing the sample analogue of a set of population moment conditions that are assumed to hold.

3.4.1 A Motivating Example

To motivate the use of GMM, consider the following example. Suppose we have a random sample $X_1, \cdots, X_n$ drawn from the Poisson distribution with parameter $\lambda > 0$. One way to estimate the parameter λ is by finding the value of λ that maximizes the joint likelihood of $X_1, \cdots, X_n$; this is the maximum likelihood estimator of λ, and it is given as the sample mean

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Another way to estimate λ is through the method of moments. Since λ is the mean of a Poisson distribution with parameter λ, it follows that the sample observations satisfy the population moment condition

$$\mathbb{E}[X_i] = \lambda.$$

The sample analogue of this moment condition is given as

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \lambda = 0;$$

therefore, by the analogy principle, the method of moments estimator of λ is given as

$$\hat{\lambda}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which in this case is the same as the MLE of λ.

However, the above moment condition is not the only one that we might use. Since the variance of a Poisson random variable with parameter λ is also equal to λ, we might consider the moment condition

$$\mathbb{E}\left[X_i^2\right] - \mathbb{E}[X_i]^2 = \lambda.$$

The sample analogue of this moment condition is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \lambda = 0,$$

but the method of moments estimator $\hat{\lambda}_{MM}$ does not satisfy this equation. Since both population moment conditions are valid, what do we do in this case?

Here is where the generalized part of GMM comes into play. Suppose we choose some weight $w \in (0,1)$ that represents the weight that we give to the first moment condition. In this case, we want to make the weighted average

$$Q(\lambda) = w\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \lambda\right)^2 + (1-w)\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \lambda\right)^2$$

of the two moment conditions as close to 0 as possible. The greater w, the more importance we are assigning to matching the first moment condition. The value of λ that minimizes the above weighted average is called the GMM estimator of λ.

Letting $\hat{\lambda}_{GMM}$ be the GMM estimator of λ, note that the first order condition for minimization tells us that

$$\frac{\partial Q(\hat{\lambda}_{GMM})}{\partial \lambda} = -2w\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \hat{\lambda}_{GMM}\right) - 2(1-w)\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \hat{\lambda}_{GMM}\right) = 0,$$

solving which we obtain

$$\hat{\lambda}_{GMM} = w\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + (1-w)\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right].$$

Note how the resulting GMM estimator of λ is a weighted average of the sample mean, which matches the first moment condition, and the sample variance, which matches the second moment condition.

3.4.2 A General Framework and Regularity Assumptions

To formalize the setting, let the parameter space Θ be a subset of the euclidean m-space $\mathbb{R}^m$, and suppose the sample $\{Y_t\}_{t\in\mathbb{Z}}$ takes values in the separable metric space (E,d) with corresponding Borel σ-algebra $\mathcal{E}$. Suppose there exists a sequence of functions $\{g_t\}_{t\in\mathbb{Z}}$ that maps $E\times\Theta$ into $\mathbb{R}^k$ such that

$$\mathbb{E}\left[g_t(Y_t,\theta)\right] = O_{k\times 1}$$

for any $t \in \mathbb{Z}$. The sample analogue of this population moment condition is

$$\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) = O_{k\times 1}.$$

We want to find the value of θ that solves this equation; if such a value exists, then it is the classical method of moments estimator of θ.

If $k < m$, that is, if there are more parameters of interest than moment conditions, then in general there does not exist a solution to the above system of equations; therefore, for a method of moments estimator to exist, it must be the case that $k \geq m$, so that there are at least as many moment conditions as there are parameters.

If $k > m$, then it could also be the case that there is no solution to the above system of equations. However, the problem in this case is quite different from the case $k < m$; in the case $k > m$, we could simply choose m moment conditions at random and find the method of moments estimator of θ that solves the chosen moment conditions. As such, the problem here is not one of insufficient information. It is one where there is too much information, and where we require a guide to help us choose which pieces of information to include in our estimator of θ.

In general, when $k \geq m$, we determine which moment conditions to value above the others by choosing a sequence of positive semidefinite weighting matrices $\{W_T\}_{T \in N_+}$. Given a sample of size T, then, we choose the value of θ that maximizes the quadratic form

$$\left(\frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \theta) \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \theta) \right)$$

with respect to θ. Here, W_T determines the weight that is assigned to each moment condition, that is, which information to include and which to discard when constructing the GMM estimator of θ.

Going forward, we will make the following regularity assumptions:

G1 : **The Sample and Parameter Space**

Let $\{Y_t\}_{t \in \mathbb{Z}}$ be a sample taking values in the separable metric space (E, d) with corresponding Borel σ-algebra $\mathcal{E}$. Furthermore, let the parameter space Θ be a non-empty convex and compact subset of $\mathbb{R}^m$. The true parameter θ_0 is assumed to be contained in the interior Θ^o of the parameter space.

G2 : **Moment Conditions and Weighting Matrices**

Suppose that there exists a sequence of functions $\{g_t\}_{t \in \mathbb{Z}}$ on $E \times \Theta$ into $\mathbb{R}^{k \times 1}$, where $k \geq m$, such that $\mathbb{E}[g_t(Y_t, \theta)]$ is well-defined and time invariant for any $\theta \in \Theta$, and

$$\mathbb{E}[g_t(Y_t, \theta_0)] = O_{k \times 1}$$

for any $t \in \mathbb{Z}$. We call the functions g_t moment functions.

Let there exist a sequence of continuous functions $\{W_T\}_{T \in N_+}$, where each W_T is defined on E^T and maps into the set of all positive semidefinite $k \times k$ matrices. We assume that

there exists some positive definite non-random matrix $W \in \mathbb{R}^{k\times k}$ such that

$$W_T(Y_1, \cdots, Y_T) \xrightarrow{p} W.$$

G3 : **Continuity Properties of the Moment Functions**

We assume that the functions $\{g_t\}_{t\in\mathbb{Z}}$ satisfy the following conditions:

– **Consistency of Sample Moments**

Defining $g : \Theta \to \mathbb{R}^k$ as

$$g(\theta) = \mathbb{E}\left[g_t(Y_t, \theta)\right]$$

for any $\theta \in \Theta$, $g(\cdot)$ is continuous on Θ and

$$\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t, \theta) \xrightarrow{p} g(\theta)$$

pointwise for any $\theta \in \Theta$.

– **Lipschitz Continuity**

There exists a measurable function $h : \mathbb{R}_+ \to \mathbb{R}_+$ that is continuous at 0 with $h(0) = 0$ and a sequence $\{b_t\}_{t\in N_+}$ of non-negative measurable functions on E such that

$$\frac{1}{T}\sum_{t=1}^{T} b_t(Y_t)^2 = O_p(1) \quad \text{as } T \to \infty$$

and

$$|g_t(y, \theta) - g_t(y, \gamma)| \leq b_t(y) \cdot h(|\theta - \gamma|)$$

for any $y \in E$ and $\theta, \gamma \in \Theta$. In addition, g_t is continuous on $E \times \Theta$.

Note that, by Hölder's inequality,

$$\frac{1}{T}\sum_{t=1}^{T} b_t(Y_t) \leq \left(\frac{1}{T}\sum_{t=1}^{T} b_t(Y_t)^2\right)^{\frac{1}{2}},$$

so that $\frac{1}{T}\sum_{t=1}^{T} b_t(Y_t) = O_p(1)$ as well.

G4 : **Identification Condition**

Note that $g(\theta_0) = O_{k\times 1}$ because of the population moment conditions. We assume that $g(\theta) \neq O_{k\times 1}$ if $\theta \neq \theta_0$.

These are the four properties that ensure the consistency of the GMM estimator, should it exist. In particular, the identification condition, while seemingly innocuous, is the most important assumption, since it characterizes the true parameter and makes it clear what our estimator is estimating.

3.4.3 Consistency of the GMM Estimator and Identification

Here we show that the GMM estimator exists and is consistent for the true parameter, if it exists. We also shed some more light on the exact meaning of the identification assumption.

Theorem 3.8 (Existence and Consistency of GMM Estimator)
Suppose that assumptions G1 to G4 hold. Then, there exists a sequence of functions $\{\hat{\theta}_T\}_{T\in N_+}$ such that each $\hat{\theta}_T$ is maps E^T into Θ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ solves the minimization problem

$$\min_{\theta\in\Theta} \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)' W_T(Y_1,\cdots,Y_T) \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right).$$

In addition, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is consistent for the true parameter, that is,

$$\hat{\theta}_T(Y_1,\cdots,Y_T) \xrightarrow{p} \theta_0.$$

Proof) **Existence**

We first establish that the GMM estimators exist. For any $T \in N_+$, define $\hat{Q}_T : E^T \times \Theta \to \mathbb{R}$ as

$$\hat{Q}_T(y_1,\cdots,y_T \mid \theta) = -\left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t,\theta)\right)' W_T(y_1,\cdots,y_T) \left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t,\theta)\right).$$

for any $(y_1, \cdots, y_T, \theta) \in E^T \times \Theta$. By the assumed continuity of g_t and W_T, $\hat{Q}_T$ is a continuous function on $E^T \times \Theta$. Furthermore, since compact spaces are complete and separable, Θ is a complete and separable space under the euclidean metric, so that, by theorem 3.1, there exists a Borel meaurable function $\hat{\theta}_T : E^T \to \Theta$ such that

$$\hat{Q}_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) = \max_{\theta\in\Theta} \hat{Q}_T(Y_1,\cdots,Y_T \mid \theta).$$

It follows by definition that $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the GMM estimator given a sample of size T; without ambiguity, we denote $\hat{\theta}_T$ for the GMM estimator and W_T for the random

matrix $W_T(Y_1, \cdots, Y_T)$ going forward.

Consistency

To prove that the GMM estimators are consistent for the true parameter θ_0, define the sequence of stochastic processes $\{\{\hat{Q}_T(\theta)\}_{\theta\in\Theta}\}_{T\in N_+}$ as

$$\hat{Q}_T(\theta) = -\left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)$$

for any $\theta \in \Theta$. Note that, by assumption, $g_t(y,\cdot)$ is continuous on Θ for any $y \in E$, so that $\{\hat{Q}_T(\theta)\}_{\theta\in\Theta}$ is a stochastic process with continuous paths.

Now define the limiting objective function $Q_0 : \Theta \to \mathbb{R}$ as

$$Q_0(\theta) = -g(\theta)' W g(\theta)$$

for any $\theta \in \Theta$; it is continuous on Θ by the continuity of g. Since $g(\theta) = O_{k\times 1}$ if and only if $\theta = \theta_0$, and W is a positive definite matrix, Q_0 is uniquely maximized at θ_0. In addition, by assumption

$$\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) \xrightarrow{p} g(\theta)$$

pointwise for any $\theta \in \Theta$, g is a continuous function on Θ, and there exists a sequence $\{b_t\}_{t\in N_+}$ of real non-negative measurable functions on E and a measurable function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that

- h is continuous at 0 with $h(0) = 0$,
- $\frac{1}{T}\sum_{t=1}^{T} b_t(Y_t) = O_p(1)$ and
- Each g_t is Lipschitz in the sense that

$$|g_t(y,\theta) - g_t(y,\gamma)| \le b_t(y)\cdot h(|\theta - \gamma|)$$

  for any $\theta, \gamma \in \Theta$ and $y \in E$.

By the uniform WLLN,

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right| \xrightarrow{p} 0.$$

For any $\theta \in \Theta$,

$$\hat{Q}_T(\theta) - Q_0(\theta) = -\left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right) + g(\theta)' W g(\theta)$$

$$
\begin{aligned}
&= -\left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right)' (W_T - W)\left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right) \\
&\quad - 2g(\theta)' W \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right) \\
&\quad - 2g(\theta)'(W_T - W)\left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right) \\
&\quad - g(\theta)'(W_T - W)g(\theta) \\
&\quad - \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right)' W \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right),
\end{aligned}
$$

so we have

$$
\begin{aligned}
\left|\hat{Q}_T(\theta) - Q_0(\theta)\right| \leq &\left|\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right|^2 \cdot (\|W_T - W\| + \|W\|) \\
&+ 2\cdot|g(\theta)|\cdot\|W\|\cdot\left|\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right| \\
&+ 2|g(\theta)|\cdot\|W_T - W\|\left|\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta) - g(\theta)\right| \\
&+ |g(\theta)|^2\|W_T - W\|.
\end{aligned}
$$

By the continuity of g on Θ, $\sup_{\theta\in\Theta}|g(\theta)| < +\infty$, so it follows from the uniform WLLN derived above and the fact that $\|W_T - W\| = o_p(1)$ that

$$
\sup_{\theta\in\Theta}\left|\hat{Q}_T(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0.
$$

We have thus seen that:

- Θ is compact
- Q_0 is a continuous function on Θ
- Q_0 is uniquely maximized at θ_0
- $\hat{Q}_T$ converges uniformly to Q_0:

$$
\sup_{\theta\in\Theta}\left|\hat{Q}_T(\theta) - Q_0(\theta)\right| \xrightarrow{p} 0.
$$

Therefore, by the basic consistency theorem, $\hat{\theta}_T \xrightarrow{p} \theta_0$.

Q.E.D.

3.4.4 Asymptotic Distribution of the GMM Estimator

To obtain the asymptotic distribution of the GMM estimator, we make use of the basic asymptotic normality theorem above. To do so, we must make some additional regularity assumptions:

G5 : **Smoothness of Moment Functions**

We assume that for each $t \in \mathbb{Z}$ and $y \in E$, $g_t(y,\cdot)$ is continuously differentiable on Θ_0, and define $G_t : E \times \Theta^o \to \mathbb{R}^{k\times m}$ as

$$G_t(y,\theta) = \frac{\partial g_t(y,\theta)}{\partial \theta}$$

for any $(y,\theta) \in E \times \Theta^o$. Furthermore, we assume that

- **Limiting Moment Derivative**

  There exists a continuous function $G : \Theta \to \mathbb{R}^{k\times m}$ that satisfies

  $$\frac{1}{T}\sum_{t=1}^{T} G_t(Y_t,\theta) \xrightarrow{p} G(\theta)$$

  for any $\theta \in \Theta^o$ and $\sup_{\theta\in\Theta^o} \|G(\theta)\| < +\infty$. Denote $G(\theta_0) = G_0$, and assume G_0 has rank m.

- **Lipschitz Continuity**

  There exists a measurable function $H : \mathbb{R}_+ \to \mathbb{R}_+$ that is continuous at 0 with $h(0) = 0$ and a sequence $\{B_t\}_{t\in N_+}$ of non-negative measurable functions on E such that

  $$\frac{1}{T}\sum_{t=1}^{T} B_t(Y_t) = O_p(1) \quad \text{as } T \to \infty$$

  and

  $$\|G_t(y,\theta) - G_t(y,\gamma)\| \le B_t(y)\cdot H(|\theta-\gamma|)$$

  for any $y \in E$, $\theta,\gamma \in \Theta^o$. G_t is also continuous on $E \times \Theta$.

G6 : **MDS Errors**

We assume that there exists some $\delta > 0$ such that $\{g_t(Y_t,\theta_0)\}_{t\in\mathbb{Z}}$ is an $L^{2+\delta}$-bounded MDS with respect to some filtration $\mathcal{F}$. In addition, let

$$\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta_0)g_t(Y_t,\theta_0)' \xrightarrow{p} \Sigma,$$

where Σ is a $k \times k$ positive definite matrix.

G7 : **Strict Convexity of Objective Function**

For any $T \in N_+$ and $(y_1, \cdots, y_T) \in E^T$, assume that the mapping

$$\theta \mapsto \left(\frac{1}{T} \sum_{t=1}^{T} g_t(y_t, \theta) \right)' W_T(y_1, \cdots, y_T) \left(\frac{1}{T} \sum_{t=1}^{T} g_t(y_t, \theta) \right).$$

is strictly convex on Θ.

By the MDS CLT, assumption G6 implies that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t(Y_t, \theta_0) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1}, \Sigma\right].$$

In addition, the strict convexity of the objective function in the parameter θ ensures that the function $\hat{\theta}_T : E^T \to \Theta$ is continuous.

Under our current assumptions, we can now establish the asymptotic normality of the GMM estimator:

Theorem 3.9 (Asymptotic Normality of GMM Estimator)
Suppose that assumptions G1 to G7 hold, and assume that there exists a sequence of GMM estimators $\{\hat{\theta}_T(Y_1, \cdots, Y_T)\}_{T \in N_+}$ such that each $\hat{\theta}_T(Y_1, \cdots, Y_T)$ solves the minimization problem

$$\min_{\theta \in \Theta} \left(\frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \theta) \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \theta) \right).$$

Suppose that each $\hat{\theta}_T$ takes values in the interior Θ^o of the parameter space. Then, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is asymptotically normal with the following asymptotic distribution:

$$\sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ (G_0' W G_0)^{-1} G_0' W \Sigma W G_0 (G_0' W G_0)^{-1}\right].$$

Furthermore, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is asymptotically linear in the following sense:

$$\sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] = -(G_0' W G_0)^{-1} G_0' W \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t(Y_t, \theta_0) + o_p(1).$$

Proof) We just established that, under assumptions G6 and G7, each $\hat{\theta}_T : E^T \to \Theta$ is a continuous function on E^T, and

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t(Y_t, \theta_0) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1}, \Sigma\right].$$

We also assumed in the claim of the theorem that $\hat{\theta}_T$ takes values in Θ^o.

Define the function $\hat{G}_T : E^T \times \Theta \to \mathbb{R}^m$ as

$$\hat{G}_T(y_1, \cdots, y_T \mid \theta) = \left(\frac{1}{T} \sum_{t=1}^{T} G_t(y_t, \hat{\theta}_T(y_1, \cdots, y_T)) \right)' W_T(y_1, \cdots, y_T) \left(\frac{1}{T} \sum_{t=1}^{T} g_t(y_t, \theta) \right)$$

for any $(y_1, \cdots, y_T) \in E^T$ and $\theta \in \Theta$. Due to the continuity of all the functions involved, $\hat{G}_T$ is continuous on $E^T \times \Theta$. It is also differentiable with derivative given by

$$\frac{\partial \hat{G}_T(y_1, \cdots, y_T \mid \theta)}{\partial \theta} = \left(\frac{1}{T} \sum_{t=1}^{T} G_t(y_t, \hat{\theta}_T(y_1, \cdots, y_T)) \right)' W_T(y_1, \cdots, y_T) \left(\frac{1}{T} \sum_{t=1}^{T} G_t(y_t, \theta) \right)$$

for any $(y_1, \cdots, y_T) \in E^T$ and $\theta \in \Theta^o$; again, this is continuous on $E^T \times \Theta^o$, so $\hat{G}_T$ is continuously differentiable with respect to θ on Θ^o. Going forward, we denote $\hat{\theta}_T$ for $\hat{\theta}_T(Y_1, \cdots, Y_T)$ and W_T for $W_T(Y_1, \cdots, Y_T)$ without ambiguity.

By the first order conditions for minimization,

$$\hat{G}_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T) = \left(\frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \hat{\theta}_T) \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} g_t(Y_t, \hat{\theta}_T) \right) = O_{m \times 1},$$

so that $\hat{\theta}_T$ is the zero estimator with zero function $\hat{G}_T$.

The existence of a continuous limiting function for $\frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \theta)$ and the Lipschitz continuity assumption now tells us, via the uniform WLLN, that

$$\sup_{\theta \in \Theta^o} \left\| \frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \theta) - G(\theta) \right\| \xrightarrow{p} 0.$$

It follows that

$$\left\| \frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \hat{\theta}_T) - G(\hat{\theta}_T) \right\| \xrightarrow{p} 0,$$

and by the continuity of G and the continuous mapping theorem, we have

$$G(\hat{\theta}_T) \xrightarrow{p} G(\theta_0) = G_0.$$

Putting these results together, we have

$$\frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \hat{\theta}_T) \xrightarrow{p} G_0.$$

This result, in conjunction with assumption G6 and Slutsky's theorem, now shows us that

$$\begin{aligned}\sqrt{T}\cdot\hat{G}_T(Y_1,\cdots,Y_T\mid\theta_0) &= \left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\hat{\theta}_T)\right)' W_T\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}g_t(Y_t,\theta_0)\right)\\ &\xrightarrow{d} G_0'W\times\mathcal{N}\left[O_{k\times 1},\Sigma\right]=\mathcal{N}\left[O_{m\times 1},G_0'W\Sigma WG_0\right].\end{aligned}$$

From the previous theorem, we know that $\hat{\theta}_T$ exists and is consistent for θ_0 under G1 to G4. Furthermore, by assumption $\theta_0\in\Theta^o$, and because $G_t(y,\cdot)$ and $g(y,\cdot)$ are continuous functions on Θ^o. We have just seen that $\hat{G}_T$ is continuously differentiable with respect to θ on Θ^o.

Finally, since

$$\begin{aligned}\frac{\partial\hat{G}_T(Y_1,\cdots,Y_T\mid\theta)}{\partial\theta}-G_0'WG(\theta) &= \left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\hat{\theta}_T)\right)' W_T\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\theta)\right)-G_0WG(\theta)\\ &= \left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\hat{\theta}_T)-G_0\right)'(W_T-W)\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\theta)-G(\theta)\right)\\ &\quad+G_0'W\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\theta)-G(\theta)\right)\\ &\quad+G_0'(W_T-W)\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\theta)-G(\theta)\right)\\ &\quad+\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\hat{\theta}_T)-G_0\right)'W\cdot G(\theta)\\ &\quad+G_0'(W_T-W)G(\theta)\\ &\quad+\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\hat{\theta}_T)-G_0\right)'W\left(\frac{1}{T}\sum_{t=1}^{T}G_t(Y_t,\theta)-G(\theta)\right)\end{aligned}$$

for any $\theta\in\Theta^o$ and $\sup_{\theta\in\Theta^o}\|G(\theta)\|<+\infty$ by assumption, it follows that

$$\sup_{\theta\in\Theta^o}\left\|\frac{\partial\hat{G}_T(Y_1,\cdots,Y_T\mid\theta)}{\partial\theta}-G_0WG(\theta)\right\|\xrightarrow{p}0.$$

Therefore, by the basic asymptotic normality theorem, we can conclude that

$$\sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T)-\theta_0\right]\xrightarrow{d}\mathcal{N}\left[O_{m\times 1},\ (G_0'WG_0)^{-1}G_0'W\Sigma WG_0(G_0'WG_0)^{-1}\right],$$

and that $\hat{\theta}_T(Y_1, \cdots, Y_T)$ has the asymptotic expansion

$$\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = -(G_0' W G(\theta_0))^{-1} \cdot \sqrt{T}\hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + o_p(1).$$

We saw above that

$$\frac{1}{T}\sum_{t=1}^{T} G_t(Y_t, \theta_0) - G_0 \xrightarrow{p} O_{k \times m}$$

and $W_T \xrightarrow{p} W$, while $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0) = O_p(1)$, so

$$\sqrt{T}\hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) - G_0' W \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)$$

$$= \left[\left(\frac{1}{T}\sum_{t=1}^{T} G_t(Y_t, \theta_0)\right)' W_T - G_0' W\right] \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right) = o_p(1).$$

It follows that

$$\sqrt{T}\left[\hat{\theta}_T - \theta_0\right] = -(G_0' W G_0)^{-1} \cdot \sqrt{T}\hat{G}_T(Y_1, \cdots, Y_T \mid \theta_0) + o_p(1)$$

$$= -(G_0' W G_0)^{-1} G_0' W \frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0) + o_p(1).$$

Q.E.D.

3.4.5 Efficient GMM Estimation

For the remainder of the section on GMM estimation we give general treatments of several topics that we will cover again when studying GMM estimation of linear regression models with instrumental variables. The first of these is the topic of efficient GMM estimation, that is, the search for the optimal asymptotic weighting matrix W. We first present some requisite lemmas, that are also of independent interest.

Lemma 3.10 (Simultaneous Diagonalization of Positive Definite Matrices)
Let $A, B \in \mathbb{R}^{k\times k}$ be symmetric matrices such that A is positive definite and B is positive semidefinite. Then, there exists a non-singular $P \in \mathbb{R}^{k\times k}$ and a diagonal matrix $D \in \mathbb{R}^{k\times k}$ with non-negative diagonal entries such that

$$A = PP' \quad \text{and} \quad B = PDP'.$$

In addition, $P'P$ has the same set of eigenvalues as A, and the diagonal elements of D are solutions to the eigenvalue equation

$$\det(B - \lambda \cdot A) = 0,$$

or equivalently, the eigenvalues of $A^{-\frac{1}{2}}BA^{-\frac{1}{2}\prime}$, where $A^{\frac{1}{2}}$ is the Cholesky factor of A.

Proof) To see this, let $A = \Lambda_1 D_1 \Lambda_1'$ be the orthogonal diagonalization of A (which exists by the principal axis theorem). Defining $D_1^{\frac{1}{2}}$ as the diagonal matrix whose diagonal elements are the square roots of the eigenvalues of A, $D_1^{\frac{1}{2}} D_1^{\frac{1}{2}\prime} = D_1$ and $D_1^{\frac{1}{2}}$ is nonsingular because A is positive definite. Defining $C = \Lambda_1 D_1^{\frac{1}{2}}$, it follows that $A = CC'$.

Since $C^{-1}B(C')^{-1}$ is a positive semidefinite matrix, it has eigendecomposition

$$C^{-1}B(C')^{-1} = \Lambda_2 D \Lambda_2',$$

where Λ_2 is an orthogonal matrix and D collects the eigenvalues of $C^{-1}B(C')^{-1}$, which are all non-negative valued by the positive semidefiniteness of $C^{-1}B(C')^{-1}$. Defining $P = C\Lambda_2$, we can immediately see that

$$B = (C\Lambda_2)D_2(C\Lambda_2)' = PD_2P',$$

and that

$$PP' = C\Lambda_2\Lambda_2'C' = CC' = A.$$

To establish the claims about the eigenvalues of $P'P$, note that

$$P'P = \Lambda_2' C' C \Lambda_2 = \Lambda_2' D_1^{\frac{1}{2}} \underbrace{\Lambda_1' \Lambda_1}_{I_k} D_1^{\frac{1}{2}} \Lambda_2 = \Lambda_2' D_1 \Lambda_2.$$

As such, the characteristic polynomial of $P'P$ is given by

$$ch_{P'P}(\lambda) = \det(P'P - \lambda \cdot I_k) = \det(\Lambda_2' D_1 \Lambda_2 - \lambda \cdot I_k) = \det(D_1 - \lambda \cdot I_k),$$

where we used the fact that Λ_2 is orthogonal to justify the last equality. Thus, $P'P$ shares the same eigenvalues as D_1 and therefore A.

Finally, the final claim can be seen by first making the observation that the diagonal elements of D are exactly equal to its eigenvalues. The characteristic polynomial of D is given as

$$ch_D(\lambda) = \det(D - \lambda \cdot I_k) = \det\Big(P^{-1}\Big)^2 \cdot \det(PDP' - \lambda \cdot PP') = \det\Big(P^{-1}\Big)^2 \cdot \det(B - \lambda A),$$

so the eigenvalues of D are exactly the solutions to the equation $\det(B - \lambda \cdot A) = 0$.

Q.E.D.

Lemma 3.11 (Evaluating Positive Definiteness of Matrix Differences)
Let $A, B \in \mathbb{R}^{k \times k}$ be positive definite matrices. Then, the following are equivalent:

i) $A - B$ is positive (semi)definite.

ii) $B^{-1} - A^{-1}$ is positive (semi)definite.

iii) The solutions to the equation $\det(B - \lambda \cdot A) = 0$, which are all real positive valued, have modulus less than (or equal to) 1.

Proof) By the preceding lemma, there exists a non-singular $P \in \mathbb{R}^{k \times k}$ and a diagonal matrix $D \in \mathbb{R}^{k \times k}$ with non-negative diagonal entries such that

$$A = PP' \quad \text{and} \quad B = PDP'.$$

Since B is positive definite, the diagonal elements of D are all positive; otherwise, B has reduced rank, which contradicts is positive definiteness. We saw above that the diagonal elements of D are the solutions to the equation $\det(B - \lambda \cdot A) = 0$; thus, the solutions to the equation $\det(B - \lambda \cdot A) = 0$ are all positive. Let $\lambda_1 \geq \cdots \geq \lambda_k > 0$ be the diagonal elements of D.

The difference $A-B$ can now be written as

$$A-B=P(I_k-D)P'=P\begin{pmatrix}1-\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1-\lambda_k\end{pmatrix}P'.$$

Therefore, for any non-zero $\alpha\in\mathbb{R}^k$,

$$\alpha'(A-B)\alpha=\alpha'P\begin{pmatrix}1-\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1-\lambda_k\end{pmatrix}P'\alpha\geq\left(1-\max_{1\leq i\leq k}\lambda_i\right)\cdot\left|P'\alpha\right|^2,$$

where $P'\alpha$ is non-zero because P' is non-singular and α is non-zero. It follows that $A-B$ is positive (semi)definite if and only if

$$1>(\geq)\max_{1\leq i\leq k}\lambda_i.$$

Similarly,

$$B^{-1}-A^{-1}=(P')^{-1}\left(D^{-1}-I_k\right)P^{-1}=(P')^{-1}\begin{pmatrix}\frac{1}{\lambda_1}-1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_k}-1\end{pmatrix}P^{-1}$$

and, for any non-zero $\alpha\in\mathbb{R}^k$,

$$\alpha'(B^{-1}-A^{-1})\alpha\geq\left(\min_{1\leq i\leq k}\frac{1}{\lambda_i}-1\right)\cdot\left|P^{-1}\alpha\right|^2.$$

We can thus see that $B^{-1}-A^{-1}$ is positive (semi)definite if and only if

$$\min_{1\leq i\leq k}\frac{1}{\lambda_i}>(\geq)1,$$

which is equivalent to the inequality

$$1>(\geq)\max_{1\leq i\leq k}\lambda_i.$$

This establishes the claim.

Q.E.D.

Theorem 3.12 (Optimal Asymptotic Weighting Matrix)
Suppose that assumptions G1 to G7 hold, and let $\{W_T^*\}_{T\in N_+}$ be a sequence of positive definite valued $k\times k$ random matrices that converge to Σ^{-1} in probability. Let $\{\hat{\theta}_T^*(\cdot)\}_{T\in N_+}$ be a sequence of continuous functions taking values in Θ^o such that $\hat{\theta}_T^*(Y_1,\cdots,Y_T)$ solves the problem

$$\min_{\theta\in\Theta} \quad \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)' W_T^* \left(\frac{1}{T}\sum_{t=1}^{T} g_t(Y_t,\theta)\right)$$

everywhere on Ω. It then follows that

$$\sqrt{T}\left[\hat{\theta}_T^*(Y_1,\cdots,Y_T)-\theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ (G_0'\Sigma^{-1}G_0)^{-1}\right],$$

and $\hat{\theta}_T^*(Y_1,\cdots,Y_T)$ is asymptotically efficient among the class of asymptotically normal GMM estimators.

Proof) This statement follows directly from theorem 3.9 except for the part where we assert that $\hat{\theta}_T^*(Y_1,\cdots,Y_T)$ is asymptotically efficient. This will be established if we can show that

$$(G_0'WG_0)^{-1}G_0'W\Sigma WG_0(G_0'WG_0)^{-1}-(G_0'\Sigma^{-1}G_0)^{-1}$$

is positive semidefinite for any positive definite $k\times k$ matrix W. In light of lemma 3.11, this follows if and only if

$$\mathcal{M}=G_0'\Sigma^{-1}G_0-(G_0'WG_0)(G_0'W\Sigma WG_0)^{-1}(G_0'WG_0)$$

is positive semidefinite. To this end, denote by $L\in\mathbb{R}^{k\times k}$ the Cholesky factor of Σ^{-1}, and note that $\Sigma=(LL')^{-1}=L^{-1\prime}L^{-1}$. Then, we can decompose $\mathcal{M}$ as

$$\begin{aligned}\mathcal{M} &= G_0'L\left[I_k-L^{-1}WG_0(G_0'WL^{-1\prime}L^{-1}WG_0)^{-1}G_0'WL^{-1\prime}\right]L'G_0\\ &= G_0'L\left[I_k-\mathcal{S}\left(\mathcal{S}'\mathcal{S}\right)^{-1}\mathcal{S}'\right]L'G_0,\end{aligned}$$

where we define $\mathcal{S}=L^{-1}WG_0$. It follows from standard results that the residual maker $M_{\mathcal{S}}=I_k-\mathcal{S}\left(\mathcal{S}'\mathcal{S}\right)^{-1}\mathcal{S}'$ is symmetric and idempotent; matrices of this form play a huge role in linear regression models, but we defer this discussion to later chapters. Now we can write "M as

$$\mathcal{M}=G_0'LM_{\mathcal{S}}L'G_0=(G_0'LM_{\mathcal{S}})(G_0'LM_{\mathcal{S}})',$$

so that $\mathcal{M}$ is clearly positive semidefinite. This completes our proof.

Q.E.D.

The above theorem shows us that the efficient GMM estimator can be computed by finding an appropriate sequence of positive definite valued random matrices $\{W_T^*\}_{T\in N_+}$ that converges in probability to Σ^{-1}. This is equivalent to searching for a consistent estimator of Σ. A natural estimator is

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T} g_t(Y_t, \hat{\theta}_T^*(Y_1, \cdots, Y_T)) g_t(Y_t, \hat{\theta}_T^*(Y_1, \cdots, Y_T))',$$

since $\hat{\theta}_T^*(Y_1, \cdots, Y_T)$ is consistent for θ_0. However, this results in a bit of circular logic; to construct $\hat{\Sigma}_T$, we need to use the efficient GMM estimator, but $\hat{\Sigma}_T$ is also used in the construction of the efficient GMM estimator. To circumvent this difficulty, we make use of the fact that any GMM estimator under assumptions G1 to G4 is consistent for θ_0. Therefore, a natural approach is to first obtain a GMM estimator under some arbitrary weighting matrix, using the estimator thus obtained to construct $\hat{\Sigma}_T$, and then using $\hat{\Sigma}_T$ to obtain the efficient GMM estimator. This is called feasible efficient GMM estimation.

We show below that $\hat{\Sigma}_T$ is consistent for Σ for any consistent estimator $\hat{\theta}_T$ plugged into θ in the above formula:

Theorem 3.13 (Feasible Efficient GMM Estimation)
Suppose that assumptions G1 to G7 hold, and let $\hat{\theta}_T$ be some consistent estimator of θ_0. Then, we have

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T} g_t(Y_t, \hat{\theta}_T) g_t(Y_t, \hat{\theta}_T)' \xrightarrow{p} \Sigma$$

$$\hat{G}_T = \frac{1}{T}\sum_{t=1}^{T} G_t(Y_t, \hat{\theta}_T) \xrightarrow{p} G_0.$$

Proof) In light of assumption G6, we need only show that

$$\Delta_T := \hat{\Sigma}_T - \frac{1}{T}\sum_{t=1}^{T} g_t(Y_t, \theta_0) g_t(Y_t, \theta_0)' \xrightarrow{p} O_{k\times k}.$$

To this end, we decompose Δ_T as

$$\begin{aligned}
\Delta_T &= \frac{1}{T}\sum_{t=1}^{T} \Big(g_t(Y_t, \hat{\theta}_T) g_t(Y_t, \hat{\theta}_T)' - g_t(Y_t, \theta_0) g_t(Y_t, \theta_0)' \Big) \\
&= \frac{1}{T}\sum_{t=1}^{T} \Big[\Big(g_t(Y_t, \hat{\theta}_T) - g_t(Y_t, \theta_0) \Big) \Big(g_t(Y_t, \hat{\theta}_T) - g_t(Y_t, \theta_0) \Big)' \\
&\quad + g_t(Y_t, \theta_0) \Big(g_t(Y_t, \hat{\theta}_T) - g_t(Y_t, \theta_0) \Big)' + \Big(g_t(Y_t, \hat{\theta}_T) - g_t(Y_t, \theta_0) \Big) g(Y_t, \theta_0)' \Big].
\end{aligned}$$

By assumption G3,

$$\left|g_t(Y_t,\hat{\theta}_T)-g_t(Y_t,\theta_0)\right| \le b_t(Y_t)\cdot h\left(\left|\hat{\theta}_T-\theta_0\right|\right).$$

for any $1 \le t \le T$. Since h is continuous at 0 with $h(0)=0$, we can see that, for any $\epsilon > 0$, there exists an $\eta > 0$ such that

$$h(d) < \epsilon$$

for any $0 \le d < \eta$. It follows that

$$\mathbb{P}\left(h(\left|\hat{\theta}_T-\theta_0\right|) > \epsilon\right) \le \mathbb{P}\left(\left|\hat{\theta}_T-\theta_0\right| > \eta\right)$$

for any $T \in N_+$. Taking $T \to \infty$ on both sides yields

$$\lim_{T\to\infty} \mathbb{P}\left(h(\left|\hat{\theta}_T-\theta_0\right|) > \epsilon\right) = 0,$$

so that

$$h(\left|\hat{\theta}_T-\theta_0\right|) \xrightarrow{p} 0.$$

Now we can bound $\|\Delta_T\|$ above as

$$\begin{aligned}
\|\Delta_T\| &\le \frac{1}{T}\sum_{t=1}^{T}\left|g_t(Y_t,\hat{\theta}_T)-g_t(Y_t,\theta_0)\right|^2 + \frac{2}{T}\sum_{t=1}^{T}|g_t(Y_t,\theta_0)|\cdot\left|g_t(Y_t,\hat{\theta}_T)-g_t(Y_t,\theta_0)\right| \\
&\le \left(\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)^2\right)h(\left|\hat{\theta}_T-\theta_0\right|)^2 + \left(\frac{2}{T}\sum_{t=1}^{T}|g_t(Y_t,\theta_0)|b_t(Y_t)\right)h(\left|\hat{\theta}_T-\theta_0\right|) \\
&\le \left(\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)^2\right)h(\left|\hat{\theta}_T-\theta_0\right|)^2 \\
&\quad + 2\left(\frac{1}{T}\sum_{t=1}^{T}|g_t(Y_t,\theta_0)|^2\right)^{\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)^2\right)^{\frac{1}{2}}h(\left|\hat{\theta}_T-\theta_0\right|),
\end{aligned}$$

where the final inequality follows from the Cauchy-Schwarz inequality for finite sums. By assumption,

$$\frac{1}{T}\sum_{t=1}^{T}b_t(Y_t)^2 = O_p(1) \quad \text{as } T\to\infty,$$

and

$$\frac{1}{T}\sum_{t=1}^{T}|g_t(Y_t,\theta_0)|^2 = \operatorname{tr}\left(\frac{1}{T}\sum_{t=1}^{T}g_t(Y_t,\theta_0)g_t(Y_t,\theta_0)'\right) = O_p(1)$$

as well. This indicates that

$$\|\Delta_T\| \leq O_p(1) \cdot o_p(1) + O_p(1) \cdot o_p(1) = o_p(1),$$

or in other words that $\Delta_T \xrightarrow{p} O_{k \times k}$, which completes the proof for the first statement.

The second statement can also be proven in a similar manner. Define

$$D_T = \frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \hat{\theta}_T) - \frac{1}{T} \sum_{t=1}^{T} G_t(Y_t, \theta_0),$$

so that

$$\|D_T\| \leq \frac{1}{T} \sum_{t=1}^{T} \Big\| G_t(Y_t, \hat{\theta}_T) - G_t(Y_t, \theta_0) \Big\|.$$

By assumption G5, we have

$$\Big\| G_t(Y_t, \hat{\theta}_T) - G_t(Y_t, \theta_0) \Big\| \leq B_t(Y_t) \cdot H(\Big| \hat{\theta}_T - \theta_0 \Big|),$$

where $H : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous at 0 and $H(0) = 0$. By the same reasoning as above with $h(\cdot)$,

$$H(\Big| \hat{\theta}_T - \theta_0 \Big|) \xrightarrow{p} 0.$$

Furthermore, by assumption $\frac{1}{T} \sum_{t=1}^{T} B_t(Y_t) = O_p(1)$. It follows that

$$\|D_T\| \leq \left(\frac{1}{T} \sum_{t=1}^{T} B_t(Y_t) \right) H(\Big| \hat{\theta}_T - \theta_0 \Big|) = O_p(1) o_p(1) = o_p(1),$$

which is exactly our desired result.

Q.E.D.

3.4.6 Inference and Weak Identification

Let $\tilde{\theta}_T(\cdot)$ be a continuous function on E^T such that$\tilde{\theta}_T(Y_1, \cdots, Y_T)$ is a consistent estimator for θ_0, and define

$$\hat{\Sigma}_T(\theta; y^T) = \frac{1}{T}\sum_{t=1}^{T} g_t(y_t, \theta) g_t(y_t, \theta)'$$

for any $y^T = (y_1, \cdots, y_T) \in E^T$ and $\theta \in \Theta$. We saw from the preceding result that $\hat{\Sigma}_T(\tilde{\theta}_T(Y^T); Y^T) \xrightarrow{p} \Sigma$; now we can define $\hat{\theta}^*_T(y_1, \cdots, y_T)$ as the solution to the problem

$$\min_{\theta \in \Theta} \quad \left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t, \theta)\right)' \hat{\Sigma}_T(\tilde{\theta}_T(y^T); y^T)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t, \theta)\right)$$

for any $y^T = (y_1, \cdots, y_T) \in E^T$. Under our assumptions, $\hat{\theta}^*_T(\cdot)$ is a continuous function on E^T and $\hat{\theta}^*_T(Y_1, \cdots, Y_T)$ is the feasible efficient GMM estimator with asymptotic distribution

$$\sqrt{T}\left[\hat{\theta}^*_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1}, (G_0'\Sigma^{-1}G_0)^{-1}\right],$$

given that $\hat{\theta}^*_T(\cdot)$ takes values in Θ^o. Our goal is to make inferences about the parameter θ_0 using this result.

Confidence Sets for θ_0

We focus on confidence sets as a means of conducting statistical inferences about θ_0. A natural means of constructing confidence sets around θ_0 is to make use of the Wald-type statistic defined as

$$\mathcal{W}_T(\theta; y^T) = T\left[\hat{\theta}^*_T(y^T) - \theta\right]' \hat{V}_T\left(\hat{\theta}^*_T(y^T); y^T\right)^{-1}\left[\hat{\theta}^*_T(y^T) - \theta\right]$$

for any $y^T \in E^T$ and $\theta \in \Theta$, where we define

$$\hat{G}_T(\theta; y^T) = \frac{1}{T}\sum_{t=1}^{T} G_t(y_t, \theta)$$

$$\hat{V}_T(\theta; y^T) = \left(\hat{G}_T(\theta; y^T)'\hat{\Sigma}_T(\theta; y^T)\hat{G}_T(\theta; y^T)\right)^{-1}$$

for any $(y^T, \theta) \in E^T \times \Theta^o$. Note that $\mathcal{W}_T(\cdot)$ is a continuous function on $\Theta \times E^T$ due to the continuity of $\hat{\theta}^*_T(\cdot)$, $G_t(\cdot)$ and $g_t(\cdot)$.

From the previous result, we can see that

$$\hat{V}_T\left(\hat{\theta}^*_T(y^T); y^T\right) \xrightarrow{p} (G_0'\Sigma^{-1}G_0)^{-1}$$

and therefore

$$\mathcal{W}_T(\theta_0; Y^T) \xrightarrow{d} \chi^2_m.$$

Given some $\alpha \in (0,1)$, letting $q_{1-\alpha}$ be the $1-\alpha$ quantile of the chi squared distribution with m degrees of freedom,

$$\lim_{T\to\infty} \mathbb{P}\left(\mathcal{W}_T(\theta_0; Y^T) \leq q_{1-\alpha}\right) = 1-\alpha$$

by the Portmanteau theorem and the fact that the chi-squared distribution function is everywhere continuous.

This suggests that we can define the correspondence $C_n : E^T \to \Theta$ as

$$C_n(y_1, \cdots, y_T) = \{\theta \in \Theta \mid \mathcal{W}_T(\theta; y_1, \cdots, y_T) \leq q_{1-\alpha}\}$$

for any $y^T \in E^T$. We will verify that C_n is a measurable correspondence and therefore that it is a well-defined confidence set:

1. C_n **is non-empty closed valued**
   Choose any $y^T \in E^T$. Because $\mathcal{W}_n(\cdot; y^T)$ is continuous on Θ, defining the mapping $w : \Theta \to \mathbb{R}_+$ as $w(\cdot) = \mathcal{W}_n(\cdot; y^T)$, it follows that $w(\cdot)$ is continuous and therefore that

   $$C_n(y^T) = w^{-1}([0, q_{1-\alpha}])$$

   is a closed subset of Θ.

   The non-emptiness of $C_n(y^T)$ follows from the fact that

   $$\mathcal{W}_n(\hat{\theta}^*_T(y^T); y^T) = 0,$$

   so that $\hat{\theta}^*_T(y^T) \in C_n(y^T)$.

2. $C_n^{-1}(\theta)$ **is closed valued**
   Now choose any $\theta \in \Theta$. Defining $w : E^T \to \mathbb{R}_+$ as $w(\cdot) = \mathcal{W}_n(\theta; \cdot)$, we can see that $w(\cdot)$ is a continuous function on E^T and therefore that

   $$C_n^{-1}(\theta) = \{y^T \in E^T \mid \theta \in C_n(y^T)\} = \{y^T \in E^T \mid w(y^T) \leq q_{1-\alpha}\} = w^{-1}([0, q_{1-\alpha}])$$

   is a closed subset of E^T.

By definition, C_n is a measurable correspondence and therefore a well-defined confidence set. Furthermore, C_n has the desired asymptotic coverage probability:

$$\mathbb{P}\left(\theta_0 \in C_n(Y^T)\right) = \mathbb{P}\left(\mathcal{W}_n(\theta_0; Y^T) \leq q_{1-\alpha}\right) \to 1-\alpha$$

as $T \to \infty$. If $m = 1$, then the sets $C_n(y^T)$ will take the form of a closed interval; on the other hand, if $m > 1$, then C_n will be ellipsoid-valued.

Weak Identification-Robust Confidence Sets

The approximate confidence set constructed above relies on the asymptotic result

$$\sqrt{T}\left[\hat{\theta}_T^*(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1}, (G_0'\Sigma^{-1}G_0)^{-1}\right],$$

which in turn relies on the fact that $G_0 \in \mathbb{R}^{k\times m}$ is a matrix of full rank m. A problem arises when G_0 is only barely full rank, that is, when its smallest non-zero eigenvalue is close to 0. There are means of formalizing the notion that G_0 is "almost" reduced rank; one way of going about this is through local-to-zero approximations, where we assume that the smallest non-zero eigenvalue of G_0 converges to 0 at a rate $\frac{1}{\sqrt{T}}$. This framework will be discussed further in the section on weak instruments, but for now, we will be content to just mention that, if G_0 is close to a reduced rank matrix, then $\hat{\theta}_T^*(Y_1, \cdots, Y_T)$ will converge very slowly to its asymptotic distribution, undermining the accuracy of the asymptotic approximation above. This is called the problem of weak identification.

Weak identification implies that the confidence set C_n constructed above will also perform poorly in finite samples, often producing coverage probabilities much smaller than $1 - \alpha$ for small T. As such, we want to construct confidence sets that are robust to weak identification, or in other words, confidence sets based on test statistics whose asymptotic distributions are valid approximations to their finite sample distributions even under weak identification.

Clearly, the key to this is to search for test statistics that do not involve G_0 or its sample analogue $\frac{1}{T}\sum_{t=1}^T \frac{\partial g_t(y_t,\theta)}{\partial \theta}$. It turns out that we have a very simple statistic at hand: namely, the efficient GMM objective function defined as

$$\hat{Q}_T(y^T \mid \theta) = -\left(\frac{1}{T}\sum_{t=1}^T g_t(y_t, \theta)\right)' \hat{\Sigma}_T\left(\tilde{\theta}_T(y^T); y^T\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T g_t(y_t, \theta)\right)$$

for any $(y^T \mid \theta) \in \Theta \times E^T$. From assumption G6 we can conclude that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T g_t(Y_t, \theta_0) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\ \Sigma\right],$$

so that

$$\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T g_t(Y_t, \theta_0)\right)' \hat{\Sigma}_T\left(\tilde{\theta}_T(Y^T); Y^T\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T g_t(Y_t, \theta_0)\right) \xrightarrow{d} \chi_k^2.$$

It follows that, if we define $\mathcal{J}_T : \Theta \times E^T \to \mathbb{R}_+$ as

$$\mathcal{J}_T(\theta; y^T) = -T \cdot \hat{Q}_T(y^T \mid \theta)$$

for any $(\theta, y^T) \in \Theta \times E^T$, then

$$\mathcal{J}_T(\theta_0; Y^T) \xrightarrow{d} \chi^2_k.$$

Note that the function $\hat{Q}_T(\cdot)$ does not make use of the functions $\{G_t(\cdot)\}_{t \in N_+}$ at all, and as such, the above asymptotic approximation is valid regardless of whether identification is weak or strong.

We can now define the correspondence $C_n^{AR} : E^T \to \Theta$ as

$$C_n^{AR}(y^T) = \{\theta \in \Theta \mid \mathcal{J}_T(\theta; y^T) \leq q_{1-\alpha}\},$$

where $q_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with k degrees of freedom. The measurability of C_n^{AR} follows from the continuity properties of $\mathcal{J}_T(\cdot)$ in the same manner as the measurability of C_n follows from the continuity of $\mathcal{W}_T(\cdot)$, so C_n^{AR} is a valid confidence set for θ_0. Furthermore, by the Portmanteau theorem we have

$$\lim_{T \to \infty} \mathbb{P}\left(\theta_0 \in C_n^{AR}(Y^T)\right) = \lim_{T \to \infty} \mathbb{P}\left(\mathcal{J}_T(\theta_0; Y^T) \leq q_{1-\alpha}\right) = 1 - \alpha,$$

so that C_n^{AR} has the desired approximate coverage rate.

The confidence set C_n^{AR} is called the weak identification-robust confidence set, or the Anderson-Rubin confidence set, after the statisticians who first developed this method of inference for the linear regression model with endogenous regressors.

3.4.7 Testing for Overidentification

In the case $k > m$, we have more moment conditions than parameters. This suggests that, given a subset of m moment conditions, the parameters continue to be identified. As such, given a subset of m moment conditions, we could test for the validity of the remaining $k-m$ moment conditions by first estimating the model using the m moment conditions that are chosen beforehand, and then investigating whether the sample analogue of the remaining moment conditions are close enough to 0. These types of tests are called tests for overidentification.

Formally, suppose that assumptions G1 to G7 hold with the exception of the identification condition, and that $k > m$. We want to test the null hypothesis

$$H_0 : \mathbb{E}\left[g_t(Y_t, \theta_0)\right] = O_{k\times 1}$$

against the alternative

$$H_1 : \mathbb{E}\left[g_t(Y_t, \theta_0)\right] = O_{k\times 1}.$$

Let $\{W_T\}_{T\in N_+}$ be a sequence of continuous $k \times k$ positive definite valued functions such that $W_T(Y^T) \xrightarrow{p} \Sigma^{-1}$ for some positive definite $k \times k$ matrix W as in assumption G2. Suppose that this convergence holds regardless of whether the null or alternative hypotheses hold true.

Let $\hat{\theta}_T(\cdot)$ be a continuous function from E^T to Θ^o such that $\hat{\theta}_T(y^T)$ is the solution to the problem

$$\min_{\theta\in\Theta} \quad \left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t,\theta)\right)' W_T(y^T) \left(\frac{1}{T}\sum_{t=1}^{T} g_t(y_t,\theta)\right)$$

for any $y^T = (y_1, \cdots, y_T) \in E^T$. We proved above that $\hat{\theta}_T(Y^T)$ has asymptotic distribution

$$\sqrt{T}\left[\hat{\theta}_T^(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1}, (G_0'\Sigma^{-1}G_0)^{-1}\right],$$

making it the efficient GMM estimator.

We will consider the J-statistic defined as follows. First, define $\mathcal{J}_T : \Theta \times E^T \to \mathbb{R}_+$ as

$$\mathcal{J}_T(\theta; y^T) = -T \cdot \hat{Q}_T(y^T \mid \theta) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(y_t,\theta)\right)' W_T(y^T) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(y_t,\theta)\right)$$

for any $(\theta, y^T) \in \Theta \times E^T$. The J-statistic is then given as

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\hat{\theta}_T(Y^T))\right)' W_T(Y^T) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\hat{\theta}_T(Y^T))\right).$$

Below we show that $\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T)$ has an asymptotic chi-squared distribution under the null, but this time with reduced degrees of freedom. The proof involves a simple second order Taylor expansion:

Theorem 3.14 (Asymptotic Properties of J Statistic)
Suppose that assumptions G1 to G7 hold with the exception of the identification condition, and that $k > m$. Let $\hat{\theta}_T : E^T \to \Theta^o$ be the continuous function constructed above, with the J statistic defined as

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \hat{\theta}_T(Y^T))\right)' W_T(Y^T) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \hat{\theta}_T(Y^T))\right).$$

Then, under the null H_0,

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T) \xrightarrow{d} \chi^2_{k-m}.$$

Proof) Suppose H_0 holds, so that the moment conditions are all valid. Then, assumptions G1 to G7 hold.

As suggested in the remark before the theorem, we start by making a second order Taylor expansion. For any $y^T \in E^T$, we can expand $\mathcal{J}_T(\theta_0; Y^T)$ around $\hat{\theta}_T(y^T)$ as

$$\mathcal{J}_T(\theta_0; y^T) = \mathcal{J}_T(\hat{\theta}_T(y^T); y^T) + \boldsymbol{\nabla}_\theta \mathcal{J}_T(\hat{\theta}_T(y^T); y^T)'(\theta_0 - \hat{\theta}_T(y^T))$$

$$+ \frac{1}{2}(\hat{\theta}_T(y^T) - \theta_0)' \frac{\partial^2 \mathcal{J}_T(\tilde{\theta}_T(y^T); y^T)}{\partial\theta\partial\theta'}(\hat{\theta}_T(y^T) - \theta_0),$$

where $\tilde{\theta}_T(\cdot)$ is a function such that

$$\left|\tilde{\theta}_T(y^t) - \theta_0\right| \leq \left|\hat{\theta}_T(y^t) - \theta_0\right|.$$

The measurability of $\tilde{\theta}_T(\cdot)$ follows from the usual application of Berge's theorem together with the measurable selection theorem. We now proceed by gradually simplifying this expression.

Dealing with the First Derivative

Since

$$\boldsymbol{\nabla}_\theta \mathcal{J}_T(\hat{\theta}_T(y^T); y^T) = 2\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} G_t(y_t, \hat{\theta}_T(y^T))\right)' W_T(y^T) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(y_t, \hat{\theta}_T(y^T))\right) = O_{k\times 1}$$

by the first order condition for minimization, and

$$\frac{1}{T}\frac{\partial^2 \mathcal{J}_T(\tilde{\theta}_T(y^T); y^T)}{\partial\theta\partial\theta'} = 2\left(\frac{1}{T}\sum_{t=1}^{T} G_t(y_t, \tilde{\theta}_T(y^T))\right)' W_T(y^T) \left(\frac{1}{T}\sum_{t=1}^{T} G_t(y_t, \tilde{\theta}_T(y^T))\right)$$

by definition, our expansion simplifies to

$$\mathcal{J}_T(\theta_0; y^T) = \mathcal{J}_T(\hat{\theta}_T(y^T); y^T) + \left[\sqrt{T}(\hat{\theta}_T(y^T) - \theta_0)\right]' G_0' \Sigma^{-1} G_0 \left[\sqrt{T}(\hat{\theta}_T(y^T) - \theta_0)\right]$$
$$+ \left[\sqrt{T}(\hat{\theta}_T(y^T) - \theta_0)\right]' \left[\frac{1}{T}\frac{\partial^2 \mathcal{J}_T(\tilde{\theta}_T(y^T); y^T)}{\partial\theta\partial\theta'} - G_0' \Sigma^{-1} G_0\right] \left[\sqrt{T}(\hat{\theta}_T(y^T) - \theta_0)\right]$$

for any $y^T \in E^T$.

Since $\hat{\theta}_T(Y^T) \xrightarrow{p} \theta_0$, we can see that $\tilde{\theta}_T(Y^T) \xrightarrow{p} \theta_0$ as well, and as such

$$\frac{1}{T}\frac{\partial^2 \mathcal{J}_T(\tilde{\theta}_T(Y^T); Y^T)}{\partial\theta\partial\theta'} \xrightarrow{p} G_0' \Sigma^{-1} G_0.$$

On the other hand, the asymptotic normality of the GMM estimator indicates that $\sqrt{T}(\hat{\theta}_T(y^T) - \theta_0) = O_p(1)$ as $T \to \infty$. As such,

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T)$$
$$= \mathcal{J}_T(\theta_0; Y^T) - \left[\sqrt{T}(\hat{\theta}_T(Y^T) - \theta_0)\right]' G_0' \Sigma^{-1} G_0 \left[\sqrt{T}(\hat{\theta}_T(Y^T) - \theta_0)\right] + o_p(1).$$

Dealing with $\mathcal{J}_T(\theta_0; Y^T)$

$\mathcal{J}_T(\theta_0; Y^T)$ can be expanded as

$$\mathcal{J}_T(\theta_0; Y^T) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)' W_T(Y^T) \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)$$
$$= \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)' \Sigma^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)$$
$$+ \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)' \left[W_T(Y^T) - \Sigma^{-1}\right] \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right);$$

since $W_T(Y^T) - \Sigma^{-1} = o_p(1)$ and $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0) = O_p(1)$ by assumption, it follows that

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)' \Sigma^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t, \theta_0)\right)$$
$$- \left[\sqrt{T}(\hat{\theta}_T(Y^T) - \theta_0)\right]' G_0' \Sigma^{-1} G_0 \left[\sqrt{T}(\hat{\theta}_T(Y^T) - \theta_0)\right] + o_p(1).$$

Dealing with $\sqrt{T}\left[\hat{\theta}_T(Y^T) - \theta_0\right]$

Finally, we use the fact that $\hat{\theta}_T(Y^T)$ is asymptotically linear in the following sense:

$$\sqrt{T}\left[\hat{\theta}_T(Y^T) - \theta_0\right] = -(G_0'\Sigma^{-1}G_0)^{-1}G_0'\Sigma^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right) + o_p(1).$$

This shows us that

$$\begin{aligned}\mathcal{J}_T(\hat{\theta}_T(Y^T);Y^T) = &\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right)' \Sigma^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right) \\ &- \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right)' \Sigma^{-1}G_0(G_0'\Sigma^{-1}G_0)^{-1}G_0'\Sigma^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right) + o_p(1) \\ = &\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right)' \left[\Sigma^{-1} - \Sigma^{-1}G_0(G_0'WG_0)^{-1}G_0'\Sigma^{-1}\right]\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right) + o_p(1).\end{aligned}$$

Limiting Distribution of the J Statistic

Defining the matrix

$$V = \Sigma^{-1} - \Sigma^{-1}G_0(G_0'WG_0)^{-1}G_0'\Sigma^{-1},$$

and letting $L \in \mathbb{R}^{k\times k}$ be the Cholesky factor of Σ^{-1}, we can see that

$$V = L\left[I_k - L'G_0(G_0'LL'G_0)^{-1}G_0'L\right]L'.$$

The matrix in the middle is the residual maker corresponding to the $k \times m$ matrix $\mathcal{S} = L'G_0$. This matrix, denoted $M_{\mathcal{S}}$, is known to be symmetric and idempotent. By results that will be proven later on (see lemma 4.2 and lemma 4.3), $M_{\mathcal{S}}$ has eigenvalues that equal either 0 or 1, and its rank is equal to its trace:

$$\operatorname{tr}(M_{\mathcal{S}}) = k - \operatorname{tr}((\mathcal{S}'\mathcal{S})^{-1}\mathcal{S}'\mathcal{S}) = k - m.$$

Furthermore, for any $Z \sim \mathcal{N}[0,1]$,

$$Z'M_{\mathcal{S}}Z \sim \chi^2_{k-m}.$$

The J statistic can now be written as

$$\mathcal{J}_T(\hat{\theta}_T(Y^T);Y^T) = \left(L'\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right)' M_{\mathcal{S}}\left(L'\frac{1}{\sqrt{T}}\sum_{t=1}^{T} g_t(Y_t,\theta_0)\right) + o_p(1),$$

where

$$L' \frac{1}{\sqrt{T}} \sum_{t=1}^{T} g_t(Y_t, \theta_0) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\ L' L^{-1\prime} L^{-1} L\right] = \mathcal{N}[0,1].$$

Therefore,

$$\mathcal{J}_T(\hat{\theta}_T(Y^T); Y^T) \xrightarrow{d} Z' M_{\mathcal{S}} Z \sim \chi^2_{k-m},$$

where $Z \sim \mathcal{N}[0,1]$.

Q.E.D.

3.5 The Asymptotics of Maximum Likelihood Methods

Another important class of models to which we can apply the general asymptotic results are those making use of maximum likelihood methods. The setting is as follows:

M1 : **The Data Generating Process**

Suppose $\{Y_t\}_{t\in\mathbb{Z}}$ is a sample taking values in the separable metric space (E,d) with corresponding Borel σ-algebra $\mathcal{E}$. As usual, let the parameter space Θ be a subset of the euclidean space $\mathbb{R}^m$, with corresponding Borel σ-algebra ϑ. Crucially, note that the support E of each sample observation does not depend on the parameter θ.

Let $\{\mu_T\}_{T\in N_+}$ be a sequence of statistical models corresponding to the sample $\{Y_t\}_{t\in\mathbb{Z}}$. Specifically, for any $T\in N_+$, μ_T is a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$ such that $\mu_T(\theta,\cdot)$ is the distribution of $(Y_1,\cdots,Y_T)$ when the parameter is $\theta\in\Theta$. We want to work with densities instead of distributions. To this end, let $\{\lambda_T\}_{T\in\mathbb{N}}$ and $\{f_T\}_{T\in N_+}$ be sequences of measures and functions such that, for any $T\in N_+$, λ_T is a σ-finite measure on $(E,\mathcal{E})^T$ and $f_T : E^T\times\Theta\to\mathbb{R}_{++}$ is a $\mathcal{E}^T\bigotimes\vartheta$-measurable function that satisfies

$$\int_A f_T(y\mid\theta)d\lambda_T(y)=\mu_T(\theta,A)$$

for any $(\theta,A)\in\Theta\times\mathcal{E}^T$. Usually, λ_T is taken to be the product of T Lebesgue and counting measures on E, which correspond to continuous and discrete random variables, respectively.

The log-likelihood is defined as the sequence of functions $\{l_T\}_{T\in N_+}$ such that l_T is a $\mathcal{E}^T\bigotimes\vartheta$-measurable function defined as

$$l_T(y\mid\theta)=\log f_T(y\mid\theta)\quad\text{for any }(y,\theta)\in E^T\times\Theta.$$

The maximum likelihood estimator (MLE) of θ is then defined as the maximizer of the log-likelihood with respect to the parameter θ.

3.5.1 The Gibbs Inequality and Motivating ML Estimation

Similarly to GMM estimation, there is a simple population moment inequality that motivates ML estimation. The inequality, attributed to Gibbs, tells us that the expected log likelihood is maximized at the true parameter value given that an appropriate identification condition holds. This serves as the inspiration for ML estimation, where we estimate the true parameter by maximizing the sample analogue of the expected log-likelihood.

Theorem 3.15 (Gibbs Inequality)
Suppose that $\{Y_t\}_{t\in\mathbb{Z}}$ is a sample generated according to the DGP in assumption M1. Then, for any $\theta,\theta_0 \in \Theta$ such that $l_T(\cdot \mid \theta)$ and $l_T(\cdot \mid \theta_0)$ are integrable with respect to $\mu_T(\theta_0,\cdot)$, we have

$$\int_{E^T} \log f_T(y \mid \theta)\mu_T(\theta_0, dy) \leq \int_{E^T} \log f_T(y \mid \theta_0)\mu_T(\theta_0, dy).$$

If, in addition, $f_T(y \mid \theta) = f_T(y \mid \theta_0)$ for λ_T-a.e. $y \in E^T$ implies $\theta = \theta_0$, then the above inequality holds as an equality if and only if $\theta = \theta_0$.

Proof) Choose any $\theta,\theta_0 \in \Theta$ such that $l_T(\cdot \mid \theta)$ and $l_T(\cdot \mid \theta_0)$ are integrable with respect to $\mu_T(\theta_0,\cdot)$. For any $T \in N_+$ and $\theta \in \Theta$, define the function $h : E^T \to \mathbb{R}_{++}$ as

$$h(y) = \frac{f_T(y \mid \theta)}{f_T(y \mid \theta_0)}$$

for any $y \in E^T$; h is clearly $\mathcal{E}^T$-measurable. Furthermore, the integral of h with respect to $\mu_T(\theta_0,\cdot)$ is well-defined by the positivity of h, and given as

$$\int_{E^T} h(y)\mu_T(\theta_0, dy) = \int_{E^T} f_T(y \mid \theta)d\lambda_T(y) = \mu_T(\theta, E^T) = 1.$$

Since the log function is strictly concave on $(0,+\infty)$, $\mu_T(\theta_0,\cdot)$ is a probability measure on $(E,\mathcal{E})^T$, and $\int_{E^T} h(y)\mu_T(\theta_0, dy) \in (0,+\infty)$, it follows from Jensen's inequality that

$$\int_{E^T} \log h(y)\mu_T(\theta_0, dy) \leq \log\left(\int_{E^T} h(y)\mu_T(\theta_0, dy)\right) = \log(1) = 0.$$

The term on the left hand side can be expanded as

$$\begin{aligned}\int_{E^T} \log h(y)\mu_T(\theta_0, dy) &= \int_{E^T} \left[\log f_T(y \mid \theta) - \log f_T(y \mid \theta_0)\right]\mu_T(\theta_0, dy) \\ &= \int_{E^T} \log f_T(y \mid \theta)\mu_T(\theta_0, dy) - \int_{E^T} \log f_T(y \mid \theta_0)\mu_T(\theta_0, dy),\end{aligned}$$

where the final equality follows from the linearity of integration for integrable functions (this is where the integrability assumption comes in). Therefore,

$$\int_{E^T} \log f_T(y \mid \theta)\mu_T(\theta_0, dy) \leq \int_{E^T} \log f_T(y \mid \theta_0)\mu_T(\theta_0, dy),$$

and we have established our desired inequality.

To prove that θ_0 uniquely maximizes the above integral under the additional assumption, note that the above inequality holds as an equality if and only if Jensen's inequality holds as an equality. Since the log function is strictly concave, this is only the case if the function $h(y)$ is equal to some constant $k \in (0, +\infty)$ for $\mu_T(\theta_0, \cdot)$-a.e. $y \in E^T$. By implication,

$$f_T(y \mid \theta) = k \cdot f_T(y \mid \theta_0)$$

for $\mu_T(\theta_0, \cdot)$-a.e. $y \in E^T$. Let the equality hold for any $y \in A \in \mathcal{E}^T$, where $\mu_T(\theta_0, A) = 1$. Then,

$$0 = \mu_T(\theta_0, A^c) = \int_{A^c} f_T(y \mid \theta_0) d\lambda_T(y)$$

implies by the vanishing property for non-negative functions that

$$f_T(y \mid \theta_0) \cdot I_{A^c}(y) = 0$$

for any $y \in B \in \mathcal{E}^T$, where $\lambda_T(B^c) = 0$. If $y \in B \cap A^c$, then $f_T(y \mid \theta_0) = 0$ for this y, which contradicts our assumption that f_T is always positive valued. Therefore, $B \cap A^c = \emptyset$, or equivalently, $A^c \subset B^c$, and by the monotonicity of measures,

$$0 \leq \lambda_T(A^c) \leq \lambda_T(B^c) = 0,$$

which implies that $\lambda_T(A^c) = 0$. We have just shown that

$$f_T(y \mid \theta) = k \cdot f_T(y \mid \theta_0)$$

for λ_T-a.e. $y \in E^T$.

Integrating both sides with respect to λ_T shows us that $k = 1$, so that

$$f_T(y \mid \theta) = f_T(y \mid \theta_0).$$

for λ_T-a.e. $y \in E^T$. By the identification assumption, this implies that $\theta = \theta_0$.

Q.E.D.

In terms of population moments, the above theorem tells us the following: letting $\theta_0 \in \Theta$ be the true parameter, if the expectation of $l_T(Y_1, \cdots, , Y_T \mid \theta)$ exists in $\mathbb{R}$ for any $T \in N_+$ and $\theta \in \Theta$,

and

$$f_T(y \mid \theta_1) = f_T(y \mid \theta_2) \quad \text{for } \lambda_T\text{-a.e. } y \in E^T$$

implies $\theta_1 = \theta_2$ for any $T \in N_+$, then θ_0 is the unique maximizer of the mapping

$$\theta \mapsto \mathbb{E}\left[\frac{1}{T} l_T(Y_1, \cdots, Y_T \mid \theta)\right].$$

Therefore, a reasonable way to estimate θ_0 is to find the value that maximizes the sample analogue of $\mathbb{E}\left[\frac{1}{T} l_T(Y_t, \cdots, Y_T \mid \theta)\right]$, which in this case is simply the integrand $\frac{1}{T} l_T(Y_t, \cdots, Y_T \mid \theta)$ itself. This shows us how the Gibbs inequality helps motivate maximum likelihood methods.

3.5.2 Identification, Existence and Consistency

Even aside from providing a motivation for ML estimation using population moments, the Gibbs inequality connects the intuitive notion of identification and its mathematical formalization. Intuitively, we would expect a parameter to be identified if there exists a one-to-one mapping between parameters and data, so that the same DGP cannot follow from different parameters. This is what the first condition implies. On the other hand, in the mathematical sense, the parameter θ is identified if some function of population moments is uniquely maximized at the true parameter; recall the identification condition for GMM estimation. The Gibbs inequality shows us that, if the intuitive notion of identificaiton holds, then so does the its mathematical counterpart.

One complication arises at this point. From the basic consistency theorem, we require that the true parameter uniquely maximizes some limit of the sequence of objective functions. In the Kullback-Leibler inequality, we have only shown that the true parameter uniquely maximizes the expectation of the objective function itself; we have yet to discuss any limit of a sequence of objective functions in any meaningful capacity.

As such, we must assume an identification condition that ensures the true parameter continues to be well-identified in the limit. To this end, we take inspiration from the Gibbs inequality, which in its simplest form states that the true parameter θ_0 is the unique maximizer of the mapping

$$\theta \mapsto \mathbb{E}\left[\frac{1}{T} l_T(Y_1, \cdots, Y_T \mid \theta)\right]$$

given the intuitive identification criterion. Suppose our sample is of size T and we have the realized observations $y \in E^T$. If a law of large numbers-type result holds, we can expect the realized sample moment $\frac{1}{T} l_T(y \mid \theta)$ to not be too far off from its population counterpart $\mathbb{E}\left[\frac{1}{T} l_T(Y_1, \cdots, Y_T \mid \theta)\right]$, and as such, we can expect $l_T(y \mid \theta_0)$ to be larger than $l_T(y \mid \theta)$ for any $\theta \neq \theta_0$. Since the log-likelihood represents the information we have on the paramter of interest given our data, it is reasonable to assume that, the larger the sample size, the more acutely we will be able to identify θ_0 from looking at the log-likelihood. In our context, this indicates that the difference

$l_T(y \mid \theta_0) - l_T(y \mid \theta)$ should become larger and larger as T grows larger. This suggests that the appropriate asymptotic identification condition is to require $l_T(y \mid \theta_0) - l_T(y \mid \theta)$ to diverge to $+\infty$ at the limit for any possible realized sample $y \in E^T$.

Below, we present the model assumptions that formalize the above intuition:

M2 : **Parameter Space**

Θ is a convex subset of $\mathbb{R}^m$. There exists a non-empty, convex and compact subset Θ_0 of Θ that is contained in the interior Θ^o of the parameter space and

$$\sup_{\theta \in \Theta_0^c} l_T(y \mid \theta) \leq \sup_{\theta \in \Theta_0} l_T(y \mid \theta)$$

for any $y \in E^T$. In addition, the true parameter θ_0 is contained in Θ_0.

M3 : **Regularity Conditions for Log-Likelihood**

For any $T \in N_+$, the function l_T is continuous on $E^T \times \Theta$, and that $l_T(y \mid \cdot)$ is a strictly concave function on Θ_0 for any $y \in E^T$.

M4 : **Asymptotic Identification Condition**

For any $\epsilon > 0$, such that $B(\theta_0, \epsilon)^c \cap \Theta_0 \neq \emptyset$, there exists a number $K(\epsilon) > 0$ such that

$$\lim_{T \to \infty} \mathbb{P}\left(\sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}} \Big[l_T(Y_1, \cdots, Y_T \mid \theta) - l_T(Y_1, \cdots, Y_T \mid \theta_0) \Big] \leq -K(\epsilon) \right) = 1,$$

where we put the supremum over an empty set equal to $-\infty$, as is customary. Note that, as long as $B(\theta_0, \epsilon)^c \cap \Theta_0 \neq \emptyset$, the mapping

$$y \mapsto \sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}} \Big[l_T(y \mid \theta) - l_T(y \mid \theta_0) \Big]$$

is equivalent to the mapping

$$y \mapsto \sup_{\theta \in B(\theta_0, \epsilon)^c \cap \Theta_0} \frac{1}{\sqrt{T}} \Big[l_T(y \mid \theta) - l_T(y \mid \theta_0) \Big]$$

in light of assumption M2. This mapping is continuous on E^T by Berge's maximum theorem because it is the value function corresponding to a maximization problem of a continuous function over a non-empty compact set, so that the statement above is well-defined.

The rate $\sqrt{T}$ at which the difference $l_T(Y_1, \cdots, Y_T \mid \theta) - l_T(Y_1, \cdots, Y_T \mid \theta_0)$ grows to $+\infty$ comes from the CLT, which shows us that, under certain regularity conditions, sums of random variables are $O_p(\sqrt{T})$, or diverge at the rate $\sqrt{T}$. Since log-likelihoods can be expressed as sums of random

variables if the data are i.i.d. or when they exhibit limited dependence, as in autoregressive models, the rate $\sqrt{T}$ of divergence applies for the log-likelihoods of most models of interest.

Below we show that the assumptions we made so far are sufficient to establish the existence and consistency of ML estimators:

Theorem 3.16 (Existence and Consistency of ML Estimator)
Suppose that assumptions M1 to M4 hold. Then, there exists a sequence of continuous functions $\{\hat{\theta}_T\}_{T\in N_+}$ such that each $\hat{\theta}_T$ maps E^T into $\Theta_0 \subset \Theta^o$ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is a solution to the problem

$$\max_{\theta\in\Theta} \quad l_T(Y_1, \cdots, Y_T \mid \theta)$$

everywhere on Ω. In addition, $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is consistent for the true parameter, that is,

$$\hat{\theta}_T(Y_1, \cdots, Y_T) \xrightarrow{p} \theta_0.$$

Proof) **Existence**

Establishing the existence of the ML estimator is routine at this point. Fix $T \in N_+$. Define $\hat{l}_T : E^T \to \Theta_0 \to \mathbb{R}$ as the restriction of l_T to $E^T \times \Theta_0$. Since Θ_0 is a compact subset of euclidean space, E^T is a metric space under the product metric d^T, and $\hat{l}_T$ is a continuous function on $E^T \times \Theta_0$, by theorem 3.1 there exists a Borel measurable function $\hat{\theta}_T : E^T \to \Theta_0$ such that

$$\hat{\theta}_T(y) \in \operatorname*{argmax}_{\theta\in\Theta_0} \hat{l}_T(y \mid \theta)$$

for any $y \in E^T$. Furthermore, because $\hat{l}_T(y \mid \cdot)$ is strictly concave on the non-empty convex set Θ_0 for any $y \in E^T$, theorem 3.1 tells us that $\hat{\theta}_T$ is a continuous function on E^T. It remains to establish that $\hat{\theta}_T(y)$ maximizes $l_T(y \mid \theta)$ with respect to θ over the entire parameter space, not just Θ_0.

Fix $y \in E^T$. By design,

$$l_T(y \mid \hat{\theta}_T(y)) = \sup_{\theta\in\Theta_0} l_T(y \mid \theta) \geq l_T(y \mid \theta')$$

for any $\theta' \in \Theta_0$. On the other hand, by assumption,

$$l_T(y \mid \hat{\theta}_T(y)) = \sup_{\theta\in\Theta_0} l_T(y \mid \theta) \geq \sup_{\theta\in\Theta_0^c} l_T(y \mid \theta) \geq l_T(y \mid \theta')$$

for any $\theta' \in \Theta_0'$ as well, so it follows that

$$l_T(y \mid \hat{\theta}_T(y)) \geq l_T(y \mid \theta)$$

for any $\theta \in \Theta_0$. Therefore,

$$l_T(y \mid \hat{\theta}_T(y)) = \max_{\theta \in \Theta} l_T(y \mid \theta).$$

Consistency

Unlike in the GMM caes, we cannot rely on the basic consistency theorem because we have not assume that the scaled log-likelihood converges uniformly in probability to some limiting objective function. However, thanks to the asymptotic identification condition, the proof for the onsistency of the ML estimators remains relatively simple. Choose any $\epsilon > 0$; we want to show that

$$\lim_{T \to \infty} \mathbb{P}\left(\left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right| > \epsilon\right) = 0,$$

or equivalently,

$$\lim_{T \to \infty} \mu_T\left(\theta_0,\ \{y \in E^T \mid \left|\hat{\theta}_T(y) - \theta_0\right| > \epsilon\}\right) = 0.$$

Suppose that $\left|\hat{\theta}_T(y) - \theta_0\right| > \epsilon$ for some $T \in N_+$ and $y \in E^T$. Then, because $\hat{\theta}_T(y) \in \Theta_0 \cap B(\theta_0, \epsilon)^c$, it follows that, by the asymptotic identification condition, there exists a $K > 0$ such that

$$\lim_{T \to \infty} \mu_T\left(\theta_0,\ \left\{y \in E^T \mid \sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}}\Big[l_T(y \mid \theta) - l_T(y \mid \theta_0)\Big] > -K\right\}\right) = 0.$$

By the definition of the ML estimator,

$$l_T(y \mid \hat{\theta}_T(y)) \geq l_T(y \mid \theta_0),$$

and because $\hat{\theta}_T(y) \in \Theta \setminus B(\theta_0, \epsilon)$, we have

$$-K < 0 \leq \frac{1}{\sqrt{T}}\Big[l_T(y \mid \hat{\theta}_T(y)) - l_T(y \mid \theta_0)\Big] \leq \sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}}\Big[l_T(y \mid \theta) - l_T(y \mid \theta_0)\Big].$$

Thus,

$$\{y \in E^T \mid \left|\hat{\theta}_T(y) - \theta_0\right| > \epsilon\} \subset \{y \in E^T \mid \sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}}\Big[l_T(y \mid \theta) - l_T(y \mid \theta_0)\Big] > -K\Big\},$$

and we have

$$\mu_T\left(\theta_0,\ \{y \in E^T \mid \left|\hat{\theta}_T(y) - \theta_0\right| > \epsilon\}\right)$$

$$\leq \mu_T \left(\theta_0,\ \left\{ y \in E^T \mid \sup_{\theta \in \Theta \setminus B(\theta_0, \epsilon)} \frac{1}{\sqrt{T}} \Big[l_T(y \mid \theta) - l_T(y \mid \theta_0) \Big] > -K \right\} \right).$$

Taking $T \to \infty$ on both sides shows us that

$$\lim_{T \to \infty} \mu_T \left(\theta_0,\ \{ y \in E^T \mid \left| \hat{\theta}_T(y) - \theta_0 \right| > \epsilon \} \right) = 0,$$

as we desired.

Q.E.D.

3.5.3 The Information Matrix and Cramer-Rao Lower Bound

To show that ML estimators are asymptotically normal, we must first introduce an object called the information matrix. First, let us make the following assumptions:

M5 : **Smoothness Properties of the Log-Likelihood**

For any $T \in N_+$ and $y \in E^T$, the function $f_T(y \mid \cdot)$ is twice continuously differentiable on Θ^o. It follows that the score is given by

$$s_T(y \mid \theta) := \frac{\partial l_T(y \mid \theta)}{\partial \theta} = \frac{1}{f_T(y \mid \theta)} \cdot \frac{\partial f_T(y \mid \theta)}{\partial \theta}$$

and the Hessian by

$$H_T(y \mid \theta) := \frac{\partial^2 l_T(y \mid \theta)}{\partial \theta \partial \theta'} = -\frac{1}{f_T(y \mid \theta)^2} \left(\frac{\partial f_T(y \mid \theta)}{\partial \theta} \right) \left(\frac{\partial f_T(y \mid \theta)}{\partial \theta} \right)' + \frac{1}{f_T(y \mid \theta)} \cdot \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'}$$

$$= -s_T(y \mid \theta) s_T(y \mid \theta)' + \frac{1}{f_T(y \mid \theta)} \cdot \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'}$$

for any $\theta \in \Theta^o$. We assume that both $s_T(\cdot \mid \theta) s_T(\cdot \mid \theta)'$ and $H_T(\cdot \mid \theta)$ are $\mu_T(\theta, \cdot)$-integrable for any $\theta \in \Theta^o$, and that they are both continuous on $E^T \times \Theta$.

M6 : **Conditions for Interchanging Integration and Differentiation**

We assume that, for any $T \in N_+$, the functions $f_T(\cdot \mid \theta)$ and $\frac{\partial f_T(\cdot \mid \theta)}{\partial \theta}$ are λ_T-integrable for any $\theta \in \Theta^o$, and that there exists some measurable non-negative and λ_T-integrable function $b_T : E^T \to \mathbb{R}_+$ such that

$$\left| \frac{\partial f_T(y \mid \theta)}{\partial \theta} \right|, \left\| \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'} \right\| \leq b_T(y)$$

for any $(y, \theta) \in E^T \times \Theta^o$.

As a consequence of assumptions M5 and M6, we can establish the following result:

Theorem 3.17 (Information Matrix Equality)
Suppose that assumptions M1 to M6 hold. Then,

$$\int_{E^T} s_T(y \mid \theta)\mu_T(\theta, dy) = O_{m\times 1}$$

and

$$\int_{E^T} s_T(y \mid \theta)s_T(y \mid \theta)'\mu_T(\theta, dy) = -\int_{E^T} H_T(y \mid \theta)\mu_T(\theta, dy).$$

for any $\theta \in \Theta^o$.

Proof) Choose any $\theta \in \Theta^o$. The two results follow easily from the Leibniz integral rule, which holds for f_T and its gradient with respect to θ by the conditions in assumption M6.

Formally, by the Leibniz integral rule, we can interchange the order of differentiatoin and integration with respect to λ_T for f_T and its gradient. Since

$$\int_{E^T} f_T(y \mid \theta)d\lambda_T(y) = \mu_T(\theta, E^T) = 1,$$

it follows that

$$\int_{E^T} \frac{\partial f_T(y \mid \theta)}{\partial \theta} d\lambda_T(y) = \frac{\partial}{\partial \theta}\left(\int_{E^T} f_T(y \mid \theta)d\lambda_T(y)\right) = O_{m\times 1}.$$

Likewise,

$$\int_{E^T} \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'} d\lambda_T(y) = \frac{\partial}{\partial \theta}\left(\int_{E^T} \frac{\partial f_T(y \mid \theta)}{\partial \theta} d\lambda_T(y)\right) = O_{m\times m}.$$

Now note that

$$\int_{E^T} s_T(y \mid \theta)\mu_T(\theta, dy) = \int_{E^T} s_T(y \mid \theta)\cdot f_T(y \mid \theta)\; d\lambda_T(y) = \int_{E^T} \frac{\partial f_T(y \mid \theta)}{\partial \theta} d\lambda_T(y) = O_{m\times 1}.$$

To show that the second equality holds, we need only note that

$$\begin{aligned}
\int_{E^T} H_T(y \mid \theta)\mu_T(\theta, dy) &= \int_{E^T} H_T(y \mid \theta)\cdot f_T(y \mid \theta)\; d\lambda_T(y) \\
&= \int_{E^T} \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'} d\lambda_T(y) - \int_{E^T} s_T(y \mid \theta)s_T(y \mid \theta)'\cdot f_T(y \mid \theta)\; d\lambda_T(y) \\
&= -\int_{E^T} s_T(y \mid \theta)s_T(y \mid \theta)'\mu_T(\theta, dy).
\end{aligned}$$

Q.E.D.

The two equations have a very intuitive interpretation. Note that they imply

$$\mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0)\right] := \int_{E^T} s_T(y \mid \theta_0)\mu_T(\theta_0, dy) = O_{m\times 1}$$

and

$$\begin{aligned}-\mathbb{E}\left[H_T(Y_1,\cdots,Y_T \mid \theta_0)\right] &:= -\int_{E^T} H_T(y \mid \theta_0)\mu_T(\theta_0, dy) \\ &= \mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0)s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] := \int_{E^T} s_T(y \mid \theta_0)s_T(y \mid \theta_0)'\mu_T(\theta_0, dy).\end{aligned}$$

The first equation tells us that the score function has mean zero under the true parameter, so that $\{s_T(Y_1,\cdots,Y_T \mid \theta_0)\}_{T\in N_+}$ is a zero mean process. Subsequently, the second equation tells us that the variance of the score function, which is

$$\mathrm{Var}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0)\right] = \mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0)s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right]$$

in light of the zero mean property, is equivalent to the expectation of the Hessian with its sign reversed.

As a function of θ, the variance of the score function can be denoted

$$\mathbb{I}_T(\theta) = \mathrm{Var}\left[s_T(Y_1,\cdots,Y_T \mid \theta)\right]$$

for any $\theta \in \Theta^o$. We define the information matrix as this variance evaluated at θ_0, that is, $\mathbb{I}_T(\theta_0)$ is the information matrix under T sample observations. The equation

$$\mathbb{I}_T(\theta_0) = -\mathbb{E}\left[H_T(Y_1,\cdots,Y_T \mid \theta_0)\right]$$

is called the information matrix equality.

Note in addition that, if the ML estimator $\hat{\theta}_T(Y_1,\cdots,Y_T)$ takes values in the interior Θ^o of the parameter space, the first order conditions for maximization tells us that

$$s_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) = \frac{\partial f_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots Y_T))}{\partial\theta} = O_{m\times 1}.$$

It follows that, in this case,

$$\mathbb{I}_T\left[\hat{\theta}_T(Y_1,\cdots,Y_T)\right] = \int_{E^T} s_T\left[y \mid \hat{\theta}_T(Y_1,\cdots,Y_T)\right] s_T\left[y \mid \hat{\theta}_T(Y_1,\cdots,Y_T)\right]' \mu_T(\theta_0, dy).$$

The reason $\mathbb{I}_T(\theta_0)$ is called the information matrix will be made clear once we show how it is related to the Cramer-Rao lower bound (CRLB). Simply put, the CRLB is the lower bound of the variance of an unbiased estimator of θ. Formally, given a sample of size T, an unbiased estimator of θ is defined as a function $W : E^T \to \Theta$ such that

$$\int_{E^T} W(y)\mu_T(\theta, dy) = \theta$$

for any $\theta \in \Theta$. In particular, this implies that

$$\mathbb{E}\left[W(Y_1, \cdots, Y_T)\right] := \int_{E^T} W(y)\mu_T(\theta_0, dy) = \theta_0.$$

In this sense, unbiasedness is a finite sample property of estimators that corresponds to the large sample property of consistency. Whereas consistency requires a sequence of estimators to converge to the true parameter, unbiasedness requires that their means are equal to the true parameter. In this sense, a consistent estimator can be viewed as an asymptotically unbiased estimator.

The theorem below shows us that the variance of an unbiased estimator of θ given a sample of size T is bounded below by the inverse of the information matrix, that is,

$$\mathrm{Var}\left[W(Y_1, \cdots, Y_T)\right] - \mathbb{I}_T(\theta_0)^{-1}$$

is positive semidefinite, given that the $\mathbb{I}_T(\theta_0)$ is positive definite.

Theorem 3.18 (Cramer-Rao Bound)
Suppose that assumptions M1 to M6 hold, and that the T-observation information matrix $\mathbb{I}_T(\theta_0)$ is positive definite. Then, for any unbiased estimator $W : E^T \to \Theta$ of θ such that

$$\int_{E^T} |W(y)b_T(y)| d\lambda_T(y) < +\infty,$$

and $W(Y_1, \cdots, Y_T)$ has finite second moments, the difference

$$\mathrm{Var}\left[W(Y_1, \cdots, Y_T)\right] - \mathbb{I}_T(\theta_0)^{-1}$$

is positive semidefinite.

Proof) This result follows easily by considering the random vector

$$X_T = (W(Y_1, \cdots, Y_T) - \theta_0) - \mathbb{I}_T(\theta_0)^{-1} \cdot s_T(Y_1, \cdots, Y_T \mid \theta_0).$$

Note that

$$\mathbb{E}\left[X_T X_T'\right] = \mathbb{E}\left[(W(Y_1, \cdots, Y_T) - \theta_0)\,(W(Y_1, \cdots, Y_T) - \theta_0)'\right]$$

$$
\begin{aligned}
&- \mathbb{E}\left[(W(Y_1,\cdots,Y_T)-\theta_0)\, s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] \cdot \mathbb{I}_T(\theta_0)^{-1} \\
&- \mathbb{I}_T(\theta_0)^{-1} \cdot \mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0)\,(W(Y_1,\cdots,Y_T)-\theta_0)'\right] \\
&+ \mathbb{I}_T(\theta_0)^{-1} \cdot \mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0) s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] \cdot \mathbb{I}_T(\theta_0)^{-1}.
\end{aligned}
$$

We study each term in turn.

The first term is just the variance of $W(Y_1,\cdots,Y_T)$ in light of its unbiasedness.Likewise, the last term can be reduced to

$$
\mathbb{I}_T(\theta_0)^{-1} \cdot \mathbb{E}\left[s_T(Y_1,\cdots,Y_T \mid \theta_0) s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] \cdot \mathbb{I}_T(\theta_0)^{-1} = \mathbb{I}_T(\theta_0)^{-1}.
$$

Finally, to deal with the second and third terms, note that

$$
\begin{aligned}
\mathbb{E}\left[(W(Y_1,\cdots,Y_T)-\theta_0)\, s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] &= \mathbb{E}\left[W(Y_1,\cdots,Y_T) s_T(Y_1,\cdots,Y_T \mid \theta_0)'\right] \\
&= \int_{E^T} W(y) s_T(y \mid \theta_0) \mu_T(\theta_0, dy) \\
&= \int_{E^T} W(y) \left[\frac{1}{f_T(y \mid \theta_0)} \frac{\partial f_T(y \mid \theta_0)}{\partial \theta}\right]' f_T(y \mid \theta_0)\, d\lambda_T(y) \\
&= \int_{E^T} W(y) \left(\frac{\partial f_T(y \mid \theta_0)}{\partial \theta}\right)' d\lambda_T(y).
\end{aligned}
$$

Defining $V : E^T \times \Theta \to \mathbb{R}^m$ as

$$
V(y,\theta) = W(y) f_T(y \mid \theta)
$$

for any $(y,\theta) \in E^T \times \Theta$,

$$
\int_{E^T} V(y,\theta) d\lambda_T(y) = \int_{E^T} W(y) \mu_T(\theta, dy) = \theta
$$

for any $\theta \in \Theta$ by the unbiasedness of W. Furthermore,

$$
\frac{\partial V(y,\theta)}{\partial \theta} = W(y) \left(\frac{\partial f_T(y \mid \theta)}{\partial \theta}\right)'
$$

for any $(y,\theta) \in E^T \times \Theta^o$, with

$$
\left\| \frac{\partial V(y,\theta)}{\partial \theta} \right\| \leq |W(y)| \cdot \left| \frac{\partial f_T(y \mid \theta)}{\partial \theta} \right| \leq |W(y)| \cdot b_T(y).
$$

By assumption, we have

$$
\int_{E^T} |W(y)| \cdot b_T(y) d\lambda_T(y) < +\infty,
$$

so the Leibniz integral rule tells us that, for any $\theta \in \Theta^o$,

$$\int_{E^T} \frac{\partial V(y,\theta)}{\partial \theta} d\lambda_T(y) = \frac{\partial}{\partial \theta} \left(\int_{E^T} V(y,\theta) d\lambda_T(y) \right) = \frac{\partial}{\partial \theta} \theta = I_m.$$

By implication,

$$\int_{E^T} W(y) \left(\frac{\partial f_T(y \mid \theta_0)}{\partial \theta} \right)' d\lambda_T(y) = \int_{E^T} \frac{\partial V(y,\theta_0)}{\partial \theta} d\lambda_T(y) = I_m,$$

and thus

$$\mathbb{E}\left[(W(Y_1, \cdots, Y_T) - \theta_0)\, s_T(Y_1, \cdots, Y_T \mid \theta_0)' \right] = I_m.$$

It follows that

$$\mathbb{E}\left[X_T X_T' \right] = \mathrm{Var}\left[W(Y_1, \cdots, Y_T) \right] - \mathbb{I}_T(\theta_0)^{-1}.$$

Since the left hand side is positive semidefinite by design, the right hand side term must also be positive semidefinite; this concludes our proof.

Q.E.D.

The theorem above shows us that the variance of an unbiased estimator of θ cannot be smaller than $\mathbb{I}_T(\theta_0)^{-1}$, so that the "larger" $\mathbb{I}_T(\theta_0)$ is, the more precisely we can estimate θ. This is why we can view $\mathbb{I}_T(\theta_0)$ as representing the information contained in a sample of size T; the more information there is, the more precisely we can potentially estimate the parameter of interest.

It turns out that ML estimators are interesting because they satisfy an asymptotic version of the Cramer-Rao bound. Specifically, under some additional assumptions on the asymptotic behavior of the scaled sample information matrix $\frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta)$, we can conclude that

$$\hat{\theta}_T(Y_1, \cdots, Y_T) \approx \mathcal{N}\left[\theta_0,\, \mathbb{I}_T(\theta_0)^{-1} \right],$$

that is, the ML estimator of θ is approximately normally distributed with mean θ_0 and variance equal to the CRLB $\mathbb{I}_T(\theta_0)^{-1}$.

Given the importance of the information matrix, we want to find some way to estimate it using the sample data given to us. To this end, we note that the T-observation information matrix $\mathbb{I}_T(\theta_0)$ is the expectation of the negative Hessian evaluated at θ_0: $-H_T(Y_1, \cdots, Y_T \mid \theta_0)$. Usually, the demeaned process

$$-H_T(Y_1, \cdots, Y_T \mid \theta_0) - \mathbb{I}_T(\theta_0)$$

ends up diverging, since the variation of $H_T(Y_1, \cdots, Y_T \mid \theta_0)$ grows as the sample size T grows larger. To make sure that the above quantity converges in probability to its mean of zero, we

must thus scale it appropriately; in many cases it turns out that it converges in probability at a rate T, owing to the law of large numbers. If, as in these cases, we have

$$\frac{1}{T}\left[-H_T(Y_1,\cdots,Y_T \mid \theta_0) - \mathbb{I}_T(\theta_0)\right] \xrightarrow{p} O_{m\times m},$$

then $-\frac{1}{T}H_T(Y_1,\cdots,Y_T \mid \theta_0)$ is a consistent estimator of the scaled information matrix $\frac{1}{T}\mathbb{I}_T(\theta_0)$. Since we do not know θ_0, we can think of replacing θ_0 with the consistent estimator $\hat{\theta}_T(Y_1,\cdots,Y_T)$. Formally, define

$$\hat{\mathbb{I}}_T(y) = -H_T(y \mid \hat{\theta}_T(y))$$

for any $y \in E^T$, where the Hessian at $\hat{\theta}_T(y)$ is well-defined because $\hat{\theta}_T$ takes values in $\Theta_0 \subset \Theta^o$. $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)$ is now our go-to estimator of the information matrix.

3.5.4 Asymptotic Distribution of the ML Estimator

To establish the asymptotic normality of the ML estimator, we must first make some assumptions. First, we must establish that the scaled sample information matrix is appropriately bounded in probability as the sample size grows to infinity, and we must also ensure that the Hessian is appropriately smooth. Finally, we require a central limit theorem for the score function. Formally, we make the following assumptions:

M7 : **Boundedness of Scaled Sample Information Matrix**

Let $\hat{\mathbb{I}}_T : E^T \to \mathbb{R}^{m \times m}$ be defined as above. Let $q_{T,min}(y)$ and $q_{T,max}(y)$ be the minimum and maximum eigenvalues of the positive definite matrix $\frac{1}{T}\hat{\mathbb{I}}_T(y)$ for any $y \in E^T$. Since the mapping

$$y \mapsto H_T(y \mid \hat{\theta}_T(y))$$

is $\mathcal{E}^T$-measurable and ordered eigenvalues are continuous functions of the corresponding symmetric matrices, the mappings

$$y \mapsto q_{T,min}(y), \quad y \mapsto q_{T,max}(y)$$

are non-negative and $\mathcal{E}^T$-measurable.

We assume that there exist some $0 < \rho_{\min} < \rho_{max} < +\infty$ such that

$$\lim_{T\to\infty} \mathbb{P}\left(\rho_{\min} \leq q_{T,min}(Y_1, \cdots, Y_T) \leq q_{T,max}(Y_1, \cdots, Y_T) \leq \rho_{max}\right) = 1.$$

M8 : **Smoothness of Hessian**

For any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that $\bar{B}(\theta_0, \delta(\epsilon)) \subset \Theta$ (since θ_0 is in the interior of Θ) and

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{\theta \in \bar{B}(\theta_0,\delta(\epsilon))} \left\| \frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T}H_T(Y_1, \cdots, Y_T \mid \theta) \right\| \leq \epsilon \right) = 1,$$

where $\bar{B}(\theta_0, \delta(\epsilon))$ is the closed $\delta(\epsilon)$-ball around θ_0. Note that the mapping

$$y \mapsto \sup_{\theta \in \bar{B}(\theta_0,\delta(\epsilon))} \left\| \frac{1}{T}\hat{\mathbb{I}}_T(y) + H_T(y \mid \theta) \right\|$$

is continuous by Berge's maximum theorem because it is the value function corresponding to the problem of maximizing a continuous function over a compact set, so that the above limit statement is well-defined.

M9 : **Asymptotic Normality of Sclaed Score**

The score function, appropriately scaled, is asymptotically normal:

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}\left[\frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T\mid\theta_0)\right]\xrightarrow{d}\mathcal{N}[O_{m\times 1},\,I_m],$$

where $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}$ is the Cholesky factor of the positive definite-valued random matrix $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)$.

Note that the scaling of the score function makes sense in light of the information matrix equality. For any $T\in N_+$, the variance of $\frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T\mid\theta_0)$ is given as

$$\mathbb{E}\left[\frac{1}{T}s_T(Y_1,\cdots,Y_T\mid\theta_0)s_T(Y_1,\cdots,Y_T\mid\theta_0)'\right]=\frac{1}{T}\mathbb{I}_T(\theta_0),$$

the scaled T-observation information matrix. By assumption,

$$\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)+\frac{1}{T}H_T(Y_1,\cdots,Y_T\mid\theta_0)\xrightarrow{p}O_{m\times m},$$

and by a law of large numbers-type argument, $-\frac{1}{T}H_T(Y_1,\cdots,Y_T\mid\theta_0)$ should be close to its expectation

$$\mathbb{E}\left[-\frac{1}{T}H_T(Y_1,\cdots,Y_T\mid\theta_0)\right]=\frac{1}{T}\mathbb{I}_T(\theta_0),$$

where the equality is precisely the information matrix equality. Therefore, for large enough T the scaled sample information matrix $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)$ is similar to the variance $\frac{1}{T}\mathbb{I}_T(\theta_0)$ of the scaled score function $\frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T\mid\theta_0)$, which motivates our decision to assume that

$$\frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T\mid\theta_0)\approx\mathcal{N}\left[O_{m\times 1},\,\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right]$$

in large samples.

We are now in a position to show that the ML estimator is asymptotically normal with asymptotic variance given as the inverse of the limiting information matrix $\mathcal{I}(\theta_0)$. First, we present a lemma that is similar in spirit to lemma 3.6, which concerns inversion at the limit.

Lemma 3.19 Let $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq m > 0$. Then,

$$\|A\| \leq \sqrt{n}\cdot\lambda_1, \quad \left\|A^{-1}\right\| \leq \frac{\sqrt{n}}{\lambda_n} \leq \frac{\sqrt{n}}{m}.$$

In addition, if $B \in \mathbb{R}^{n\times n}$ a square matrix such that

$$\|A-B\| \leq \alpha\frac{m}{\sqrt{n}}$$

for some $\alpha \in (0,1)$, then B is non-singular and

$$\left\|A^{-1}-B^{-1}\right\| \leq \frac{\sqrt{n}}{m}\cdot\frac{\alpha}{1-\alpha}.$$

Proof) Let A be a positive definite $n\times n$ matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq m > 0$. Then, letting PDP' be the orthogonal eigendecomposition of A, we have

$$\|A\|^2 = \mathrm{tr}(A'A) = \mathrm{tr}\left(PD^2P'\right) = \mathrm{tr}(D^2) = \sum_{i=1}^{n}\lambda_i^2.$$

Similalry, because $A^{-1} = PD^{-1}P'$, we have

$$\left\|A^{-1}\right\|^2 = \mathrm{tr}(D^{-2}) = \sum_{i=1}^{n}\frac{1}{\lambda_i^2}.$$

It follows that

$$\|A\| \leq \left(n\cdot\lambda_1^2\right)^{\frac{1}{2}} = \sqrt{n}\cdot\lambda_1$$

$$\left\|A^{-1}\right\| \leq \frac{\sqrt{n}}{\lambda_n} \leq \frac{\sqrt{n}}{m}.$$

Recall that, for any $n\times n$ matrix B such that

$$\|A-B\| < \frac{1}{\|A^{-1}\|},$$

B is also non-singular and its inverse B^{-1} satisfies the inequality

$$\left\|B^{-1}-A^{-1}\right\| \leq \frac{\left\|A^{-1}\right\|^2\cdot\|A-B\|}{1-\|A^{-1}\|\cdot\|A-B\|}.$$

Now let B be an $n\times n$ matrix such that

$$\|A-B\| < \alpha\frac{m}{\sqrt{n}}.$$

Then, because

$$\|A-B\| < \alpha \cdot \frac{1}{\|A^{-1}\|} < \frac{1}{\|A^{-1}\|},$$

B is non-singular. In addition,

$$\begin{aligned}\left\|A^{-1}-B^{-1}\right\| &\leq \frac{\left\|A^{-1}\right\|^2 \cdot \|A-B\|}{1-\|A^{-1}\| \cdot \|A-B\|} \\ &\leq \frac{\sqrt{n}}{m} \cdot \left(\frac{\|A^{-1}\| \cdot \|A-B\|}{1-\|A^{-1}\| \cdot \|A-B\|}\right).\end{aligned}$$

Since the mapping $x \mapsto \frac{x}{1-x}$ is increasing on $[0,1)$, and

$$\left\|A^{-1}\right\| \cdot \|A-B\| \leq \alpha,$$

we can see that

$$\left\|A^{-1}-B^{-1}\right\| \leq \frac{\sqrt{n}}{m} \cdot \frac{\alpha}{1-\alpha}.$$

Q.E.D.

Lemma 3.20 (Generalized Inversion at the Limit)
Suppose $\{X_T\}_{T\in N_+}$ and $\{Y_T\}_{T\in N_+}$ are sequences of n-dimensional random vectors, $\{W_T\}_{T\in N_+}$ a sequence of $n\times n$ random (not necessarily square) matrices and $\{H_T\}_{T\in N_+}$ a sequence of $n\times n$ real symmetric matrices such that

$$\|W_T - H_T\| \leq \alpha\frac{m}{\sqrt{n}} \quad \text{and} \quad 0 < m \leq q_{T,min}$$

with probability approaching 1 as $T\to\infty$, where $\alpha\in(0,1)$ and $q_{T,min}$ is the smallest eigenvalue of H_T for any $T\in N_+$. If $Y_T = W_T X_T$ and Y_T is $O_p(1)$ as $T\to\infty$, then so is X_T.

Proof) Define the event

$$A_T = \left\{\omega\in\Omega \mid \|W_T(\omega) - H_T(\omega)\| \leq \alpha\frac{m}{\sqrt{n}} \quad \text{and} \quad 0 < m \leq q_{T,min}(\omega)\right\} \in \mathcal{H}.$$

By assumption,

$$\mathbb{P}(A_T)\to 1 \quad \text{as } T\to\infty.$$

Furthemore, for any $\omega\in A_T$, since

$$\|W_T(\omega) - H_T(\omega)\| \leq \alpha\frac{m}{\sqrt{n}}$$

and $H_T(\omega)$ is a positive definite matrix whose smallest eigenvalue is bounded below by $m>0$, lemma 3.14 tells us that

$$\left\|H_T(\omega)^{-1}\right\| \leq \frac{\sqrt{n}}{m},$$

$W_T(\omega)$ is non-singular, and

$$\left\|W_T(\omega)^{-1} - H_T(\omega)^{-1}\right\| \leq \frac{\sqrt{n}}{m}\cdot\frac{\alpha}{1-\alpha}.$$

In particular,

$$\left\|W_T(\omega)^{-1}\right\| \leq \frac{\sqrt{n}}{m}\cdot\frac{\alpha}{1-\alpha} + \left\|H_T(\omega)^{-1}\right\| \leq \frac{\sqrt{n}}{m}\cdot\frac{1}{1-\alpha}.$$

Finally, since $Y_T(\omega) = W_T(\omega)X_T(\omega)$, we have

$$\begin{aligned}|X_T(\omega)| = \left|W_T(\omega)^{-1}W_T(\omega)X_T(\omega)\right| &= \left|W_T(\omega)^{-1}Y_T(\omega)\right| \\ &\leq \left\|W_T(\omega)^{-1}\right\|\cdot|Y_T(\omega)| \leq \left(\frac{\sqrt{n}}{m}\cdot\frac{1}{1-\alpha}\right)|Y_T(\omega)|.\end{aligned}$$

By assumption, Y_T is $O_p(1)$, so that, for any $\epsilon > 0$, there exists a $K > 0$ and an $N \in N_+$ such that

$$\mathbb{P}(|Y_T| > K) < \frac{\epsilon}{2} \quad \text{and} \quad \mathbb{P}(A_T^c) < \frac{\epsilon}{2}$$

for any $T \geq N$. Define

$$K_1 = \left(\frac{\sqrt{n}}{m} \cdot \frac{1}{1-\alpha} \right) K > 0.$$

Then, for any $T \geq N$,

$$\mathbb{P}(|X_T| > K_1) = \mathbb{P}(\{|X_T| > K_1\} \cap A_T) + \mathbb{P}(\{|X_T| > K_1\} \cap A_T^c)$$

$$\leq \mathbb{P}(\{|X_T| > K_1\} \cap A_T) + \mathbb{P}(A_T^c).$$

For any $\omega \in A_T$, $K_1 < |X_T(\omega)|$ implies

$$K_1 < \left(\frac{\sqrt{n}}{m} \cdot \frac{1}{1-\alpha} \right) |Y_T(\omega)|,$$

or equivalently, $|Y_T(\omega)| > K$, so that

$$\mathbb{P}(|X_T| > K_1) \leq \mathbb{P}(\{|Y_T| > K\} \cap A_T) + \mathbb{P}(A_T^c) \leq \mathbb{P}(|Y_T| > K) + \mathbb{P}(A_T^c) < \epsilon.$$

We can find such a $K_1 > 0$ and $N \in N_+$ for any $\epsilon > 0$, so by definition, X_T is $O_p(1)$ as $T \to \infty$.

Q.E.D.

Note that this result contains, as a special case, the inversion at the limit result proved in lemma 3.6 when the limit is positive definite. To see this, note that lemma 3.6 can be formulated in terms of the hypotheses of lemma 3.15 by putting $H_T = W$ for any $T \in N_+$, where W is the positive definite probability limit of the sequence $\{W_T\}_{T \in N_+}$.

Theorem 3.21 (Asymptotic Normality of the ML Estimator)
Suppose that assumptions M1 to M9 hold, and let $\{\hat{\theta}_T\}_{T\in N_+}$ be a sequence of continuous functions such that $\hat{\theta}_T : E^T \to \Theta_0$ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the T-observation ML estimator of θ. Then,

$$\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 = O_p(T^{-1/2}) \quad \text{as } T \to \infty$$

with asymptotic distribution given as

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right].$$

Proof) The proof is similar, but not identical to, the proof of the basic asymptotic normality theorem, since we have abstained from assuming the existence of a "limiting information matrix", that is, a uniform limit of $-\frac{1}{T}H_T(Y_1, \cdots, Y_T \mid \theta)$.

Note first that, under assumptions M1 to M4, there exists a sequence $\{\hat{\theta}_T\}_{T\in N_+}$ of continuous functions such that $\hat{\theta}_T : E^T \to \Theta_0 \subset \Theta^o$ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the T-observation ML estimator of θ. The fact that the ML estimator takes values in the interior of the parameter space simplifies the proof somewhat.

Choose any $1 \le i \le m$ and let $s_{i,T}(Y_1, \cdots, Y_T \mid \theta)$ be the ith component of the score function $s_T(Y_1, \cdots, Y_T \mid \theta)$. By assumption M5, the function $s_{i,T} : E^T \times \Theta^o \to \mathbb{R}$ is continuous on $E^T \times \Theta$ and continuously differentiable with respect to its second argument, so by the stochastic mean value theorem,

$$\begin{aligned} s_{i,T}(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) &= s_{i,T}(Y_1, \cdots, Y_T \mid \theta_0) \\ &\quad + \left(\boldsymbol{\nabla}_\theta\ s_{i,T}(Y_1, \cdots, Y_T \mid \theta_T^{(i)})\right)' \left(\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right), \end{aligned}$$

where $\theta_T^{(i)}$ is a random variable such that

$$\left|\theta_T^{(i)} - \theta_0\right| \le \left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right|.$$

We can write

$$\boldsymbol{\nabla}_\theta\ s_{i,T}(Y_1, \cdots, Y_T \mid \theta_T^{(i)}) = \begin{pmatrix} \frac{\partial^2 l_T(Y_1, \cdots, Y_T \mid \theta_T^{(i)})}{\partial\theta_1 \partial\theta_i} \\ \vdots \\ \frac{\partial^2 l_T(Y_1, \cdots, Y_T \mid \theta_T^{(i)})}{\partial\theta_m \partial\theta_i} \end{pmatrix} = H_{i,T}(Y_1, \cdots, Y_T \mid \theta_T^{(i)}),$$

where $H_{i,T}(Y_1, \cdots, Y_T \mid \theta_T^{(i)})$ is the ith column/row of the Hessian $H_T(Y_1, \cdots, Y_T \mid \theta_T^{(i)})$

(it doesn't matter whether it's the column or row because the Hessian is symmetric).

Define the $m \times m$ matrix H_T^* as

$$H_T^* = \Big(H_{1,T}(Y_1, \cdots, Y_T \mid \theta_T^{(1)})' \quad \cdots \quad H_{m,T}(Y_1, \cdots, Y_T \mid \theta_T^{(m)}) \Big),$$

we can now see that

$$s_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) = s_T(Y_1, \cdots, Y_T \mid \theta_0) + H_T^* \Big(\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big).$$

By the first order condition for maximization, the left hand side equals $O_{m \times 1}$, so we have

$$\Big(-\frac{1}{T} H_T^* \Big) \cdot \sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] = \frac{1}{\sqrt{T}} s_T(Y_1, \cdots, Y_T \mid \theta_0).$$

Step 1: Asymptotic Behavior of $-\frac{1}{T} H_T^*$

We first show that the distance between $-\frac{1}{T} H_T^*$ and the sample information matrix $\frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)$ reduces to 0 in probability.

Choose any $\epsilon > 0$. By assumption M8, there exists a $\delta > 0$ such that

$$\lim_{T \to \infty} \mathbb{P} \left(\sup_{\theta \in \bar{B}(\theta_0, \delta)} \left\| \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta) \right\| \leq \epsilon \right) = 1,$$

and because the ML estimator is consistent for θ_0,

$$\lim_{T \to \infty} \mathbb{P} \Big(\Big| \hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big| \leq \delta \Big) = 1.$$

It follows from the fact that $\Big| \theta_T^{(i)} - \theta_0 \Big| \leq \Big| \hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big|$ for each $1 \leq i \leq m$ that

$$\lim_{T \to \infty} \mathbb{P} \Big(\Big| \theta_T^{(i)} - \theta_0 \Big| \leq \delta \Big) = 1$$

for all $1 \leq i \leq m$. Define the event A_T as

$$A_T = \left[\bigcap_{i=1}^{m} \Big\{ \Big| \theta_T^{(i)} - \theta_0 \Big| \leq \delta \Big\} \right] \cap \Big\{ \sup_{\theta \in \bar{B}(\theta_0, \delta)} \left\| \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta) \right\| \leq \epsilon \Big\}.$$

We just showed above that $\mathbb{P}(A_T) \to 1$ as $T \to \infty$. Note that, if $\omega \in A_T$, then $\theta_T^{(i)}(\omega) \in \bar{B}(\theta_0, \delta)$ for $1 \leq i \leq m$ and therefore

$$A_T \subset \Big\{ \left\| \frac{1}{T} H_T^* + \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) \right\| \leq \epsilon \sqrt{m} \Big\},$$

which implies that

$$\lim_{T\to\infty} \mathbb{P}\left(\left\|\frac{1}{T}H_T^* + \frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right\| \leq \epsilon\sqrt{m}\right) = 1.$$

This holds for any $\epsilon > 0$, so

$$\frac{1}{T}H_T^* + \frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) \xrightarrow{p} O_{m\times m}$$

as $T \to \infty$.

Step 2: Rate of Convergence of ML Estimator

Now we want to show that the $o_p(1)$ term $\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0$ has a rate of convergence $\frac{1}{\sqrt{T}}$, that is, $\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0 = O_p(T^{-1/2})$.

Note that

$$\begin{aligned}\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}\right\|^2 &= \operatorname{tr}\left[\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}\prime}\right] \\ &= \operatorname{tr}\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right) \leq m\cdot q_{T,max}(Y_1,\cdots,Y_T),\end{aligned}$$

so that

$$\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}\right\| \leq \sqrt{m\cdot q_{T,max}(Y_1,\cdots,Y_T)}.$$

Assumption M7 now tells us that

$$\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}\right\| \leq \sqrt{m\cdot\rho_{max}}$$

with probability approaching 1 as $T \to \infty$. This indicates that $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}} = O_p(1)$ as $T \to \infty$.

For any $T \in N_+$, we can see that

$$\frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T \mid \theta_0) = \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}}\left[\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \times \frac{1}{\sqrt{T}}s_T(Y_1,\cdots,Y_T \mid \theta_0)\right].$$

By assumption M9, the term

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \times \frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0) = O_p(1)$$

as $T \to \infty$, and we just saw that $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}} = O_p(1)$. Therefore,

$$\frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0) = O_p(1) \quad \text{as } T \to \infty.$$

From the result shown above on the asymptotic behavior of the random matrix $-\frac{1}{T}H_T^*$, as well as assumption M7,

$$\left\| -\frac{1}{T}H_T^* - \frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) \right\| \leq \frac{\rho_{min}}{2\sqrt{m}} \quad \text{and} \quad 0 < \rho_{min} \leq q_{T,min}(Y_1,\cdots,Y_T)$$

with probability approaching 1 as $T \to \infty$, where $\{\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\}_{T\in N_+}$ is a sequence of $m \times m$ real symmetric matrices (positive semidefinite matrices, to be exact). Moreover,

$$\left(-\frac{1}{T}H_T^*\right) \cdot \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right] = \frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0),$$

where the term on the right hand side is $O_p(1)$ as $T \to \infty$. By the generalized inversion at the limit, it follows that

$$\sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right] = O_p(1) \quad \text{as } T \to \infty.$$

Step 3: Asymptotic Distribution of ML Estimator

Now that we have established the rate of convergence of the ML estimator, it remains to obtain its asymptotic distribution. Pre-multiplying both sides of

$$\left(-\frac{1}{T}H_T^*\right) \cdot \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right] = \frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0),$$

by $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}$ and expanding suitably yields the equation

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \times \frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0)$$

$$= \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \left(-\frac{1}{T}H_T^*\right) \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$$

$$= \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$$
$$- \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T^*\right) \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right].$$

Rearranging, we can see that

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$$
$$= \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \times \frac{1}{\sqrt{T}} s_T(Y_1,\cdots,Y_T \mid \theta_0) + \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T^*\right) \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right].$$

We want to show that the second term is negligible as $T \to \infty$.

Because

$$\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}\right\|^2 = \operatorname{tr}\left(\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}\prime}\right)$$
$$= \operatorname{tr}\left[\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-1}\right] \leq \frac{m}{q_{T,min}(Y_1,\cdots,Y_T)},$$

we have

$$\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}\right\| \leq \sqrt{\frac{m}{q_{T,min}(Y_1,\cdots,Y_T)}}$$

for any $T \in N_+$. In light of assumption M7,

$$\left\|\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}\right\| \leq \sqrt{\frac{m}{\rho_{min}}}$$

with probability approaching 1 as $T \to \infty$. This implies that

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} = O_p(1) \quad \text{as } T \to \infty.$$

Therefore, the term

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}} \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T^*\right) \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$$

is the product of two $O_p(1)$ terms $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{-\frac{1}{2}}$ and $\sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$, and one $o_p(1)$ term $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T^*$. It follows that the entire term is $o_p(1)$ as $T \to \infty$, so that

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\right]$$

$$= \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{-\frac{1}{2}} \times \frac{1}{\sqrt{T}} s_T(Y_1, \cdots, Y_T \mid \theta_0) + o_p(1).$$

By Slutsky's theorem and assumption A9,

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}'} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right],$$

as desired.

Q.E.D.

As we noted in the previous section, this result suggests that, in large samples,

$$\hat{\theta}_T(Y_1, \cdots, Y_T) \approx \mathcal{N}\left[\theta_0,\ \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)^{-1}\right],$$

so that the ML estimator is asymptotically unbiased (=consistent) and attains the Cramer-Rao Bound. in the limit This result is what econometricians refer to when they say that the ML estimator is asymptotically efficient.

3.5.5 The Trinity: Wald, LR and LM

The asymptotic properties of the ML estimator derived above has many useful statistical applications, in particular when constructing asymptotic test statistics. For the sake of simplicity, we consider here tests with simple null hypotheses, that is, we consider testing

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta \neq \theta_0$$

for some $\theta_0 \in \Theta$.

There are three types of tests that are often brought up in the context of ML estimation: the Wald, LR and LM tests, collectively referred to as the trinity tests. We consider each test in turn.

The Wald Test

The Wald test is perhaps the simplest and most intuitive of the trinity. It is motivated as follows. Suppose that the null were true, that is, θ_0 were the true parameter. Then, in light of the consistency of the ML estimator, the difference $\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0$ would be close to 0, so that quadratic forms of the form

$$\left(\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right)' W_T \left(\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right)$$

would be close to 0 for an appropriately chosen weighting matrix W_T. On the other hand, if any alternative $H_1 : \theta = \theta_1 \neq \theta_0$ were true, then the difference $\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0$ would converge to the non-zero value $\theta_1 - \theta_0$, meaning that the quadratic form above will diverge to $+\infty$ in probability if the weighting matrix W_T also diverges. This shows us that using a test statistic based on the quadratic form above will yield a test whose significance level we can control and which is consistent.

The formal definition of the Wald test statistic, as well as its asymptotic properties, are given below:

Theorem 3.16 (Asymptotic Properties of the Wald Test Statistic)
Suppose that assumptions M1 to M9 hold, and let $\{\hat{\theta}_T\}_{T \in N_+}$ be a sequence of continuous functions such that $\hat{\theta}_T : E^T \to \Theta_0$ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the T-observation ML estimator of θ.

The Wald statistic for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is defined as

$$W_T(Y_1, \cdots, Y_T) = T \cdot \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]' \left(\frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right) \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right].$$

If $\theta_0 \in \Theta_0$, then under the null, this test statistic has asymptotic distribution

$$W_T(Y_1, \cdots, Y_T) \xrightarrow{d} \chi^2_m,$$

so that the null hypothesis is rejected for large values of the Wald statistic.

Under the alternative $H_1 : \theta = \theta_1 \neq \theta_0$, where $\theta_1 \in \Theta_0$, the test statistic $W_T(Y_1, \cdots, Y_T)$ diverges in probability to $+\infty$.

Proof) Suppose that the null is true and that $\theta_0 \in \Theta_0$. By assumptions M1 to M9, the ML estimator exists, is consistent for θ_0, the difference $\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0$ is $O_p(T^{-1/2})$, and its asymptotic distribution is given as

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right].$$

Therefore,

$$W_T(Y_1, \cdots, Y_T) = T \cdot \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]' \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right) \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]$$

$$= \left[\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]\right]' \left[\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]\right]$$

$$\xrightarrow{d} \chi^2_m.$$

Now suppose that the alternative $H_1 : \theta = \theta_1 \neq \theta_0$ is true, where $\theta_1 \in \Theta_0$. Then, by assumptions M1 to M9 again and the consistency of the ML estimator,

$$\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_1 \xrightarrow{p} O_{m\times 1},$$

so that

$$\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \xrightarrow{p} \theta_1 - \theta_0 \neq O_{m\times 1}.$$

For any positive definite $A \in \mathbb{R}^{m\times m}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m > 0$ and non-zero $v \in \mathbb{R}^m$, letting $A = PDP'$ be the orthogonal eigendecomposition of A and $z = \frac{v}{|v|}$, since

$$\left|P'z\right|^2 = z'PP'z = z'z = 1,$$

it follows that

$$v'Av = |v|^2 (z'Az) = |v|^2 \left[(z'P)D(P'z)\right]$$

$$= |v|^2 \cdot \sum_{i=1}^{m} \lambda_i \cdot (P'z)_i^2 \geq |v|^2 \cdot \lambda_1 \left(\sum_{i=1}^{m} (P'z)_i^2\right)$$

$$= |v|^2 \lambda_m \cdot \left|P'z\right|^2 = |v|^2 \lambda_m.$$

If $v = O_{m\times 1}$, then the above inequality holds trivially as an equality.

For any $T \in N_+$, the random matrix $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)$ is positive definite valued with minimum eigenvalue $q_{T,min}(Y_1, \cdots, Y_T)$,the above result shows us that

$$W_T(Y_1, \cdots, Y_T) = T \cdot \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]' \left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right) \left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right]$$

$$\geq T \cdot \left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right|^2 \cdot q_{T,min}(Y_1, \cdots, Y_T).$$

By assumption M7, $q_{T,min}(Y_1, \cdots, Y_T) \geq \rho_{min} > 0$ with probability approaching 1 as $T \to \infty$, and because

$$\lim_{T\to\infty} \mathbb{P}\left(\left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 - (\theta_1 - \theta_0)\right| \leq \frac{1}{2}|\theta_1 - \theta_0|\right) = 1$$

by the consistency of the ML estimator, the triange inequality implies that

$$\lim_{T\to\infty} \mathbb{P}\left(\left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right| \geq \frac{1}{2}|\theta_1 - \theta_0|\right) = 1$$

as well. Putting the two results together, we can now see that

$$\lim_{T\to\infty} \mathbb{P}\left(\left|\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right|^2 \cdot q_{T,min}(Y_1, \cdots, Y_T) \geq \frac{1}{4}|\theta_1 - \theta_0|^2 \cdot \rho_{min}\right) = 1,$$

where $\frac{1}{4}|\theta_1 - \theta_0|^2 \cdot \rho_{min} > 0$. It follows from lemma 2.4 that $W_T(Y_1, \cdots, Y_T)$ diverges in probability to $+\infty$ under θ_1.

Q.E.D.

The divergence in probability result shown above implies that the Wald test is consistent against alternatives $H_1 : \theta = \theta_1 \neq \theta_0$ such that $\theta_1 \in \Theta_0$.

The Likelihood Ratio Test

The Likelihood Ratio (LR) test is perhaps the most fundamental test in statistical theory. It is based on a simple intuition. Gibbs' inequality tells us that, for any $T \in N_+$, the mapping

$$\theta \mapsto \mathbb{E}\left[l_T(Y_1, \cdots, Y_T \mid \theta)\right]$$

is uniquely maximized at the true parameter value θ_0. Since the ML estimator $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is close to the true parameter θ_0 under large enough samples, and $l_T(Y_1, \cdots, Y_T \mid \theta_0)$ is close to its mean $\mathbb{E}\left[l_T(Y_1, \cdots, Y_T \mid \theta_0)\right]$, if θ_0 is indeed the true parameter, then the difference

$$l_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) - l_T(Y_1, \cdots, Y_T \mid \theta_0)$$

should be close to 0. On the other hand, if θ_0 is far from the true parameter, then $l_T(Y_1, \cdots, Y_T \mid \theta)$ should be small compared to $l_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T))$, so that the above difference should also be large. This allows us to think of a test for the null $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ by testing whether the above likelihood difference is larger than some appropriately chosen critical value; the null is rejected if the likelihood ratio exceeds this critical value. The test is called a likelihood "ratio" test because the difference in likelihoods is just the log of the ratio of densities.

We have already studied versions of the likelihood ratio test, although we did not explicitly refer to them as such. For instance, the UMP test for simple null hypotheses against simple alternatives proposed in the Neyman-Pearson lemma can be viewed as a likelihood ratio test; recall how the NP test rejects the null if and only if the ratio of the density under the true parameter to the density under the alternative is larger than some critical value. Similarly, the Karlin-Rubin one-sided UMP test is also a likelihood ratio-type test, since the MLR property ensures that the test statistic is proportional to the ratio of densities.

The formal definition of the likelihood test statistic, as well as its asymptotic properties, are given below. Once again, the test statistic has an asymptotic chi-squared distribution with m degrees of freedom under the null, while it diverges to $+\infty$ under parameter values which belong to the alternative hypothesis, demonstrating that the test is consistent.

Theorem 3.22 (Asymptotic Properties of the LR Test Statistic)
Suppose that assumptions M1 to M9 hold, and let $\{\hat{\theta}_T\}_{T\in N_+}$ be a sequence of continuous functions such that $\hat{\theta}_T : E^T \to \Theta_0$ and $\hat{\theta}_T(Y_1,\cdots,Y_T)$ is the T-observation ML estimator of θ.

The LR statistic for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is defined as

$$LR_T(Y_1,\cdots,Y_T) = 2\left[l_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) - l_T(Y_1,\cdots,Y_T \mid \theta)\right].$$

If $\theta_0 \in \Theta_0$, then under the null, this test statistic has asymptotic distribution

$$LR_T(Y_1,\cdots,Y_T) \xrightarrow{d} \chi^2_m,$$

so that the null hypothesis is rejected for large values of the LR statistic.

Under the alternative $H_1 : \theta = \theta_1 \neq \theta_0$, where $\theta_1 \in \Theta_0$, the test statistic $LR_T(Y_1,\cdots,Y_T)$ diverges in probability to $+\infty$.

Proof) **Asymptotic Distribution of LR Statistic under the Null**

Suppose that the null is true and that $\theta_0 \in \Theta_0$. For any $T \in N_+$, define the function $h : E^T \times [0,1] \to \mathbb{R}$ as

$$h(y,t) = -\left|l_T(y \mid \theta_0) - l_T(y \mid \hat{\theta}_T(y)) - \frac{1}{2}\left[\hat{\theta}_T(y) - \theta_0\right]' H_T(y \mid t\cdot\hat{\theta}(y) + (1-t)\theta_0)\left[\hat{\theta}_T(y) - \theta_0\right]\right|$$

for any $(y,t) \in E^T \times [0,1]$. Then, since h is continuous on $E^T \times [0,1]$ by assumption and $[0,1]$ is a compact set, by Berge's maximum theorem the set of solutions $G_T : E^T \to [0,1]$ to the maximization problem

$$\max_{t\in[0,1]} \quad h(y,t)$$

is non-empty compact valued and upper hemicontinuous. Furthermore, because $[0,1]$ is a complete and separable metric space under the euclidean metric, and E^T is a topological space by assumption, the measurable selection theorem tells us that there exists a measurable selection $g_T : E^T \to [0,1]$ of G_T, that is, a measurable function such that $g_T(y) \in G_T(y)$ for any $y \in [0,1]$.

Now note that, for any $y \in E^T$, by a second-order Taylor expansion of $l_T(y \mid \theta_0)$ around $\hat{\theta}_T(y)$, there exists some $t^* \in [0,1]$ such that

$$\begin{aligned} l_T(y \mid \theta_0) &= l_T(y \mid \hat{\theta}_T(y)) + s_T(y \mid \hat{\theta}_T(y))'\left[\theta_0 - \hat{\theta}_T(y)\right] \\ &\quad + \frac{1}{2}\left[\hat{\theta}_T(y) - \theta_0\right]' H_T(y \mid t^*\cdot\hat{\theta}(y) + (1-t^*)\theta_0)\left[\hat{\theta}_T(y) - \theta_0\right] \\ &= l_T(y \mid \hat{\theta}_T(y)) + \frac{1}{2}\left[\hat{\theta}_T(y) - \theta_0\right]' H_T(y \mid t^*\cdot\hat{\theta}(y) + (1-t^*)\theta_0)\left[\hat{\theta}_T(y) - \theta_0\right], \end{aligned}$$

where the last equality follows because $s_T(y \mid \hat{\theta}_T(y)) = O_{m\times 1}$ by the first order conditions for maximization. As such, $h(y,t^*) = 0$, which means that

$$\sup_{t\in[0,1]} h(y,t) = 0 = h(y, g_T(y))$$

by definition. This holds for any $y \in E^T$, so defining the function $\theta_T^* : E^T \to \Theta_0$ as

$$\theta_T^*(y) = g_T(y)\hat{\theta}_T(y) + (1 - g_T(y))\theta_0$$

for any $y \in E^T$, θ_T^* is a measurable function by the measurability of g and the continuity of $\hat{\theta}_T$, which satisfies

$$2\Big[l_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) - l_T(Y_1,\cdots,Y_T \mid \theta_0)\Big]$$

$$= \Big(\sqrt{T}\Big[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\Big]\Big)' \Big[-\frac{1}{T}H_T(Y_1,\cdots,Y_T \mid \theta_T^*(Y_1,\cdots,Y_T))\Big] \Big(\sqrt{T}\Big[\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\Big]\Big)$$

everywhere on the sample space.

By construction, g_T takes values in $[0,1]$, so that

$$|\theta_T^*(Y_1,\cdots,Y_T) - \theta_0| = (1 - g_T(Y_1,\cdots,Y_T)) \cdot \Big|\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\Big| \leq \Big|\hat{\theta}_T(Y_1,\cdots,Y_T) - \theta_0\Big|$$

for any $T \in N_+$. It follows from the consistency of the ML estimator, which assumptions M1 to M4 ensure, that $\theta_T^* \xrightarrow{p} \theta_0$.

Choose any $\epsilon > 0$. By assumption M8, there exists a $\delta > 0$ such that

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{\theta\in\bar{B}(\theta_0,\delta)} \left\|\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T(Y_1,\cdots,Y_T \mid \theta)\right\| \leq \epsilon\right) = 1.$$

By the definition of convergence in probability,

$$\lim_{T\to\infty} \mathbb{P}(|\theta_T^*(Y_1,\cdots,Y_T) - \theta_0| \leq \delta) = 1.$$

Define the event $A_T \in \mathcal{H}$ as

$$A_T = \{|\theta_T^*(Y_1,\cdots,Y_T) - \theta_0| \leq \delta\} \cap \Big\{\sup_{\theta\in\bar{B}(\theta_0,\delta)} \left\|\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T(Y_1,\cdots,Y_T \mid \theta)\right\| \leq \epsilon\Big\}.$$

From what we have just observed, it follows that $\mathbb{P}(A_T) \to 1$ as $T \to \infty$. Furthermore, for any outcome contained in the event A_T, since $\theta_T^*(Y_1,\cdots,Y_T)$ is contained in the closed ball of radius δ around θ_0,

$$\left\|\frac{1}{T}\hat{\mathbb{I}}_T(Y_1,\cdots,Y_T) + \frac{1}{T}H_T(Y_1,\cdots,Y_T \mid \theta_T^*(Y_1,\cdots,Y_T))\right\| \leq \frac{\epsilon}{2}.$$

It follows that

$$A_T \subset \Big\{ \Big\| \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta_T^*(Y_1, \cdots, Y_T)) \Big\| \leq \epsilon \Big\},$$

and therefore

$$\lim_{T \to \infty} \mathbb{P} \left(\Big\| \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta_T^*(Y_1, \cdots, Y_T)) \Big\| \leq \epsilon \right) = 1.$$

This holds for any $\epsilon > 0$, so by definition

$$\mathcal{M}_T := \frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta_T^*(Y_1, \cdots, Y_T)) \xrightarrow{p} O_{m \times m}.$$

Putting all that we have derived together, we have

$$\begin{aligned}
&2 \Big[l_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) - l_T(Y_1, \cdots, Y_T \mid \theta_0) \Big] \\
&= \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big)' \Big[-\frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta_T^*(Y_1, \cdots, Y_T)) \Big] \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big) \\
&= \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big)' \Big[\frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) \Big] \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big) \\
&\qquad - \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big)' \mathcal{M}_T \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big).
\end{aligned}$$

Under assumptions M1 to M9, we showed in the previous section that $\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] = O_p(1)$, and we have just shown that $\mathcal{M}_T = o_p(1)$. It follows that

$$\Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big)' \mathcal{M}_T \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big) = o_p(1) \quad \text{as } T \to \infty,$$

which allows us to express the LR test statistic as

$$\begin{aligned}
&2 \Big[l_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) - l_T(Y_1, \cdots, Y_T \mid \theta_0) \Big] \\
&= \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big)' \Big[\frac{1}{T} \hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) \Big] \Big(\sqrt{T} \Big[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0 \Big] \Big) + o_p(1) \\
&= W_T(Y_1, \cdots, Y_T) + o_p(1),
\end{aligned}$$

where $W_T(Y_1, \cdots, Y_T)$ is just the Wald test statistic. By Slutsky's theorem, we can conclude that

$$2 \Big[l_T(Y_1, \cdots, Y_T \mid \hat{\theta}_T(Y_1, \cdots, Y_T)) - l_T(Y_1, \cdots, Y_T \mid \theta_0) \Big] \xrightarrow{d} \chi^2_m.$$

Asymptotic Behavior of LR Statistic under the Alternative

Now suppose that $\theta_1 \neq \theta_0$ is the true parameter, where $\theta_1 \in \Theta_0$. From what we showed above,

$$2\left[l_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) - l_T(Y_1,\cdots,Y_T \mid \theta_1)\right] = O_p(1).$$

Define $\epsilon = |\theta_1 - \theta_0| > 0$. Then,

$$\theta_0 \in \Theta_0 \cap B(\theta_1,\epsilon)^c,$$

so that, by the asymptotic identification condition, there exists a $K > 0$ such that

$$\lim_{T\to\infty} \mathbb{P}\left(\frac{1}{\sqrt{T}}\left[l_T(Y_1,\cdots,Y_T \mid \theta_0) - l_T(Y_1,\cdots,Y_T \mid \theta_1)\right] \leq -K\right) = 1.$$

It follows that, for any $M > 0$, letting $N \in N_+$ be chosen so that $-K\sqrt{T} < -M$, we have

$$\lim_{T\to\infty} \mathbb{P}\left(l_T(Y_1,\cdots,Y_T \mid \theta_0) - l_T(Y_1,\cdots,Y_T \mid \theta_1) < -M\right) = 1,$$

so that, by definition,

$$l_T(Y_1,\cdots,Y_T \mid \theta_1) - l_T(Y_1,\cdots,Y_T \mid \theta_0)$$

diverges in probability to $+\infty$.

The LR test statistic can now be written as

$$LR_T(Y_1,\cdots,Y_T) = 2\left[l_T(Y_1,\cdots,Y_T \mid \hat{\theta}_T(Y_1,\cdots,Y_T)) - l_T(Y_1,\cdots,Y_T \mid \theta_1)\right]$$

$$+ 2\left[l_T(Y_1,\cdots,Y_T \mid \theta_1) - l_T(Y_1,\cdots,Y_T \mid \theta_0)\right].$$

Since this is the sum of an $O_p(1)$ term and a term that diverges in probability to $+\infty$, by lemma 2.4 we can conclude that $LR_T(Y_1,\cdots,Y_T)$ diverges in probability to $+\infty$.

Q.E.D.

In the process of the above proof, we showed that it is possible to express

$$LR_T(Y_1, \cdots, Y_T) = W_T(Y_1, \cdots, Y_T) - r_T(Y_1, \cdots, Y_T),$$

where $r_T(Y_1, \cdots, Y_T)$ is a remainder term that is a quadratic form of a positive definite matrix. Therefore, $r_T(Y_1, \cdots, Y_T) \geq 0$, so that

$$LR_T(Y_1, \cdots, Y_T) \leq W_T(Y_1, \cdots, Y_T).$$

In other words, the LR test statistic is generally smaller than the Wald test statistic.

The Lagrange Multiplier/Score Test

The final test in the trinity is the Lagrange Multiplier (LM) test, also referred to as the score test. The score test is motivated in a similar manner to the Wald test. When testing the null hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can consider the unrestricted and restricted estimators of θ. The unrestricted estimator of θ the estimator obtained by maximizing the likelihood without imposing any restrictions; as such, it is simply the ML estimator of θ. Meanwhile, the restricted estimator of θ is the estimator obtained by once again maximizing the likelihood, only this time we maximize it subject to the restriction that the null is true. In the case of the simple null hypothesis above, the restricted estimator of θ is simply θ_0, while the problem is considerably more complicated when we are faced with a composite null hypothesis.

The Wald test statistic is a test that uses only the unrestricted estimator of θ, since it is derived solely from the result

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}\prime} \times \sqrt{T}\left[\hat{\theta}_T(Y_1, \cdots, Y_T) - \theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right],$$

or in other words, from the asymptotic distribution of the ML estimator. Meanwhile, the LR test statistic utilizes asymptotic results pertaining to both the unrestricted estimator $\hat{\theta}_T(Y_1, \cdots, Y_T)$ and the restricted estimator θ_0, since it is given as the difference in the log-likelihood under the unrestricted and restricted estimators of θ, respectively.

In this context, the score test statistic uses only the asymptotic result pertaining to the restricted estimator, namely assmption M9:

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{-\frac{1}{2}} \left[\frac{1}{\sqrt{T}} s_T(Y_1, \cdots, Y_T \mid \theta_0)\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right].$$

Note the conspicuous absence here of the ML estimator $\hat{\theta}_T(Y_1, \cdots, Y_T)$. This also motivates the name of the score test; the score test statistic is derived based on the asymptotic behavior of the score at the restricted estimator.

The score test has another, more widely used name: the LM test. To see why, consider the

constrained maximization problem

$$\begin{aligned} \min_{\theta \in \Theta} \quad & l_T(Y_1, \cdots, Y_T \mid \theta) \\ \text{subject to} \quad & \theta = \theta_0. \end{aligned}$$

The solution to this problem is clearly $\theta = \theta_0 \in \Theta_0 \subset \Theta^o$, and the Lagrangian is given by

$$\mathcal{L}(\theta, q_T) = l_T(Y_1, \cdots, Y_T \mid \theta) + q_T'(\theta_0 - \theta).$$

The first order conditions for constrained maximization now tell us that

$$\frac{\partial l_T(Y_1, \cdots, Y_T \mid \theta_0)}{\partial \theta_0} = q_T,$$

so that the Lagrange multiplier q_T is exactly equal to the score function evaluated at θ_0. Since the score test is based on the asymptotic distribution of the score function, which is exactly the Lagrange multiplier q_T, the score test is also called the LM test.

The formal definition of the score test statistic, as well as its asymptotic properties, are given below. We can see again that this test statistic has exactly the same asymptotic distribution as the Wald and LR test statistics under the null.

Theorem 3.23 (Asymptotic Properties of the LM/Score Test Statistic) Suppose that assumptions M1 to M9 hold, and let $\{\hat{\theta}_T\}_{T \in N_+}$ be a sequence of continuous functions such that $\hat{\theta}_T : E^T \to \Theta_0$ and $\hat{\theta}_T(Y_1, \cdots, Y_T)$ is the T-observation ML estimator of θ.

The LM/Score test statistic for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is defined as

$$LM_T(Y_1, \cdots, Y_T) = s_T(Y_1, \cdots, Y_T \mid \theta_0)' \left[\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right]^{-1} s_T(Y_1, \cdots, Y_T \mid \theta_0).$$

If $\theta_0 \in \Theta_0$, then under the null, this test statistic has asymptotic distribution

$$LM_T(Y_1, \cdots, Y_T) \xrightarrow{d} \chi^2_m,$$

so that the null hypothesis is rejected for large values of the LM statistic.

Proof) Suppose that the null is true and that $\theta_0 \in \Theta_0$. By assumption M9,

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{-\frac{1}{2}} \left[\frac{1}{\sqrt{T}} s_T(Y_1, \cdots, Y_T \mid \theta_0)\right] \xrightarrow{d} \mathcal{N}\left[O_{m \times 1},\ I_m\right],$$

so that

$$\frac{1}{T} s_T(Y_1, \cdots, Y_T \mid \theta_0)' \left[\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right]^{-1} s_T(Y_1, \cdots, Y_T \mid \theta_0) \xrightarrow{d} \chi^2_m$$

by the continuous mapping theorem. The left hand side is simply the LM test statistic,

so we have the desired result

$$LM_T(Y_1, \cdots, Y_T) \xrightarrow{d} \chi^2_m.$$

Q.E.D.

Note that we are unable to establish the consistency of the LM test under the current assumptions, since we did not impose sufficient regularity conditions on the score function to be able to discuss its asymptotic properties at points other than the true parameter. Heuristically, this is a consequence of the fact that the LM test statistic does not involve the unrestricted estimator; therefore, it contains only information on the null and not the alternative. This feature sometimes causes the LM test to display erratic behavior, such as the non-monotonic power property found in many LM-type stationarity tests (most notably, the KPSS test).

3.5.6 A Special Case: I.I.D. Data

So far, we have studied the properties of the ML estimator and associated tests under the following regularity conditions:

M1 : **The Data Generating Process**

Suppose $\{Y_t\}_{t\in\mathbb{Z}}$ is a sample taking values in the separable metric space (E,d) with corresponding Borel σ-algebra $\mathcal{E}$. The parameter space Θ is a subset of the euclidean space $\mathbb{R}^m$, with corresponding Borel σ-algebra ϑ.

Let $\{\mu_T\}_{T\in N_+}$ be a sequence of statistical models corresponding to the sample $\{Y_t\}_{t\in\mathbb{Z}}$. Specifically, for any $T\in N_+$, μ_T is a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$ such that $\mu_T(\theta,\cdot)$ is the distribution of $(Y_1,\cdots,Y_T)$ when the parameter is $\theta\in\Theta$.

$\{\lambda_T\}_{T\in\mathbb{N}}$ and $\{f_T\}_{T\in N_+}$ are sequences of measures and functions such that, for any $T\in N_+$, λ_T is a σ-finite measure on $(E,\mathcal{E})^T$ and $f_T : E^T\times\Theta\to\mathbb{R}_{++}$ is a $\mathcal{E}^T\bigotimes\vartheta$-measurable function that satisfies

$$\int_A f_T(y\mid\theta)d\lambda_T(y)=\mu_T(\theta,A)$$

for any $(\theta,A)\in\Theta\times\mathcal{E}^T$.

The log-likelihood of the T-observation sample $(Y_1,\cdots,Y_T)$ is defined as

$$l_T(y\mid\theta)=\log f_T(y\mid\theta)$$

for any $(y,\theta)\in E^T\times\Theta$.

M2 : **Parameter Space**

Θ is a convex subset of $\mathbb{R}^m$. There exists a non-empty, convex and compact subset Θ_0 of Θ that is contained in the interior Θ^o of the parameter space and

$$\sup_{\theta\in\Theta_0^c} l_T(y\mid\theta)\le\sup_{\theta\in\Theta_0} l_T(y\mid\theta)$$

for any $y\in E^T$. In addition, the true parameter θ_0 is contained in Θ_0.

M3 : **Regularity Conditions for Log-Likelihood**

We assume that, for any $T\in N_+$, the function l_T is continuous on $E^T\times\Theta$, and that $l_T(y\mid\cdot)$ is a strictly concave function on Θ_0 for any $y\in E^T$.

M4 : **Asymptotic Identification Condition**

For any $\epsilon > 0$, such that $B(\theta_0, \epsilon)^c \cap \Theta_0 \neq \emptyset$, there exists a number $K(\epsilon) > 0$ such that

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{\theta\in\Theta\setminus B(\theta_0,\epsilon)} \frac{1}{\sqrt{T}} \Big[l_T(Y_1, \cdots, Y_T \mid \theta) - l_T(Y_1, \cdots, Y_T \mid \theta_0) \Big] \leq -K(\epsilon) \right) = 1,$$

where we put the supremum over an empty set equal to $-\infty$, as is customary.

M5 : **Smoothness Properties of the Log-Likelihood**

For any $T \in N_+$ and $y \in E^T$, the function $f_T(y \mid \cdot)$ is twice continuously differentiable on Θ^o. Define the score as

$$s_T(y \mid \theta) := \frac{\partial l_T(y \mid \theta)}{\partial \theta}$$

and the Hessian as

$$H_T(y \mid \theta) := \frac{\partial^2 l_T(y \mid \theta)}{\partial \theta \partial \theta'}$$

for any $\theta \in \Theta^o$. We assume that both $s_T(\cdot \mid \theta) s_T(\cdot \mid \theta)'$ and $H_T(\cdot \mid \theta)$ are $\mu_T(\theta, \cdot)$-integrable for any $\theta \in \Theta^o$, and that they are both continuous on $E^T \times \Theta^o$.

M6 : **Conditions for Interchanging Integration and Differentiation**

We assume that, for any $T \in N_+$, the functions $f_T(\cdot \mid \theta)$ and $\frac{\partial f_T(\cdot|\theta)}{\partial \theta}$ are λ_T-integrable for any $\theta \in \Theta^o$, and that there exists some measurable non-negative and λ_T-integrable function $b_T : E^T \to \mathbb{R}_+$ such that

$$\left| \frac{\partial f_T(y \mid \theta)}{\partial \theta} \right|, \left\| \frac{\partial^2 f_T(y \mid \theta)}{\partial \theta \partial \theta'} \right\| \leq b_T(y)$$

for any $(y, \theta) \in E^T \times \Theta^o$.

M7 : **Boundedness of Scaled Sample Information Matrix**

Let $\hat{\mathbb{I}}_T : E^T \to \mathbb{R}^{m\times m}$ be defined as

$$\hat{\mathbb{I}}_T(y) = -H_T(y \mid \hat{\theta}_T(y))$$

for any $y \in E^T$, and let $q_{T,min}(y)$ and $q_{T,max}(y)$ be the minimum and maximum eigenvalues of the positive definite matrix $\frac{1}{T}\hat{\mathbb{I}}_T(y)$ for any $y \in E^T$.

We assume that there exist some $0 < \rho_{\min} < \rho_{max} < +\infty$ such that

$$\lim_{T\to\infty} \mathbb{P}\left(\rho_{\min} \leq q_{T,min}(Y_1, \cdots, Y_T) \leq q_{T,max}(Y_1, \cdots, Y_T) \leq \rho_{max}\right) = 1.$$

M8 : **Smoothness of Hessian**

For any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that $\bar{B}(\theta_0, \delta(\epsilon)) \subset \Theta$ (since θ_0 is in the interior of Θ) and

$$\lim_{T\to\infty} \mathbb{P}\left(\sup_{\theta \in \bar{B}(\theta_0,\delta(\epsilon))} \left\| \frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T) + \frac{1}{T} H_T(Y_1, \cdots, Y_T \mid \theta) \right\| \leq \epsilon \right) = 1,$$

where $\bar{B}(\theta_0, \delta(\epsilon))$ is the closed $\delta(\epsilon)$-ball around θ_0.

M9 : **Asymptotic Normality of Sclaed Score**

The score function, appropriately scaled, is asymptotically normal:

$$\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{-\frac{1}{2}} \left[\frac{1}{\sqrt{T}} s_T(Y_1, \cdots, Y_T \mid \theta_0)\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\, I_m\right],$$

where $\left(\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)\right)^{\frac{1}{2}}$ is the Cholesky factor of the positive definite-valued random matrix $\frac{1}{T}\hat{\mathbb{I}}_T(Y_1, \cdots, Y_T)$.

These are a lot of assumptions. The question now is: can these assumptions be satisfied by any reasonable statistical model? If some of these assumptions end up contradicting one another, or if they are too stringent that statistical models can seldom satisfy them, then the theory we have developed so far will be rendered obsolete.

In this section we verify that assumptions M1 to M9 are satisfied by an important class of statistical models that satisfy weaker primitive conditions. Specifically, we consider the case where the data are i.i.d., and show that the required assumptions are satisfied when we assume the regularity conditions of Newey and McFadden (1994) hold. The assumptions of the i.i.d. model are as follows:

IID.M1 : **The Data Generating Process**

The sample $\{X_i\}_{i\in N_+}$ is an i.i.d. process taking values in the separable metric space (E,d), where $\mathcal{E}$ is the corresponding Borel σ-algebra. The parameter space Θ is a subset of the euclidean space $\mathbb{R}^m$, with corresponding Borel σ-algebra ϑ.

The statistical model is summarized by μ, which is a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$. For any $\theta\in\Theta$, the probability measure $\mu(\theta,\cdot)$ is the distribution of each X_i given that the parameter is equal to θ, so that the distribution of the n-observation sample $(X_1,\cdots,X_n)$ given that the true parameter is θ is given as the product measure

$$\underbrace{\mu(\theta,\cdot)\times\cdots\times\mu(\theta,\cdot)}_{n}.$$

λ is a σ-finite measure on $(E,\mathcal{E})$, and $f_X : E\times\Theta\to\mathbb{R}_{++}$ a $\mathcal{E}\bigotimes\vartheta$-measurable function such that

$$\mu(\theta,A)=\int_A f_X(x\mid\theta)d\lambda(x)$$

for any $(\theta,A)\in\Theta\times\mathcal{E}$. This tells us that $f_X(\cdot\mid\theta)$ is the density of each X_i with respect to the measure λ, given that the true parameter is θ. For any $n\in N_+$, we define $f_n : E^n\times\Theta\to\mathbb{R}_{++}$ as

$$f_n(x_1,\cdots,x_n\mid\theta)=\prod_{i=1}^{n}f_X(x_i\mid\theta)$$

for any $(x_1,\cdots,x_n)\in E^n$ and $\theta\in\Theta$ by the independence of the sample. $f_n(\cdot\mid\theta)$ is the density of $(X_1,\cdots,X_n)$ with respect to the product measure $\lambda^n=\underbrace{\lambda\times\cdots\times\lambda}_{n}$ given that θ is the true parameter.

The log-likelihood is defined as

$$l(x\mid\theta)=\log f_X(x\mid\theta)$$

for any $(x,\theta)\in E\times\Theta$. It follows that the n-observation log-likelihood is

$$l_n(x_1,\cdots,x_n\mid\theta)=\log f_n(x_1,\cdots,x_n\mid\theta)=\sum_{i=1}^{n}\log f_X(x_i\mid\theta)=\sum_{i=1}^{n}l(x_i\mid\theta)$$

for any $(x_1,\cdots,x_n)\in E^n$ and $\theta\in\Theta$.

IID.M2 : **Parameter Space**

Θ is a convex subset of $\mathbb{R}^m$. There exists a non-empty, convex and compact subset Θ_0 of Θ that is contained in the interior Θ^o of the parameter space and

$$\sup_{\theta \in \Theta_0^c} l_n(x \mid \theta) \leq \sup_{\theta \in \Theta_0} l_n(x \mid \theta)$$

for any $n \in N_+$ and $x \in E^n$. In addition, the true parameter θ_0 is contained in Θ_0.

IID.M3 : **Regularity Conditions for Log-Likelihood**

The function l is continuous on $E \times \Theta$, and $l(x \mid \cdot)$ is a strictly concave function on Θ_0 for any $x \in E$. We also assume that $l(\mid \theta)$ is $\mu(\theta_0, \cdot)$-integrable for any $\theta \in \Theta_0$.

IID.M4 : **Intuitive Identification Condition**

For any $\theta_1, \theta_2 \in \Theta$,

$$l(x \mid \theta_1) = l(x \mid \theta_2)$$

for λ-a.e. $x \in E$ implies $\theta_1 = \theta_2$.

IID.M5 : **Smoothness Properties of the Log-Likelihood**

For any $n \in N_+$ and $x \in E$, the function $f_X(x \mid \cdot)$ is twice continuously differentiable on Θ^o. Define the score as

$$s(x \mid \theta) := \frac{\partial l(x \mid \theta)}{\partial \theta}$$

and the Hessian as

$$H(x \mid \theta) := \frac{\partial^2 l(x \mid \theta)}{\partial \theta \partial \theta'}$$

for any $\theta \in \Theta^o$. We assume that both $s(\cdot \mid \theta)s(\cdot \mid \theta)'$ and $H(\cdot \mid \theta)$ are $\mu(\theta, \cdot)$-integrable for any $\theta \in \Theta^o$, and that they are both continuous on $E \times \Theta$.

IID.M6 : **Conditions for Interchanging Integration and Differentiation**

For any $n \in N_+$, the functions $f_X(\cdot \mid \theta)$ and $\frac{\partial f_X(\cdot|\theta)}{\partial \theta}$ are λ-integrable for any $\theta \in \Theta^o$, and that there exist some measurable non-negative and λ-integrable function $b : E \to \mathbb{R}_+$ such that

$$|f_X(x \mid \theta)|, \left|\frac{\partial f_X(x \mid \theta)}{\partial \theta}\right|, \left\|\frac{\partial^2 f_X(x \mid \theta)}{\partial \theta \partial \theta'}\right\| \leq b(x)$$

for any $(x, \theta) \in E \times \Theta^o$.

Define $h : E \to \mathbb{R}$ as

$$h(x) = \inf_{\theta \in \Theta_0} f_X(x \mid \theta)$$

for any $x \in E$, which is a continuous and positive-valued function by Berge's theorem and the extreme value theorem, respectively. Assume that the mapping $x \mapsto \left(\frac{b(x)}{h(x)}\right)^2$ is $\mu(\theta_0, \cdot)$-integrable. Because $\mu(\theta_0, \cdot)$ is a finite measure on $(E, \mathcal{E})$, it follows from Hölder's inequality that the mapping $x \mapsto \frac{b(x)}{h(x)}$ is also $\mu(\theta_0, \cdot)$-integrable.

IID.M7 : **Smoothness Properties of Hessian**

There exist functions $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ and $\beta : E \to \mathbb{R}_+$ such that

- α is continuous at 0 with $\alpha(0) = 0$
- β is $\mu(\theta_0, \cdot)$ integrable
- For any $x \in E$ and $\theta_1, \theta_2 \in \Theta_0$,

$$\|H(x \mid \theta_1) - H(x \mid \theta_2)\| \leq \beta(x) \cdot \alpha(|\theta_1 - \theta_2|).$$

We pause here to show that the four conditions above imply assumptions M1 to M6 thanks to the Gibbs inequality. This demonstrates that the intuitive identification condition is indeed sufficient to statistically identify the model when our data is i.i.d., and that the information matrix equality, as well as the Cramer-Rao Bound, hold as well. We first show that the log-likelihood satisfies the following properties:

Lemma 3.24 (Continuity Properties of the Log-Likelihood in the IID Case) Suppose that assumptions IID.M1 to IID.M6 hold. Then, for any $x \in E$ and $\theta_1, \theta_2 \in \Theta_0$,

$$|l(x \mid \theta_1) - l(x \mid \theta_2)| \leq \frac{b(x)}{h(x)} \cdot |\theta_1 - \theta_2|.$$

In addition, defining the function $g : \Theta_0 \to \mathbb{R}$ as

$$g(\theta) = \mathbb{E}\left[l(X_i \mid \theta)\right] := \int_E l(x \mid \theta)\mu(\theta_0, dx)$$

for any $\theta \in \Theta_0$, g is a continuous function on Θ_0.

Proof) Choose any $x \in E$ and $\theta_1, \theta_2 \in \Theta_0$. By a second-order Taylor expansion of $l(x \mid \theta_1)$ around θ_2, we have

$$l(x \mid \theta_1) = l(x \mid \theta_2) + s(x \mid \theta_2)'(\theta_1 - \theta_2) + \frac{1}{2}(\theta_1 - \theta_2)' H(x \mid \theta^*)(\theta_1 - \theta_2),$$

where θ^* is some convex combination of θ_1 and θ_2. Since $l(x \mid \cdot)$ is strictly concave by assumption IID. M2, it follows that

$$H(x \mid \theta^*) = \frac{\partial^2 l(x \mid \theta^*)}{\partial \theta \partial \theta'}$$

is a negative definite matrix, and as such that

$$\frac{1}{2}(\theta_1 - \theta_2)' H(x \mid \theta^*)(\theta_1 - \theta_2) \leq 0.$$

As such,

$$l(x \mid \theta_1) - l(x \mid \theta_2) = s(x \mid \theta_2)'(\theta_1 - \theta_2) + \frac{1}{2}(\theta_1 - \theta_2)' H(x \mid \theta^*)(\theta_1 - \theta_2)$$

$$\leq s(x \mid \theta_2)'(\theta_1 - \theta_2) \leq |s(x \mid \theta_2)| \cdot |\theta_1 - \theta_2|,$$

where the last inequality is just the Cauchy-Schwarz inequality for inner products. Note that

$$s(x \mid \theta_2) = \frac{\partial f_X(x \mid \theta_2)}{\partial \theta} \cdot \frac{1}{f_X(x \mid \theta_2)}.$$

We can now see that

$$|s(x \mid \theta_2)| = \left|\frac{\partial f_X(x \mid \theta_2)}{\partial \theta}\right| \cdot \frac{1}{f_X(x \mid \theta_2)}$$

$$\leq \frac{b(x)}{f_X(x \mid \theta_2)} \leq \frac{b(x)}{h(x)}.$$

Therefore,

$$l(x \mid \theta_1) - l(x \mid \theta_2) \leq |s(x \mid \theta_2)| \cdot |\theta_1 - \theta_2| \leq \frac{b(x)}{h(x)} \cdot |\theta_1 - \theta_2|.$$

Similarly, by expanding $l(x \mid \theta_2)$ around θ_1, we can see that the inequality

$$l(x \mid \theta_2) - l(x \mid \theta_1) \leq \frac{b(x)}{h(x)} \cdot |\theta_1 - \theta_2|$$

holds, from which we can conclude that

$$|l(x \mid \theta_1) - l(x \mid \theta_2)| \leq \frac{b(x)}{h(x)} \cdot |\theta_1 - \theta_2|.$$

It is now a simple matter to establish the continuity of the function g. Choose any $\theta \in \Theta_0$, and let $\{\theta_n\}_{n \in N_+}$ be a sequence in Θ_0 converging to θ. Define the function $G_n : E \to \mathbb{R}$ as

$$G_n(x) = l(x \mid \theta_n) \quad \text{for any } n \in N_+, x \in E.$$

Then, $\{G_n\}_{n \in N_+}$ is a sequence of measurable functions on E that converges pointwsie to $l(x \mid \theta)$ by the continuity of the log-likelihood. Furthermore,

$$|G_n(x)| = |l(x \mid \theta_n)| \leq \frac{b(x)}{h(x)} \cdot |\theta_n - \theta| + |l(x \mid \theta)|$$

for any $x \in E$ and $n \in N_+$ by the inequality we just derived above. Since $l(\cdot \mid \theta)$ and the mapping $x \mapsto \frac{b(x)}{h(x)}$ are both $\mu(\theta_0, \cdot)$-integrable by assumption, it follows that the function $\beta : E \to \mathbb{R}_+$ defined as

$$\beta(x) = \frac{b(x)}{h(x)} \cdot |\theta_n - \theta| + |l(x \mid \theta)|$$

for any $x \in E$ is a $\mu(\theta_0, \cdot)$-integrable non-negative function. By the DCT, we now have

$$\lim_{n \to \infty} g(\theta_n) = \lim_{n \to \infty} \int_E G_n(x) \mu(\theta_0, dx)$$

$$= \int_E \Big(\lim_{n \to \infty} G_n(x) \Big) \mu(\theta_0, dx) = \int_E l(x \mid \theta) \mu(\theta_0, dx) = g(\theta).$$

It follows that g is continuous on Θ_0.

Q.E.D.

Corollary to Lemma 3.24 (Uniform WLLN for Log-Likelihood)
Suppose that assumptions IID.M1 to IID.M6 hold. Then, we have

$$\sup_{\theta\in\Theta_0}\left|\frac{1}{n}\sum_{i=1}^{n} l(X_i \mid \theta) - g(\theta)\right| \xrightarrow{p} 0.$$

as $n \to \infty$.

Proof) To establish the desired result, we need only rely on theorem 3.4, the uniform WLLN. We verify the assumptions one by one.

First, Θ_0 is a comapct convex subset of $\mathbb{R}^m$. We also established that $g : \Theta_0 \to \mathbb{R}$ is continuous in the preceding lemma.

Define the function $\gamma : E \to \mathbb{R}_+$ as

$$\gamma(x) = \frac{b(x)}{h(x)} \quad \text{for any } x \in E.$$

By the $\mu(\theta_0, \cdot)$-integrability of the function γ and the i.i.d. property of $\{X_i\}_{i\in N_+}$, $\{\gamma(X_i)\}_{i\in N_+}$ is an i.i.d. sequence of real random variables with common mean $\mathbb{E}[\gamma(X_1)]$. By Khinchin's WLLN,

$$\frac{1}{n}\sum_{i=1}^{n}\gamma(X_i) \xrightarrow{p} \mathbb{E}[\gamma(X_1)],$$

so that

$$\frac{1}{n}\sum_{i=1}^{n}\gamma(X_i) = O_p(1) \quad \text{as } n \to \infty.$$

We showed in the preceding lemma that

$$|l(x \mid \theta_1) - l(x \mid \theta_2)| \le \gamma(x) \cdot |\theta_1 - \theta_2|$$

for any $x \in E$ and $\theta_1, \theta_2 \in \Theta_0$. This demonstrates that the log-likelihood satisfies the requisite Lipschitz continuity assumption.

Finally, for any $\theta \in \Theta_0$, because $\{X_i\}_{i\in N_+}$ is i.i.d., so is $\{l(X_i \mid \theta)\}_{i\in N_+}$. In addition, by the integrability assumption, each $l(X_i \mid \theta)$ is an integrable real random variable with

common mean

$$\mathbb{E}\left[l(X_i \mid \theta)\right] = g(\theta).$$

Khinchin's WLLN for i.i.d. sequences of random variables with finite first moments now tells us that

$$\frac{1}{n}\sum_{i=1}^{n} l(X_i \mid \theta) \xrightarrow{p} g(\theta).$$

We have shown that assumptions 1 to 5 of theorem 3.4. are satisfied; therefore, by the uniform WLLN,

$$\sup_{\theta\in\Theta_0}\left|\frac{1}{n}\sum_{i=1}^{n} l(X_i \mid \theta) - g(\theta)\right| \xrightarrow{p} 0$$

as $n \to \infty$,where the quantity on the left hand side is properly $\mathcal{H}$-measurable.

Q.E.D.

It is now easy to establish that the asymptotic identification condition follows from the intuitive identification condition. Formally showing this is the content of the next theorem:

Theorem 3.25 (Asymptotic Identification Condition in the IID Case)
Suppose that assumptions IID.M1 to IID.M6 hold. Then, the asymptotic identification condition holds, that is, for any $\epsilon > 0$ such that $\Theta_0 \setminus B(\theta_0, \epsilon) \neq \emptyset$, there exists some $K(\epsilon) > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\theta\in\Theta_0\setminus B(\theta_0,\epsilon)} \frac{1}{\sqrt{n}}\Big[l_n(X_1,\cdots,X_n \mid \theta) - l_n(X_1,\cdots,X_n \mid \theta_0)\Big] \leq -K(\epsilon)\right) = 1.$$

Proof) For any $\theta \in \Theta_0$, since $l(\cdot \mid \theta)$ and $l(\cdot \mid \theta_0)$ are $\mu(\theta_0, \cdot)$-integrable by assumption IID.M2, the version of the Gibbs inequality with $n = 1$ shows us that

$$\int_E l(x \mid \theta)\mu(\theta_0, dx) \leq \int_E l(x \mid \theta_0)\mu(\theta_0, dx),$$

with the inequality holding as an equality if and only if $\theta = \theta_0$. Therefore, the continuous function $g : \Theta_0 \to \mathbb{R}$ defined as

$$g(\theta) = \mathbb{E}\left[l(X_i \mid \theta)\right] := \int_E l(x \mid \theta)\mu(\theta_0, dx)$$

for any $\theta \in \Theta_0$ is uniquely maximized at θ_0.

Choose any $\epsilon > 0$ such that $\Theta_0 \setminus B(\theta_0, \epsilon) \neq \emptyset$. Denoting $\Theta_1 = \Theta_0 \setminus B(\theta_0, \epsilon)$, Θ_1 is a

non-empty closed subset of the compact set Θ_0 and as such is compact itself. Since

$$\sup_{\theta\in\Theta_1}\left|\frac{1}{n}\sum_{i=1}^{n}l(X_i\mid\theta)-g(\theta)\right|\leq\sup_{\theta\in\Theta_0}\left|\frac{1}{n}\sum_{i=1}^{n}l(X_i\mid\theta)-g(\theta)\right|,$$

it follows from the uniform WLLN above that

$$\sup_{\theta\in\Theta_1}\left|\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta)\right|\xrightarrow{p}0.$$

Since g is continuous on Θ_0 and Θ_1 is a compact subset of Θ_0, there exists some $\theta^*\in\Theta_1$ such that

$$g(\theta^*)=\sup_{\theta\in\Theta_1}g(\theta).$$

This θ^* is not equal to θ_0 by the way that Θ_1 is constructed, so

$$K(\epsilon)=\frac{1}{2}\left(g(\theta_0)-g(\theta^*)\right)>0.$$

Meanwhile, by Khinchin's WLLN,

$$\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta_0)-g(\theta_0)\xrightarrow{p}0$$

It follows that

$$\sup_{\theta\in\Theta_1}\left|\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta)\right|-\left[\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta_0)-g(\theta_0)\right]\xrightarrow{p}0.$$

For any $n\in N_+$, if

$$\sup_{\theta\in\Theta_1}\left|\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta)\right|-\left[\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta_0)-g(\theta_0)\right]\leq K(\epsilon),$$

then because

$$\begin{aligned}\sup_{\theta\in\Theta_1}\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta^*)&\leq\sup_{\theta\in\Theta_1}\left(\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta)\right)\\&\leq\sup_{\theta\in\Theta_1}\left|\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta)\right|,\end{aligned}$$

we have

$$\sup_{\theta\in\Theta_1}\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta)-g(\theta^*)-\frac{1}{n}l_n(X_1,\cdots,X_n\mid\theta_0)+g(\theta_0)\leq K(\epsilon).$$

Since $g(\theta_0) - g(\theta^*) = 2K(\epsilon)$, subtracting $2K(\epsilon)$ from both siees shows us that

$$\sup_{\theta \in \Theta_1} \frac{1}{n} l_n(X_1, \cdots, X_n \mid \theta) - \frac{1}{n} l_n(X_1, \cdots, X_n \mid \theta_0) \leq -K(\epsilon).$$

Finally, pre-multiplying both sides by $\sqrt{n}$ shows us that

$$\sup_{\theta \in \Theta_1} \frac{1}{\sqrt{n}} \Big[l_n(X_1, \cdots, X_n \mid \theta) - l_n(X_1, \cdots, X_n \mid \theta_0) \Big] \leq -\sqrt{n} \cdot K(\epsilon) \leq -K(\epsilon),$$

where the last inequality follows because

$$\sqrt{n} K(\epsilon) \geq K(\epsilon).$$

We have just shown that

$$\mathbb{P}\left(\sup_{\theta \in \Theta_1} \left| \frac{1}{n} l_n(X_1, \cdots, X_n \mid \theta) - g(\theta) \right| - \left[\frac{1}{n} l_n(X_1, \cdots, X_n \mid \theta_0) - g(\theta_0) \right] \leq K(\epsilon \right)$$

$$\leq \mathbb{P}\left(\sup_{\theta \in \Theta_1} \frac{1}{\sqrt{n}} \Big[l_n(X_1, \cdots, X_n \mid \theta) - l_n(X_1, \cdots, X_n \mid \theta_0) \Big] \leq -K(\epsilon) \right)$$

for any $n \in N_+$. By the convergence in probability result derived above, along with the fact that $K(\epsilon) > 0$, the left hand side goes to 1 as $n \to \infty$. Therefore,

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{\theta \in \Theta_1} \frac{1}{\sqrt{n}} \Big[l_n(X_1, \cdots, X_n \mid \theta) - l_n(X_1, \cdots, X_n \mid \theta_0) \Big] \leq -K(\epsilon) \right) = 1.$$

Such a $K(\epsilon)$ exists for any $\epsilon > 0$ such that $\Theta_1 = \Theta_0 \setminus B(\theta_0, \epsilon)^c \neq \emptyset$, so the proof is complete.

Q.E.D.

We actually showed a stronger result above than the asymptotic identification condition, namely that for any $\epsilon > 0$ there exists a $K(\epsilon) > 0$ such that

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{\theta \in \Theta_0 \setminus B(\theta_0, \epsilon)} \frac{1}{n} \Big[l_n(X_1, \cdots, X_n \mid \theta) - l_n(X_1, \cdots, X_n \mid \theta_0) \Big] \leq -K(\epsilon) \right) = 1.$$

Thus, in our i.i.d. case the log-likelihood becomes stable at the rate n instead of the slower rate $\sqrt{n}$.

As for the remaining assumptions, assumptions IID.M1, IID.M2, and IID.M3 all immediately imply their general counterparts M1, M2 and M3. Meanwhile, we can see that the n-observations score and Hessian functions are given as

$$s_n(x_1, \cdots, x_n \mid \theta) = \sum_{i=1}^{n} s(x_i \mid \theta)$$

$$H_n(x_1, \cdots, x_n \mid \theta) = \sum_{i=1}^{n} H(x_i \mid \theta)$$

for any $(x_1, \cdots, x_n) \in E^n$ and $\theta \in \Theta^o$, so that $s_n(\cdot \mid \theta) s_n(\cdot \mid \theta)'$ and $H_n(\cdot \mid \theta)$ are both integrable with respect to the product measure

$$\underbrace{\mu(\theta, \cdot) \times \cdots \times \mu(\theta, \cdot)}_{n}$$

for any $\theta \in \Theta^o$. In addition, for any $(x_1, \cdots, x_n) \in E^n$ and $\theta \in \Theta^o$, by the product rule

$$\frac{\partial f_n(x_1, \cdots, x_n \mid \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial f_X(x_i \mid \theta)}{\partial \theta} \cdot \left(\prod_{j \neq i} f_X(x_j \mid \theta) \right).$$

As such,

$$\begin{aligned} \left| \frac{\partial f_n(x_1, \cdots, x_n \mid \theta)}{\partial \theta} \right| &\leq \sum_{i=1}^{n} \left| \frac{\partial f_X(x_i \mid \theta)}{\partial \theta} \right| \left(\prod_{j \neq i} |f_X(x_j \mid \theta)| \right) \\ &\leq \sum_{i=1}^{n} \left(\prod_{j=1}^{n} b(x_j) \right) = n \cdot \prod_{i=1}^{n} b(x_i). \end{aligned}$$

Here,

$$\mathbb{E}\left[n \cdot \prod_{i=1}^{n} b(X_i) \right] = n \cdot \prod_{i=1}^{n} \mathbb{E}\left[b(X_i) \right]$$

by the independence of $X_1, \cdots, X_n$, which shows us that the mapping

$$(x_1, \cdots, x_n) \mapsto n \cdot \prod_{i=1}^{n} b(x_i)$$

is appropriately integrable.

Since

$$\frac{\partial f_n(x_1, \cdots, x_n \mid \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial f_X(x_i \mid \theta)}{\partial \theta} \left(\prod_{j \neq i} f_X(x_j \mid \theta) \right),$$

we can see from the product rule of differentiation that

$$\begin{aligned} \frac{\partial^2 f_n(x_1, \cdots, x_n \mid \theta)}{\partial \theta \partial \theta'} &= \frac{\partial}{\partial \theta} \left(\frac{\partial f_n(x_1, \cdots, x_n \mid \theta)}{\partial \theta} \right) \\ &= \sum_{i=1}^{n} \frac{\partial^2 f_X(x_i \mid \theta)}{\partial \theta \partial \theta'} \left(\prod_{j \neq i} f_X(x_j \mid \theta) \right) \end{aligned}$$

$$+\sum_{i=1}^{n}\sum_{j\neq i}\left(\frac{\partial f_X(x_i \mid \theta)}{\partial \theta}\right)\left(\frac{\partial f_X(x_j \mid \theta)}{\partial \theta}\right)'\left(\prod_{k\neq j,i} f_X(x_k \mid \theta)\right).$$

It follows that

$$\begin{aligned}\left\|\frac{\partial^2 f_n(x_1,\cdots,x_n \mid \theta)}{\partial\theta\partial\theta'}\right\| &\leq \sum_{i=1}^{n}\left|\frac{\partial^2 f_X(x_i \mid \theta)}{\partial\theta\partial\theta'}\right|\cdot\left(\prod_{j\neq i}|f_X(x_j \mid \theta)|\right)\\ &\quad+\sum_{i=1}^{n}\sum_{j\neq i}\left|\frac{\partial f_X(x_i \mid \theta)}{\partial\theta}\right|\cdot\left|\frac{\partial f_X(x_j \mid \theta)}{\partial\theta}\right|\cdot\left(\prod_{k\neq j,i}|f_X(x_k \mid \theta)|\right)\\ &\leq 2\sum_{i=1}^{n}\prod_{j=1}^{n} b(x_j) = 2n\cdot\prod_{i=1}^{n} b(x_i).\end{aligned}$$

As such, $\frac{\partial^2 f_n(x_1,\cdots,x_n|\theta)}{\partial\theta\partial\theta'}$ is also dominated by an integrable function on E^n for any $\theta\in\Theta$. We have just shown that the general assumptions M5 and M6 follow from the IID assumptions IID.M5 and IID. M6.

By implication, we can see that the information matrix equality holds:

$$\mathbb{E}\left[s_n(X_1,\cdots,X_n \mid \theta)s(X_1,\cdots,X_n \mid \theta)'\right] = -\mathbb{E}\left[H_n(X_1,\cdots,X_n \mid \theta)\right]$$

for any $n\in N_+$ and $\theta\in\Theta^o$. Defining

$$\mathbb{I}_n(\theta) = \mathbb{E}\left[s_n(X_1,\cdots,X_n \mid \theta)s(X_1,\cdots,X_n \mid \theta)'\right]$$

as the n-observation information matrix, and letting $\mathbb{I}(\theta)$ denote the single observation information matrix, the information matrix equality implies that

$$\begin{aligned}\mathbb{I}_n(\theta) = -\mathbb{E}\left[H_n(X_1,\cdots,X_n \mid \theta)\right] &= -\sum_{i=1}^{n}\mathbb{E}\left[H(X_i \mid \theta)\right]\\ &= \sum_{i=1}^{n}\mathbb{E}\left[s(X_i \mid \theta)s(X_i \mid \theta)'\right] = n\cdot\mathbb{I}(\theta),\end{aligned}$$

so that the n-observation information matrix is simply n times the single observation information matrix in the i.i.d. case. We can therefore see that the Cramer-Rao Bound for the variance of an unbiased estimator is given as

$$\mathbb{I}_n(\theta_0)^{-1} = \frac{1}{n}\mathbb{I}(\theta_0)^{-1}.$$

Finally, by Khinchin's WLLN,

$$-\frac{1}{n}\sum_{i=1}^{n} H(x_i \mid \theta) \xrightarrow{p} -\mathbb{E}\left[H(X_i \mid \theta)\right] = \mathbb{I}(\theta)$$

for any $\theta\in\Theta$. It turns out, as we show below, that the single observation information matrix is

continuous on the compact set Θ_0.

Lemma 3.26 (Continuity of Information Matrix)
Suppose that assumptions IID.M1 to IID.M6 hold. Then, the function $\mathbb{I} : \Theta_0 \to \mathbb{R}^{m \times m}$ defined as

$$\mathbb{I}(\theta) = \mathbb{E}\left[s(X_i \mid \theta) s(X_i \mid \theta)'\right] := \int_E s(x \mid \theta) s(x \mid \theta)' \mu(\theta_0, dx)$$

for any $\theta \in \Theta_0$ is continuous on Θ_0.

Proof) Note first that, for any $x \in E$ and $\theta \in \Theta_0$,

$$s(x \mid \theta) = \frac{1}{f_X(x \mid \theta)} \cdot \frac{\partial f_X(x \mid \theta)}{\partial \theta},$$

so that

$$s(x \mid \theta) s(x \mid \theta)' = \frac{1}{f_X(x \mid \theta)^2} \left(\frac{\partial f_X(x \mid \theta)}{\partial \theta}\right) \left(\frac{\partial f_X(x \mid \theta)}{\partial \theta}\right)'$$

and

$$|s(x \mid \theta)| \leq \left(\frac{b(x)}{h(x)}\right)^2$$

by assumption IID.M6.

Now choose any $\theta \in \Theta_0$, and let $\{\theta_n\}_{n \in N_+}$ be a sequence in Θ_0 that converges to θ. Defining $\{g_n\}_{n \in N_+}$ as the sequence of functions on E such that

$$g_n(x) = s(x \mid \theta_n) s(x \mid \theta_n)' \quad \text{for any } x \in E, n \in N_+,$$

we can see that

$$\|g_n(x)\| \leq \left(\frac{b(x)}{h(x)}\right)^2$$

for any $x \in E$ and $n \in N_+$. By assumption, the mapping $x \mapsto \left(\frac{b(x)}{h(x)}\right)^2$ is $\mu(\theta_0, \cdot)$-integrable, so that $\{g_n\}_{n \in N_+}$ is pointwise dominated by an integrable non-negative function. Furthermore, by the continuity of the score function on $E \times \Theta^o$, $\{g_n\}_{n \in N_+}$ converges pointwise to $s(x \mid \theta) s(x \mid \theta)'$. By the DCT, we now have

$$\begin{aligned}
\lim_{n \to \infty} \int_E g_n(x) \mu(\theta_0, dx) &= \int_E \left(\lim_{n \to \infty} g_n(x)\right) \mu(\theta_0, dx) \\
&= \int_E s(x \mid \theta) s(x \mid \theta)' \mu(\theta_0, dx) = \mathbb{I}(\theta).
\end{aligned}$$

Since

$$\mathbb{I}(\theta_n) = \mathbb{E}\left[s(X_i \mid \theta_n)s(X_i \mid \theta_n)'\right] = \int_E g_n(x)\mu(\theta_0, dx)$$

for any $n \in N_+$, the result above can be written as

$$\lim_{n\to\infty} \mathbb{I}(\theta_n) = \mathbb{I}(\theta).$$

This shows us that $\mathbb{I}$ is a continuous function on Θ_0.

Q.E.D.

Corollary to Lemma 3.26 (Uniform WLLN for Hessian)
Suppose that assumptions IID.M1 to IID.M7 hold. Then,

$$\sup_{\theta\in\Theta_0} \left\| \frac{1}{n} H_n(X_1, \cdots, X_n \mid \theta) + \mathbb{I}(\theta) \right\| \xrightarrow{p} 0$$

as $n \to \infty$.

Proof) Once again, we need only verify the conditions of theorem 3.4, the uniform WLLN.

First, Θ_0 is a compact convex subset of $\mathbb{R}^m$, and $\mathbb{I} : \Theta_0 \to \mathbb{R}^{m\times m}$ is a continuous function. Furthermore, by the $\mu(\theta_0, \cdot)$-integrability of the function β and the i.i.d. property of $\{X_i\}_{i\in N_+}$, $\{\beta(X_i)\}_{i\in N_+}$ is an i.i.d. sequence of real random variables with common mean $\mathbb{E}[\beta(X_1)]$. By Khinchin's WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} \beta(X_i) \xrightarrow{p} \mathbb{E}\left[\beta(X_1)\right],$$

so that

$$\frac{1}{n}\sum_{i=1}^{n} \beta(X_i) = O_p(1) \quad \text{as } n \to \infty.$$

By assumption IID.M7, there exists a measurable function $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ that is continuous at 0 with $\alpha(0) = 0$ such that

$$\|H(x \mid \theta_1) - H(x \mid \theta_2)\| \le \beta(x) \cdot \alpha(|\theta_1 - \theta_2|)$$

for any $x \in E$ and $\theta_1, \theta_2 \in \Theta_0$.

Finally, for any $\theta \in \Theta_0$, $\{H(X_i \mid \theta)\}_{i\in N_+}$ is an i.i.d. sequence of $m \times m$ random matrices

with common mean

$$\mathbb{E}\left[H(X_i \mid \theta)\right] = -\mathbb{I}(\theta).$$

Therefore, by Khinchin's WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} H(X_i \mid \theta) \xrightarrow{p} -\mathbb{I}(\theta).$$

We have shown that assumptions 1 to 5 of the uniform WLLN have been satisfied; therefore,

$$\sup_{\theta \in \Theta_0} \left\| \frac{1}{n}\sum_{i=1}^{n} H(X_i \mid \theta) + \mathbb{I}(\theta) \right\| \xrightarrow{p} 0$$

as $n \to \infty$, as we desired.

Q.E.D.

As usual, we define the sample information matrix as

$$\begin{aligned}\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n) &= -\frac{1}{n} H_n(X_1, \cdots, X_n \mid \hat{\theta}_n(X_1, \cdots, X_n)) \\ &= -\frac{1}{n}\sum_{i=1}^{n} H(X_i \mid \hat{\theta}_n(X_1, \cdots, X_n)).\end{aligned}$$

By the consistency of the ML estimator and a WLLN-type argument, we would expect $\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n)$ to converge in probability to the information matrix $\mathbb{I}(\theta_0)$, which we can show is positive definite. In addition, the uniform WLLN result above suggests that $\frac{1}{n} H_n(X_1, \cdots, X_n \mid \theta)$ can be made arbitrarily close to $-\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n)$ in a neighborhood of θ_0 when n is large. These correspond to general assumptions M7 and M8, and the theorem below formally show that they hold.

Theorem 3.27 (Asymptotic Properties of the Hessian)
Suppose that assumptions IID.M1 to IID.M7 hold. Then, $\mathbb{I}(\theta_0)$ is a positive definite $m \times m$ matrix and

$$\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n) \xrightarrow{p} \mathbb{I}(\theta_0).$$

Furthermore, for any $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\theta \in \Theta_0 \cap \bar{B}(\theta_0, \delta(\epsilon))} \left\| \frac{1}{n} H_n(X_1, \cdots, X_n \mid \theta) + \frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n) \right\| \leq \epsilon \right) = 1.$$

Proof) **Positive Definiteness of Information Matrix**

First, note that, because $l(x \mid \cdot)$ is a strictly concave function on Θ_0 for any $x \in E$, its second derivative $H(x \mid \cdot)$ is negative definite-valued on Θ_0. In particular, $H(x \mid \theta_0)$ is negative definite for any $x \in E$, which implies that

$$\mathbb{I}(\theta_0) = -\mathbb{E}\left[H(X_i \mid \theta_0)\right]$$

is positive semidefinite-valued; to see this, note that, for any $v \in \mathbb{R}^m$,

$$v'\mathbb{I}(\theta_0)v = \mathbb{E}\left[v'(-H(X_i \mid \theta_0))v\right] \geq 0$$

because the postive definite-valuedness of $-H(\cdot \mid \theta_0)$ on E implies that the mapping $x \mapsto v'(-H(x \mid \theta_0))v$ on E is non-negative valued.

Now suppose that $\mathbb{I}(\theta_0)$ is not positive definite. Then, there exists some non-zero $v \in \mathbb{R}^m$ such that

$$v'\mathbb{I}(\theta_0)v = \mathbb{E}\left[v'(-H(X_i \mid \theta_0))v\right] = 0.$$

By the vanishing property for non-negative functions,

$$v'(-H(x \mid \theta_0))v = 0$$

for any $x \in A \in \mathcal{E}$, where $\mu(\theta_0, A) = 1$. However, since $-H(x \mid \theta_0)$ is positive definite for any $x \in E$, it follows that $v'(-H(x \mid \theta_0))v > 0$ for any $x \in E$, or in other words that $A = \emptyset$. This contradicts the fact that $\mu(\theta_0, A) = 1$, so that $\mathbb{I}(\theta_0)$ must be positive definite.

Probability Limit of Sample Information Matrix

By the uniform WLLN result, we can see that

$$\sup_{\theta\in\Theta_0}\left\|\left[-\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)\right]-\mathbb{I}(\theta)\right\|\xrightarrow{p}0$$

as $n\to\infty$. Since the function $\hat{\theta}_n : E^n\to\Theta$ takes values in $\Theta_0\subset\Theta^o$, it follows that

$$\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)-\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n))\xrightarrow{p}O_{m\times m}.$$

By the continuity of the function $\mathbb{I}:\Theta_0\to\mathbb{R}^{m\times m}$, the continuous mapping theorem, and the consistency of the ML estimator,

$$\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n))\xrightarrow{p}\mathbb{I}(\theta_0).$$

Putting these two results together, we have

$$\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\xrightarrow{p}\mathbb{I}(\theta_0).$$

Uniform Continuity of Hessian

Choose any $\epsilon>0$. Then, by the continuity of the function $\mathbb{I}:\Theta_0\to\mathbb{R}^{m\times m}$ (which is actually uniformly continuous because Θ_0 is compact), there exists a $\delta>0$ such that

$$\|\mathbb{I}(\theta_1)-\mathbb{I}(\theta_2)\|\le\frac{\epsilon}{3}$$

for any $\theta_1,\theta_2\in\Theta_0$ such that $|\theta_1-\theta_2|\le 2\delta$. Furthermore, by the consistency of the ML estimator,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right|\le\delta\right)=1.$$

Finally, by the uniform WLLN for the Hessian,

$$\lim_{n\to\infty}\mathbb{P}\left(\sup_{\theta\in\Theta_0}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\mathbb{I}(\theta)\right\|\le\frac{\epsilon}{2}\right)=1.$$

Suppose $\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right|\le\delta$ and

$$\sup_{\theta\in\Theta_0}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\mathbb{I}(\theta)\right\|\le\frac{\epsilon}{3}.$$

Then,

$$\left|\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)-\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n))\right|\leq\frac{\epsilon}{3},$$

where we used the fact that the ML estimator takes values in Θ_0.

Next, for any $\theta\in\Theta_0\cap\bar{B}(\theta_0,\delta)$, since

$$\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta\right|\leq\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right|+|\theta-\theta_0|\leq 2\delta,$$

we have

$$\left\|\mathbb{I}(\theta)-\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n))\right\|\leq\frac{\epsilon}{3}.$$

It follows from the triangle inequality that

$$\begin{aligned}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right\|&\leq\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\mathbb{I}(\theta)\right\|\\&\quad+\left\|\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n))-\mathbb{I}(\theta)\right\|\\&\quad+\left\|\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)-\mathbb{I}(\hat{\theta}_n(X_1,\cdots,X_n)\right\|\\&\leq\frac{\epsilon}{3}+\frac{\epsilon}{3}+\frac{\epsilon}{3}\leq\epsilon.\end{aligned}$$

This holds for any $\theta\in\Theta_0\cap\bar{B}(\theta_0,\delta)$, so

$$\sup_{\theta\in\Theta_0\cap\bar{B}(\theta_0,\delta)}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right\|\leq\epsilon.$$

We have just shown that

$$\begin{aligned}&\left\{\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right|\leq\delta\right\}\cap\left\{\sup_{\theta\in\Theta_0}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\mathbb{I}(\theta)\right\|\leq\frac{\epsilon}{3}\right\}\\&\qquad\subset\left\{\sup_{\theta\in\Theta_0\cap\bar{B}(\theta_0,\delta)}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right\|\leq\epsilon\right\},\end{aligned}$$

and in particular, taking expectations on both sides, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right|\leq\delta\right)+\mathbb{P}\left(\sup_{\theta\in\Theta_0}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\mathbb{I}(\theta)\right\|\leq\frac{\epsilon}{3}\right)-1\\&\qquad\leq\mathbb{P}\left(\sup_{\theta\in\Theta_0\cap\bar{B}(\theta_0,\delta)}\left\|\frac{1}{n}H_n(X_1,\cdots,X_n\mid\theta)+\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right\|\leq\epsilon\right).\end{aligned}$$

Taking $n \to \infty$ on both sides shows us that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\theta \in \Theta_0 \cap \bar{B}(\theta_0,\delta)} \left\| \frac{1}{n} H_n(X_1, \cdots, X_n \mid \theta) + \frac{1}{n} \hat{\mathbb{I}}_n(X_1, \cdots, X_n) \right\| \leq \epsilon \right) = 1,$$

as we desired.

Q.E.D.

Corollary to Theorem 3.28 (Boundedness of Eigenvalues of Sample Information Matrix)
Suppose that assumptions IID.M1 to IID.M7 hold, and let $q_{n,min}(x)$ and $q_{n,max}(x)$ be the smallest and largest eigenvalues of $\frac{1}{n}\hat{\mathbb{I}}_n(x)$ for any $n \in N_+$ and $x \in E^n$. Then, there exist $0 < \rho_{\min} < \rho_{max}$ such that

$$\lim_{n\to\infty} \mathbb{P}(\rho_{\min} \leq q_{n,min}(X_1, \cdots, X_n) \leq q_{n,max}(X_1, \cdots, X_n) \leq \rho_{max}) = 1.$$

Proof) By the preceding result, we showed that

$$\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n) \xrightarrow{p} \mathbb{I}(\theta_0),$$

where $\mathbb{I}(\theta_0)$ is a positive definite matrix. Let ρ_{min} and ρ_{max} be the smallest and largest eigenvalues of $\mathbb{I}(\theta_0)$, respectively; by the positive definiteness of $\mathbb{I}(\theta_0)$, $\rho_{min} > 0$.

Since the ordered eigenvalues of a matrix are continuous functions of that matrix, the continuous mapping theorem shows us that

$$q_{n,min}(X_1, \cdots, X_n) \xrightarrow{p} \rho_{min} \quad \text{and} \quad q_{n,max}(X_1, \cdots, X_n) \xrightarrow{p} \rho_{max}.$$

Since $\frac{1}{2}\rho_{min} > 0$, by the definition of convergence in probability

$$\lim_{n\to\infty} \mathbb{P}\left(|q_{n,min}(X_1, \cdots, X_n) - \rho_{min}| \leq \frac{1}{2}\rho_{min} \right) = 1.$$

By implication,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{2}\rho_{min} \leq q_{n,min}(X_1, \cdots, X_n) \right) = 1.$$

Similarly,

$$\lim_{n\to\infty} \mathbb{P}(|q_{n,max}(X_1, \cdots, X_n) - \rho_{max}| \leq \rho_{max}) = 1$$

and thus

$$\lim_{n\to\infty} \mathbb{P}(q_{n,max}(X_1, \cdots, X_n) \leq 2\rho_{max}) = 1.$$

Putting these results together, we can conclude that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{2}\rho_{min} \leq q_{n,min}(X_1, \cdots, X_n) \leq q_{n,max}(X_1, \cdots, X_n) \leq 2\rho_{max}\right) = 1,$$

where $0 < \frac{1}{2}\rho_{min} < 2\rho_{max}$.

Q.E.D.

We have shown that general assumptions M7 and M8 hold under IID.M1 to IID.M7. Finally, general assumption M9 follows from the Lindeberg-Levy CLT; since $\{s(X_i \mid \theta_0)\}_{i\in N_+}$ is a sequence of i.i.d. random vectors with common mean 0 (by the information matrix equality) and variance

$$\mathbb{E}\left[s(X_i \mid \theta_0)s(X_i \mid \theta_0)'\right] = \mathbb{I}(\theta_0),$$

the CLT shows us that

$$\frac{1}{\sqrt{n}} s_n(X_1, \cdots, X_n \mid \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i \mid \theta_0) \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ \mathbb{I}(\theta_0)\right].$$

Since we saw above that

$$\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n) \xrightarrow{p} \mathbb{I}(\theta_0),$$

by the continuity of the inversion operation and the Cholesky decomposition operation, along with the continuous mapping theorem, we have

$$\left(\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n)\right)^{-\frac{1}{2}} \xrightarrow{p} \mathbb{I}(\theta_0)^{-\frac{1}{2}}.$$

As such, Slutsky's theorem tells us that

$$\left(\frac{1}{n}\hat{\mathbb{I}}_n(X_1, \cdots, X_n)\right)^{-\frac{1}{2}} \left[\frac{1}{\sqrt{n}} s_n(X_1, \cdots, X_n \mid \theta_0)\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right],$$

which is exactly general assumption M9.

By implication,

$$\sqrt{n}\left[\hat{\theta}_n(X_1, \cdots, X_n) - \theta_0\right] = O_p(1) \quad \text{as } n \to \infty,$$

and the asymptotic distribution of the ML estimator is given as

$$\left(\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right)^{\frac{1}{2}\prime} \times \sqrt{n}\left[\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ I_m\right].$$

By Slutsky's theorem and the fact that

$$\left(\frac{1}{n}\hat{\mathbb{I}}_n(X_1,\cdots,X_n)\right)^{-\frac{1}{2}\prime} \xrightarrow{p} \mathbb{I}(\theta_0)^{-\frac{1}{2}\prime},$$

we can now conclude that

$$\sqrt{n}\left[\hat{\theta}_n(X_1,\cdots,X_n)-\theta_0\right] \xrightarrow{d} \mathcal{N}\left[O_{m\times 1},\ \mathbb{I}(\theta_0)^{-1}\right].$$

In other words, for large n,

$$\hat{\theta}_n(X_1,\cdots,X_n) \approx \mathcal{N}\left[\theta_0,\ \frac{1}{n}\mathbb{I}(\theta_0)^{-1}\right],$$

where $\frac{1}{n}\mathbb{I}(\theta_0)^{-1}$ is the n-observation Cramer-Rao bound. As such, it is clearer in the i.i.d. case that the ML estimator is asymptotically unbiased and asymptotically attains the CRLB.

3.6 Extremum Estimation with Convex Objective Functions

So far, we have assumed the compactness of the parameter space and uniform convergence of the objective function to establish consistency, and required smoothness properties to establish asymptotic normality. In the special case when the objective function is convex in the parameters, it turns out that we can relax these assumptions, so that similar consistency and asymptotic normality results hold even when the parameter space is non-compact, the objective function does not necessarily converge uniformly in probability, and the objective function is non-differentiable at certain points.

3.6.1 Topological Properties of Sequence Spaces

To establish our core result, we must first study the topological properties of the space of sequences, which occupy a middle ground between finite-dimensional euclidean spaces and function spaces. Formally, let $\{(E_i, d_i)\}_{i \in N_+}$ be a collection of metric spaces, with the metric topology induced by d_i denoted by τ_i. Let (E^∞, τ^∞) be the topological space where

$$E^\infty = \{(x_i)_{i \in N_+} \mid x_i \in E_i \quad \forall i \in N_+\},$$

and τ^∞ the product topology $\prod_i \tau_i$.

Suppose each E_i is separable, that is, there exists a countable subset $E_{0,i}$ of E_i that is dense in E_i in the topology τ_i. Then, because separable metric spaces are also second countable, there exists a countable base $\mathbb{B}_i \subset \tau_i$ that generates τ_i for each $i \in N_+$. It follows from the second countability of each topological space and the countability of the index N_+ that the product Borel σ-algebra is exactly the Borel σ-algebra generated by the product topology:

$$\bigotimes_n \mathcal{B}(E_n, \tau_n) = \mathcal{B}(E^\infty, \tau^\infty).$$

To obtain sharper topological results for E^∞, we define a metric on E^∞. First, we define the metric $\tilde{d}_i : E_i \times E_i \to \mathbb{R}_+$ on E_i as

$$\tilde{d}_i(x, y) = \min(d(x, y), 1)$$

for any $x, y \in E_i$. Note that $\tilde{d}_i(\cdot)$ is a well-defined metric on E_i:

Positive Definiteness For any $x, y \in E_i$, if $\tilde{d}_i(x, y) = 0$, then $d(x, y) = 0$ and therefore $x = y$. On the other hand, if $x = y$, then $d(x, y) = 0$ and thus $\tilde{d}_i(x, y) = 0$. In other words, $\tilde{d}_i(x, y) = 0$ if and only if $x = y$.

Symmetry For any $x, y \in E_i$,

$$\tilde{d}_i(x, y) = \min(d(x, y), 1) = \min(d(y, x), 1) = \tilde{d}_i(y, x).$$

Triangle Inequality For any $x, y, z \in E_i$,

$$\tilde{d}_i(x,y) = \min(d(x,y),1) \leq \min(d(x,z)+d(z,y),1).$$

For any $a, b \geq 0$, suppose $\min(a+b,1) > \min(a,1)+\min(b,1)$. Then, because $\min(a+b,1) \leq a+b$, it must be the case that $\min(a,1)+\min(b,1) < a+b$, so that either $\min(a,1) < a$ or $\min(b,1) < b$. Without loss of generality, let $\min(a,1) < a$. Then,

$$1 \geq \min(a+b,1) > \min(a,1)+\min(b,1) = 1+\min(b,1) \geq 1,$$

a contradiction. Therefore, we must have $\min(a+b,1) \leq \min(a,1)+\min(b,1)$, which implies that

$$\tilde{d}_i(x,y) \leq \min(d(x,z),1)+\min(d(z,y),1) = \tilde{d}_i(x,z)+\tilde{d}_i(z,y).$$

Furthermore, and this is the interesting part, the metric topology induced by $\tilde{d}_i$ is precisely that induced by d_i. To see this, choose any sequence $\{x_n\}_{n\in N_+}$ in E_i that converges to some $x \in E_i$ in the metric d_i. Then, the continuity of the minimum function shows us that

$$\lim_{n\to\infty} \tilde{d}_i(x_n,x) = \min\Big(\lim_{n\to\infty} d_i(x_n,x),1\Big) = 0,$$

so that $\{x_n\}_{n\in N_+}$ converges to x in the metric $\tilde{d}_i$ as well. Conversely, suppose that $\{x_n\}_{n\in N_+}$ converges to x in the metric $\tilde{d}_i$. Then, for any $\epsilon \in (0,1)$ there exists an $N \in N_+$ such that

$$d_i(x_n,x) = \tilde{d}_i(x_n,x) < \epsilon \quad \text{for any } n \geq N.$$

Taking $n \to \infty$ on both sides yields $\limsup_{n\to\infty} d_i(x_n,x) \leq \epsilon$. This holds for any $\epsilon > 0$, so $d_i(x_n,x) \to 0$ as $n \to \infty$ and therefore $\{x_n\}_{n\in N_+}$ converges to x in the metric d_i. In other words,

$$x_n \to x \text{ in the metric } d_i \text{ if and only if } x_n \to x \text{ in the metric } \tilde{d}_i.$$

Now choose any subset $B \subset E_i$. B is closed in the metric topology induced by d_i if and only if, for any sequence $\{x_n\}_{n\in N_+}$ that is convergent in d_i, its limit is also contained in B. Because a sequence converges to some limit in d_i if and only if it converges to the same limit in $\tilde{d}_i$, it follows that B is closed in the metric topology induced by d_i if and only if B is closed in the metric topology induced by $\tilde{d}_i$. This proves that d_i and $\tilde{d}_i$ induce exactly the same closed sets, and therefore exactly the same open sets.

Using the so-called minimum metric $\tilde{d}_i$ defined above, we can now construct a metric on E^∞ that induces the product topology τ^∞. It is a generalization of the usual product, or box, metric used to induce the product topology when the product is finite.

Theorem 3.29 (Metric on E^∞)
Let $\{(E_i, d_i)\}_{i \in N_+}$ be a collection of metric spaces and (E^∞, τ^∞) their product. For any $i \in N_+$, let $\tilde{d}_i : E_i \times E_i \to \mathbb{R}_+$ be the minimum metric defined above, and define $d : E^\infty \times E^\infty \to \mathbb{R}_+$ as

$$d(x, y) = \sup_{i \in N_+} \frac{\tilde{d}_i(x_i, y_i)}{i}$$

for any $x, y \in E^\infty$. Then, d is a metric on E^∞ that induces a metric topology exactly equal to the product topology d^∞. Furthermore, if each (E_i, τ_i) is a separable space, then so is (E^∞, d^∞).

Proof) We first show that $d(\cdot)$ is a well-defined metric on E^∞. Note first that, because $\frac{\tilde{d}_i(x_i, y_i)}{i} \leq \frac{1}{i} \leq 1$ for any $x_i, y_i \in E_i$ and $i \in N_+$, it follows that $d(x, y) \leq 1$ for any $x, y \in E^\infty$ and therefore that $d(\cdot)$ is a real-valued non-negative function. Now we check that it satisfies the properties of a metric:

1. For any $x, y \in E^\infty$, if $d(x, y) = 0$ then $\tilde{d}_i(x_i, y_i) = 0$ for any $i \in N_+$, which implies that $x_i = y_i$ for each i and thus that $x = y$. Conversely, if $x = y$, then $d(x, y) = 0$ trivially.
2. For any $x, y \in E^\infty$, $\tilde{d}_i(x_i, y_i) = \tilde{d}_i(y_i, x_i)$ for each i, so that $d(x, y) = d(y, x)$ as well.
3. For any $x, y, z \in E^\infty$, the triangle inequality for $\tilde{d}_i$ tells us that

$$\frac{\tilde{d}_i(x_i, y_i)}{i} \leq \frac{\tilde{d}_i(x_i, z_i)}{i} + \frac{\tilde{d}_i(z_i, y_i)}{i} \leq d(x, z) + d(z, y).$$

   This holds for any i, so we have $d(x, y) \leq d(x, z) + d(z, y)$.

Therefore, d is a well-defined metric on E^∞.

To show that the metric topology induced by d is exactly the product topology τ^∞, first define τ^* as the metric topology induced by d. Letting $\mathcal{M}_0$ be the set of all open rectangles on E^∞, choose some $A \in \mathcal{M}_0$. Then, letting $A = \prod_{i \in N_+} A_i$, there exists an $N \in N_+$ such that $A_i = E$ for any $i > N$. For any $x \in A$, note that $x_i \in A_i$ for $1 \leq i \leq N$, and because each A_i is an open set in τ_i, which is the metric topology induced by $\tilde{d}_i$, there exists an $\epsilon_i > 0$ such that $B_{\tilde{d}_i}(x_i, \epsilon_i) \subset A_i$. Let

$$\epsilon = \min_{1 \leq i \leq N} \frac{\epsilon_i}{i} > 0.$$

Then, for any $z \in B_d(x, \epsilon)$, we can see that

$$\frac{\tilde{d}_i(x_i, z_i)}{i} \leq d(x, y) < \epsilon \leq \frac{\epsilon_i}{i},$$

or equivalently, $\tilde{d}_i(x_i, z_i) < \epsilon_i$, so that $z_i \in B_{\tilde{d}_i}(x_i, \epsilon_i)$ for each $1 \leq i \leq N$. It follows that $z \in A$, so we have shown that $B_d(x, \epsilon) \subset A$. This holds for any $x \in A$, so by the definition of the metric topology, we have $A \in \tau^*$. This in turn holds for any $A \in \mathcal{M}_0$, so $\mathcal{M}_0 \subset \tau^*$

and we have $\tau^\infty \subset \tau^*$.

To show the reverse inclusion, choose any $A \in \tau^*$ and $x \in A$. Then, there exists an $\epsilon > 0$ such that $B_d(x, \epsilon) \subset A$. Now choose any $N \in N_+$ such that $\frac{1}{N} < \epsilon$, and define the open rectangle A_0 as

$$A_0 = B_{\tilde{d}_1}(x_1, 1/N) \times \cdots \times B_{\tilde{d}_N}(x_N, 1/N) \times E \times E \times \cdots .$$

Clearly, $x \in A_0$. Furthermore, for any $z \in A_0$, we can see that

$$\frac{\tilde{d}_i(z_i, x_i)}{i} \leq \frac{1}{Ni} < \frac{1}{N} < \epsilon$$

if $1 \leq i \leq N$ by definition, while if $i > N$,

$$\frac{\tilde{d}_i(z_i, x_i)}{i} \leq \frac{1}{i} < \frac{1}{N} < \epsilon,$$

where the first inequality follows because $\tilde{d}_i$ is bounded above by 1. Therefore,

$$d(x, z) = \sup_{i \in N_+} \frac{\tilde{d}_i(z_i, x_i)}{i} \leq \frac{1}{N} < \epsilon,$$

so that $z \in B_d(x, \epsilon)$. It follows that $x \in A_0 \subset B_d(x, \epsilon) \subset A$. We have just shown that, for any $x \in A$, we can find an $A_x \in \mathcal{M}_0$ such that $x \in A_x \subset A$; therefore,

$$A = \bigcup_{x \in A} A_x \in \tau^\infty.$$

This holds for any $A \in \tau^*$, so we can see that $\tau^* \subset \tau^\infty$.

Finally, to establish the separability claim, suppose that $E_{0,i}$ is a countable dense set in E_i for any $i \in N_+$. We define the set E_0^∞ as

$$E_0^\infty = \{x \in E^\infty \mid x_i \in E_{0,i} \quad \forall i \in N_+\};$$

being the countable union of countable sets, E_0^∞ is itself a countable subset of E^∞. It remains to show that E_0^∞ is dense in E^∞ under the product topology τ^∞. To this end, choose any $x \in E^\infty$ and $\epsilon > 0$. Because $E_{0,i}$ is dense in E_i, we can find an $x_i^0 \in E_{0,i}$ such that

$$\tilde{d}_i(x_i^0, x_i) < \frac{\epsilon i}{2}$$

for any $i \in N_+$. Defining $x_0 = \{x_i^0\}_{i \in N_+} \in E_0^\infty$, we can now see that

$$d(x, x_0) = \sup_{i \in N_+} \frac{\tilde{d}_i(x_i^0, x_i)}{i} \leq \frac{\epsilon}{2} < \epsilon.$$

This proves that E_0^∞ is a countable dense set in E^∞, and therefore that (E^∞, d^∞) is separable.

Q.E.D.

Having established that the product space (E^∞, τ^∞) is metrizable, the next natural step is to investigate how sequences of random variables defined on (E^∞, τ^∞) behave in the limit. It turns out that there exists a very close relationship between the weak convergence of random variables defined on (E^∞, τ^∞) and the weak convergence of their elements.

To this end, we introduce a familiar element: projection functions and finite dimensional distributions. For any $i_1, \cdots, i_k \in N_+$, recall that the projection function $\pi_{i_1,\cdots,i_k} : E^\infty \to E_{i_1} \times \cdots \times E_{i_k}$ is defined as

$$\pi_{i_1,\cdots,i_k}(x) = (x_{i_1}, \cdots, x_{i_k})$$

for any $x \in E^\infty$. It is easy to see that $\pi_{i_1,\cdots,i_k}$ is continuous relative to the product topologies on E^∞ and $E_{i_1} \times \cdots \times E_{i_k}$: for any open rectangle $A = A_{i_1} \times \cdots \times A_{i_k}$ on $\prod_{j=1}^k E_{i_j}$, it follows that

$$\pi^{-1}_{i_1,\cdots,i_k}(A) = \prod_{i \in N_+} B_i,$$

where $B_i = A_{i_j} \in \tau_{i_j}$ if $i = i_j$ for some $1 \leq j \leq k$ and $B_i = E_i$ otherwise. By definition, $\prod_{i \in N_+} B_i$ is an open rectangle on E^∞, so $\pi^{-1}_{i_1,\cdots,i_k}(A) \in \tau^\infty$. Given a probability measure on $(E^\infty, \mathcal{B}(E^\infty, \tau^\infty))$, the collection of pushforward measures

$$\{\mu \circ \pi^{-1}_{i_1,\cdots,i_k} \mid i_1, \cdots, i_k \in N_+\}$$

are the familiar finite-dimensional distributions of μ.

Inverse images of open sets under projection functions assume a cylinder form. Formally, for any $N \in N_+$ we define the collection of all open cylinder sets as

$$\mathcal{C}_N = \Big\{A \times E_{-1:N} \mid A \in \tau_1 \times \cdots \times \tau_N\Big\},$$

where $E_{-1:N} = E_{N+1} \times E_{N+2} \times \cdots$. It follows that, for any $A \in \tau_1 \times \cdots \times \tau_N$,

$$\pi^{-1}_{1,\cdots,N}(A) = A \times E_{-1:N}.$$

Cylinder sets are useful because set operations applied to any finite collection of open rectangles on E^∞ can be written as an open cylinder set. Let $B_1, \cdots, B_m$ be open rectangles on E^∞ such that $B_i = \prod_{j \in N_+} A_{ij}$ for any $1 \leq i \leq m$, where $A_{ij} \in \tau_j$. Note that there must now exist some $N \in N_+$ such that $A_{ij} = E_j$ for any $j > N$ and $1 \leq i \leq m$. Then, for any $1 \leq i \leq m$ we can write

$$B_i = \underbrace{(A_{i1} \times \cdots \times A_{iN})}_{A_i} \times E_{-1:N}$$

for each $1 \leq i \leq m$; since $A_i = A_{i1} \times \cdots \times A_{iN}$ is an open rectangle on $E_1 \times \cdots \times E_N$, it follows that $B_i \in \mathcal{C}_N$. In particular,

$$B_m \setminus \left(\sum_{i=1}^{m-1} B_i \right) = \left[A_m \setminus \left(\sum_{i=1}^{m-1} A_i \right) \right] \times E_{-1:N} \in \mathcal{C}_N.$$

Now we are ready to prove the result of interest.

Theorem 3.30 (Characterization of Weak Convergence on E^∞)
Let $\{(E_i, d_i)\}_{i \in N_+}$ be a collection of separable metric spaces and (E^∞, τ^∞) their product. Let d be the metric on E^∞ constructed above that metrizes the separable topological space (E^∞, τ^∞).

Let $\{\mu_n\}_{n \in N_+}$ be a sequence of probability measures on $(E^\infty, \mathcal{B}(E^\infty, \tau^\infty))$, and μ a probability measure on the same space. Suppose that, for any distinct $i_1, \cdots, i_k \in N_+$,

$$\mu_n \circ \pi^{-1}_{i_1, \cdots, i_k} \to \mu \circ \pi^{-1}_{i_1, \cdots, i_k}$$

weakly as $n \to \infty$. Then, $\mu_n \to \mu$ weakly as $n \to \infty$.

Proof) We will use the Portmanteau theorem to prove the result of interest. Specifically, we will show that, for any $A \in \tau^\infty$,

$$\liminf_{n \to \infty} \mu_n(A) \geq \mu(A).$$

First, we show that this inequality holds for open cylinder sets. For any $N \in N_+$ and $A \times E_{-1:N} \in \mathcal{C}_N$, we showed above that

$$\pi^{-1}_{1, \cdots, N}(A) = A \times E_{-1:N}.$$

As such,

$$\mu_n(A \times E_{-1:N}) = \left(\mu_n \circ \pi^{-1}_{1, \cdots, N} \right)(A)$$

for any $A \in \tau_1 \times \cdots \times \tau_N$, and likewise for μ. It follows from the weak convergence of the finite-dimensional distributions that

$$\liminf_{n \to \infty} \mu_n(A \times E_{-1:N}) = \liminf_{n \to \infty} \left(\mu_n \circ \pi^{-1}_{1, \cdots, N} \right)(A) \geq \left(\mu \circ \pi^{-1}_{1, \cdots, N} \right)(A) = \mu(A \times E_{-1:N}),$$

where we used the Portmanteau theorem to justify the inequality.

We now extend this result to arbitrary open sets in τ^∞. The separability of each (E_i, d_i) implies that they are second countable (by the property of metric spaces), and letting

$\mathbb{B}_i$ be the countable base generating τ_i, we know that the collection of open rectangles

$$\mathcal{M}_0 = \Big\{ \prod_{i \in N_+} A_i \mid A_i \in \mathbb{B}_i \cup \{E_i\}, \quad A_i \neq E_i \quad \text{for finitely many } i \Big\}$$

is a countable base on E^∞ that generates τ^∞.

Choosing any $A \in \tau^\infty$, this shows that there exists a countable collection $\{B_m\}_{m \in N_+}$ of open rectangles in $\mathcal{M}_0$ such that $A = \bigcup_m B_m$. Now we can define the sequence of open rectangles $\{C_m\}_{m \in N_+}$ as follows: let $C_1 = B_1$ and

$$C_m = B_m \setminus \left(\bigcup_{i=1}^{m-1} B_i \right)$$

for any $m \geq 2$. We showed above that each C_m is an open cylinder set, and this implies, in conjunction with the preceding result, that

$$\mu(C_m) \leq \liminf_{n \to \infty} \mu_n(C_m)$$

for any $m \in N_+$. Furthermore, because $\{C_m\}_{m \in N_+}$ is a sequence of sets that increases to A, by sequential continuity we have

$$\lim_{m \to \infty} \mu(C_m) = \mu(A).$$

Using the fact that $C_m \subset A$ for any $m \in N_+$, we can see that

$$\mu(C_m) \leq \liminf_{n \to \infty} \mu_n(C_m) \leq \liminf_{n \to \infty} \mu_n(A).$$

Finally, this holds for any $m \in N_+$, so taking $m \to \infty$ on both sides yields

$$\mu(A) \leq \liminf_{n \to \infty} \mu_n(A),$$

which is exactly the result we desired.

Q.E.D.

3.6.2 Measurability of Convex Maps

Our goal in this section is to study the behavior of extremum estimators when the objective function is convex-valued. Working with convex objective functions is convenient because of the following property of convex functions: for any open convex subset E of $\mathbb{R}^d$ that admits a countable dense subset $E_0 = \{x_i\}_{i \in N_+}$ and a convex function $f : E \to \mathbb{R}$,

$$f(x) = \sup\{g(x) \mid g \text{ is affine, } g(x_i) \leq f(x_i) \quad \forall i \in N_+\}$$

for any $x \in E$. In other words, the values of f on the countable dense subset E_0 determine the values of f on the entire domain E. This reduces problems involving the entire domain to problems involving only the countable set E_0, which is very useful since E is usually uncountably infinite.

For this property to be useful in the context of extremum estimation and probability theory in general, we need to show that $f(x)$ is not only a function but also a measurable function of the values $\{f(x_i)\}_{i \in N_+}$. Furthermore, we need to establish that the minimizers of f, if they exist, are also measurable functions of $\{f(x_i)\}_{i \in N_+}$. The results in this subsection are devoted to establishing these mathematical preliminaries.

First, we show a boundedness result that will prove useful when proving that $f(x)$ is a measurable function of the values $\{f(x_i)\}_{i \in N_+}$. The collection of affine functions g such that $g(x_i) \leq f(x_i)$ for a finite number of i can be expressed as a set of vectors in $\mathbb{R}^{d+1}$ that satisfy certain linear constraints, that is, as the constraint set of a linear program. The following lemma shows us that this set is bounded under some additional linear constraints.

Lemma 3.31 (Boundedness of Constraint Set of Linear Program)
Let $x_1, \cdots, x_{n_0}$ be a set in $\mathbb{R}^d$ whose convex hull contains the point $x \in \mathbb{R}^d$ in its interior. Define the set

$$H = \{(a, b) \in \mathbb{R}^{d+1} \mid a'x_i + b \leq y_i \quad \forall 1 \leq i \leq n_0, \quad m < a'x + b < M\}.$$

Then, H is a bounded set.

Proof) Suppose that H is not a bounded set. In this case, there exists a sequence $\{z_n = (a_n, b_n)\}_{n \in N_+} \subset H$ of non-zero vectors such that $|z_n| \to +\infty$ as $n \to \infty$. Define $\hat{z}_n = \frac{z_n}{|z_n|}$ for any $n \in N_+$. Then, $\{\hat{z}_n\}_{n \in N_+}$ is a sequence of d-dimensional vectors on the unit circle, which is bounded, so by the Bolzano-Weierstrass theorem there exists a subsequence $\{\hat{z}_{n_k}\}_{k \in N_+}$ of $\{\hat{z}_n\}_{n \in N_+}$ that converges to some unit vector $z_\infty = (a_\infty, b_\infty)$.

Note that, for any $k \in N_+$,

$$\frac{a'_{n_k} x_i + b_{n_k}}{|z_{n_k}|} \leq \frac{y_i}{|z_{n_k}|} \quad \forall 1 \leq i \leq n_0$$

and

$$\frac{m}{|z_{n_k}|} < \frac{a'_{n_k}x + b_{n_k}}{|z_{n_k}|} < \frac{M}{|z_{n_k}|}.$$

Taking $k \to \infty$ shows us that

$$a'_\infty x_i + b_\infty \leq 0 \quad \forall 1 \leq i \leq n_0$$

and $a'_\infty x + b_\infty = 0$.

This indicates that x lies on the $d-1$-dimensional hyperplane $V = \{z \in \mathbb{R}^d \mid a'_\infty z = -b_\infty\}$, while $x_1, \cdots, x_{n_0}$ lie completely on one side of V. If $a_\infty = O_{d\times 1}$, then $b_\infty = 0$, which contradicts the fact that $|z_\infty| = 1$. Thus, it must be the case that $a_\infty \neq O_{d\times 1}$, that is, $|a_\infty| > 0$.

Letting $C = \text{conv}(x_1, \cdots, x_{n_0})$ be the convex hull of C, for any $z \in C$ there exists $(\lambda_1, \cdots, \lambda_{n_0}) \in \Delta^{n_0-1}$ such that $z = \sum_{i=1}^{n_0} \lambda_i x_i$, so that

$$a'_\infty z + b_\infty = \sum_{i=1}^{n_0} \lambda_i (a'_\infty x_i + b_\infty) \leq 0.$$

It follows that C also lies entirely to one side of V.

By assumption, x lies in the interior of C, so there exists some $\delta > 0$ such that $B(x, \delta) \subset C$. Define $z^* = \eta \cdot a_\infty + x$, where $\eta = \frac{\delta}{|a_\infty|+1} > 0$. It follows that $z^* \in B(x, \delta) \subset C$, and that

$$a'_\infty z^* = \eta |a_\infty|^2 + a'_\infty x = \eta |a_\infty|^2 - b_\infty,$$

so that $a'_\infty z^* + b_\infty = \eta |a_\infty|^2 > 0$. This contradicts the fact that z^*, being a point in C, must satisfy $a'_\infty z^* + b_\infty \leq 0$. Therefore, it must be the case that H is bounded.

Q.E.D.

Next, we show that the subset F of $\mathbb{R}^\infty$ that collects points $y \in \mathbb{R}^\infty$ such that there exists a convex function $f : E \to \mathbb{R}$ such that $f(x_i) = y_i$ for any i is a closed and therefore measurable subset of $\mathbb{R}^\infty$. This is an important intermediary result because we will be interested in functions defined on F.

Lemma 3.32 Let E be an open convex subset of $\mathbb{R}^d$ equipped with the standard euclidean topology τ_E on E, and let $\mathbb{R}^\infty$ be the space of real valued sequences equipped with the product topology τ^∞, which is induced by the metric d defined earlier. Let $E_0 = \{x_i\}_{i \in N_+}$ be a countable dense subset of E. Then, the set F defined as

$$F = \{y \in \mathbb{R}^\infty \mid \exists \text{ convex function } f : E \to \mathbb{R} \text{ s.t. } f(x_i) = y_i \quad \forall i \in N_+\}$$

is a closed subset of $\mathbb{R}^\infty$, and for any $y \in F$, there exists a unique convex function $f : E \to \mathbb{R}$ such that $f(x_i) = y_i$ for any $i \in N_+$.

Proof) The first claim follows from the fact that pointwise convergence of a sequence of convex functions on a dense subset implies pointwise convergence everywhere on the domain. Formally, suppose $\{y^n\}_{n \in N_+}$ is a sequence contained in F that converges to $y \in \mathbb{R}^\infty$ in the metric d. This implies that, for any $i \in N_+$ and $\epsilon > 0$, there exists some $N \in N_+$ such that

$$\frac{\min(|y_i^n - y_i|, 1)}{i} \leq d(y^n, y) < \frac{\epsilon}{i}$$

for any $n \geq N$. It follows that $\min(|y_i^n - y_i|, 1) < \epsilon$ for any $n \geq N$, so that

$$\lim_{n \to \infty} \min(|y_i^n - y_i|, 1) = 0.$$

Since the topology induced by the standard bounded metric is equivalent to the euclidean topology on $\mathbb{R}$, it follows that $y_i^n \to y_i$ as $n \to \infty$ in the usual sense of the convergence of real numbers.

By the definition of F, for any $n \in N_+$ there exists a convex function $f_n : E \to \mathbb{R}$ such that $f_n(x_i) = y_i^n$ for any $i \in N_+$. In particular, this indicates that the sequence $\{f_n\}_{n \in N_+}$ of convex functions converges pointwise on the dense subset E_0 of E. This now implies that $\{f_n\}_{n \in N_+}$ converges pointwise on E to some convex function $f : E \to \mathbb{R}$: by the uniqueness of limits, it follows that $f(x_i) = y_i$ for any $i \in N_+$, and therefore $y \in F$. This proves that F is a closed subset of $\mathbb{R}^\infty$.

The uniqueness claim follows easily. For any $y \in F$, suppose that f, g are convex functions such that $f(x_i) = g(x_i) = y_i$ for any $i \in N_+$. By the characterization of convex functions via convex minorants, it follows that

$$f(x) = \sup\{g(x) \mid g \text{ is convex, } g(x_i) \leq y_i \quad \forall i \in N_+\} = g(x)$$

for any $x \in E$, so that $f = g$.

Q.E.D.

Now we are ready to show that the values of a convex function on the entire domain E is a measurable function of its values on a countable dense subset E_0 of E. The idea of proof is to approximate, in the sense of pointwise convergence, the desired function by a sequence of functions that involve the linear constraints $g(x_i) \leq f(x_i)$ for only a finite number of points x_i in E_0. This then reduces the problem of showing measurability of a function defined on $E \times \mathbb{R}^\infty$ to showing that the value function of a linear program is measurable, which is a much easier task.

Theorem 3.33 (Measurability of the Affine Approximation)
Let E be an open convex subset of $\mathbb{R}^d$ equipped with the standard euclidean topology τ_E on E, and let $\mathbb{R}^\infty$ be the space of real valued sequences equipped with the product topology τ^∞, which is induced by the metric d defined earlier. Let $E_0 = \{x_i\}_{i \in N_+}$ be a countable dense subset of E, and define the function $\Phi : E \times \mathbb{R}^\infty \to \mathbb{R}$ defined as

$$\Phi(x, y) = \sup\{g(x) \mid g \text{ is affine, } g(x_i) \leq y_i \quad \forall i \in N_+\}.$$

for any $(x, y) \in E \times \mathbb{R}^\infty$. Then,

i) For any $y \in \mathbb{R}^\infty$, $\Phi(\cdot, y)$ is a convex function, and

ii) For any $x \in E$, $\Phi(x, \cdot)$ is upper semicontinuous relative to the product topology τ^∞.

iii) $\Phi(\cdot)$ is jointly Borel measurable on $E \times \mathbb{R}^\infty$.

Proof) **An Approximation Scheme**

We first define the mapping $\Phi^n : E \times \mathbb{R}^\infty \to (-\infty, +\infty]$ as

$$\Phi^n(x, y) = \sup\{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall 1 \leq i \leq n\}$$

for any $(x, y) \in E \times \mathbb{R}^\infty$. We show that $\Phi(\cdot)$ is the infimum of the sequence of functions $\Phi^n(\cdot)$.

First, fix $(x, y) \in E \times \mathbb{R}^\infty$. Note that, for any $n \in N_+$,

$$\{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall i \in N_+\} \subset \{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall 1 \leq i \leq n\}.$$

It follows that

$$\begin{aligned} \Phi(x, y) &= \sup\{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall i \in N_+\} \\ &\leq \sup\{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall 1 \leq i \leq n\} = \Phi^n(x, y). \end{aligned}$$

This holds for any $n \in N_+$, so we can see that

$$\Phi(x,y) \leq \inf_{n \in N_+} \Phi^n(x,y).$$

It remains to show that the reverse inequality holds as well. First we show that there exists some $n_0 \in N_+$ such that $\Phi^n(x,y)$ is finite for any $n \geq n_0$. Since E_0 is a set that is dense in E, there exist vectors in E_0 such that the convex hull of these vectors contains x in its interior[2]. As such, there must exist some finite number $n_0 \in N_+$ such that $\{x_1, \cdots, x_{n_0}\}$ contains the vectors whose convex hull contains x.

Consider any affine function h such that $h(x_i) \leq y_i$ for any $1 \leq i \leq n_0$. Then, because affine functions are convex, and there exist some $\lambda_1, \cdots, \lambda_{n_0} \in [0,1]$ such that $x = \sum_{i=1}^{n_0} \lambda_i x_i$, it follows that

$$h(x) \leq \sum_{i=1}^{n_0} \lambda_i h(x_i) \leq \sum_{i=1}^{n_0} \lambda_i y_i \leq \max(y_1, \cdots, y_{n_0}).$$

Therefore,

$$\Phi^{n_0}(x,y) = \sup\{h(x) \mid h \text{ is affine, } h(x_i) \leq y_i \quad \forall 1 \leq i \leq n_0\} \leq \max(y_1, \cdots, y_{n_0}) < +\infty.$$

Furthermore, for any $n \geq n_0$, since $\Phi^n(x,y) \leq \Phi^{n_0}(x,y)$, $\Phi^n(x,y)$ is finite as well.

Now choose some $n \geq n_0$. By the definition of the supremum there exists an affine function h_n defined as $h_n(z) = a_n' z + b_n$ for any $z \in \mathbb{R}^d$ such that $h_n(x_i) \leq y_i$ for $1 \leq i \leq n$ and

$$\Phi(x,y) - 1 \leq \Phi^n(x,y) - \frac{1}{n} < h_n(x) \leq \Phi^n(x,y) \leq \Phi^{n_0}(x,y) < +\infty.$$

We can find a vector $(a_n, b_n) \in \mathbb{R}^{d+1}$ satisfying the above conditions for any $n \geq n_0$, so that we can construct a sequence $\{(a_n, b_n)\}_{n \geq n_0}$ such that

$$h_n(x_i) = a_n' x_i + b_n \leq y_i \quad \text{for any } 1 \leq i \leq n$$

for any $n \in N_+$ and which is also contained in the set

$$H = \{(a,b) \in \mathbb{R}^{d+1} \mid a' x_i + b \leq y_i \quad \forall 1 \leq i \leq n_0, \quad \Phi(x,y) - 1 < a' x + b < \Phi^{n_0}(x,y)\},$$

where $\Phi(x,y) - 1, \Phi^{n_0}(x,y)$ are finite constants and $x_1, \cdots, x_{n_0}$ are vectors such that x is contained in the interior of their convex hull. It follows from the preceding lemma that H is a bounded set. By the Bolzano-Weierstrass theorem, there exists some subsequence

[2]For a construction, see theorem 2.13 of the convex analysis and optimization text.

$\{(a_{n_k}, b_{n_k})\}_{k \in N_+}$ of $\{(a_n, b_n)\}_{n \geq n_0}$ such that

$$(a_{n_k}, b_{n_k}) \to (a, b)$$

as $k \to \infty$. Define $h : E \to \mathbb{R}$ as $h(x) = a'x + b$ for any $x \in \mathbb{R}^d$. Fix any $i \in N_+$. Since

$$a'_{n_k} x_i + b_{n_k} \leq y_i$$

for any $k \geq i$ by construction, taking $k \to \infty$ shows us that $h(x) = a'x_i + b \leq y_i$. This holds for any $i \in N_+$, so h is an affine function such that $h(x_i) \leq y_i$ for any $i \in N_+$, which tells us that $h(x) \leq \Phi(x, y)$. Finally, because

$$\inf_{n \in N_+} \Phi^n(x, y) - \frac{1}{n_k} \leq \Phi^{n_k}(x, y) - \frac{1}{n_k} < h_{n_k}(x) = a'_{n_k} x + b_{n_k},$$

taking $k \to \infty$ on both sides implies that

$$\inf_{n \in N_+} \Phi^n(x, y) \leq h(x) = a'x + b \leq \Phi(x, y).$$

This establishes that

$$\Phi(x, y) = \inf_{n \in N_+} \Phi^n(x, y) \in \mathbb{R}$$

for any $(x, y) \in E \times \mathbb{R}^\infty$.

We have just shown that $\Phi(x, y)$ can be represented as the infimum of a sequence of functions, and that it is real valued for any $(x, y) \in E \times \mathbb{R}^\infty$.

Convexity of $\Phi(\cdot, y)$

Fix $y \in \mathbb{R}^\infty$, and choose any $x^1, x^2 \in E$ and $\lambda \in [0, 1]$. Defining $\tilde{x} = \lambda x^1 + (1 - \lambda)x^2$, choose any affine function h such that $h(x_i) \leq y_i$ for any $i \in N_+$. Then, by affinity,

$$\lambda h(x^1) + (1 - \lambda) h(x^2) = h(\tilde{x}) \leq \Phi(\tilde{x}, y),$$

where the last inequality follows by definition. This holds for any affine function h such that $h(x_i) \leq y_i$ for any $i \in N_+$, so taking suprema over both sides with respect to h shows us that

$$\lambda \Phi(x^1) + (1 - \lambda) \Phi(x^2) \leq \Phi(\tilde{x}, y).$$

This proves that $\Phi(\cdot, y)$ is a real-valued convex function on E, and as such that it is continuous on E.

Upper Semicontinuity of $\Phi(x,\cdot)$

Now fix $x \in E$. We showed above that there exists some $n_0 \in N_+$ such that $\Phi^n(x,y) < +\infty$ for any $n \geq n_0$. For each $n \geq n_0$, $\Phi^n(x,y)$ is given as the value function to the following linear program:

$$\begin{aligned} \max_{(a,b)\in\mathbb{R}^{d+1}} \quad & a'x + b \\ \text{subject to} \quad & \underbrace{\begin{pmatrix} \tilde{X}_n & \iota_n \end{pmatrix}}_{A_n} \begin{pmatrix} a \\ b \end{pmatrix} \leq \tilde{Y}_n, \end{aligned}$$

where we define

$$\tilde{X}_n = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}, \quad \tilde{Y}_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

and $\iota_n \in \mathbb{R}^n$ is the vector of ones. It follows that, letting $\pi_{1,\cdots,n} : \mathbb{R}^\infty \to \mathbb{R}^n$ be the projection function defined in the previous section,

$$\Phi^n(x,y) = (v_x \circ \pi_{1,\cdots,n})(y)$$

for any $y \in \mathbb{R}^\infty$, where $v_x(\cdot)$ is the value function defined as

$$v_x(\hat{y}) = \sup\{c'\mathbf{x} \mid c \in \mathbb{R}^{d+1},\ A_n c \leq \hat{y}\},$$

for any $\hat{y} \in \mathbb{R}^n$, where $\mathbf{x} = (x', 1)'$. Note that $v_x(\cdot)$ is finite valued because we saw above that

$$v_x(\hat{y}) \leq \max(\hat{y}_1, \cdots, \hat{y}_n)$$

for any $\hat{y} \in \mathbb{R}^n$. Furthermore, $v_x(\cdot)$ is a concave function on $\mathbb{R}^n$: for any $\hat{y}_1, \hat{y}_2 \in \mathbb{R}^n$ and $\lambda \in [0,1]$, if $A_n c_1 \leq \hat{y}_1$ and $A_n c_2 \leq \hat{y}_2$, we have

$$\lambda(A_n c_1) + (1-\lambda)(A_n c_2) = A_n(\lambda c_1 + (1-\lambda)c_2) \leq \lambda\hat{y}_1 + (1-\lambda)\hat{y}_2.$$

It follows that

$$A_n(\lambda c_1 + (1-\lambda)c_2) \leq v_x(\lambda\hat{y}_1 + (1-\lambda)\hat{y}_2),$$

and since this holds for any $c_1, c_2 \in \mathbb{R}^{d+1}$ such that $A_n c_1 \leq \hat{y}_1$ and $A_n c_2 \leq \hat{y}_2$, we have

$$\lambda v_x(\hat{y}_1) + (1-\lambda)v_x(\hat{y}_2) \leq v_x(\lambda\hat{y}_1 + (1-\lambda)\hat{y}_2).$$

Therefore, $v_x(\cdot)$ is a concave and therefore continuous function on $\mathbb{R}^n$. Because conti-

nuity is preserved across compositions and the projection function is continuous, the mapping $\Phi^n(x,\cdot)$ is also continuous relative to the product topology τ^∞. This result holds for any $n \geq n_0$.

Finally, recall that $\{\Phi^n(x,y)\}_{n\in N_+}$ is a decreasing sequence in $(-\infty,+\infty]$, with elements corresponding to $n \geq n_0$ being finite valued. Therefore, for any $y \in \mathbb{R}^\infty$, we can write

$$\Phi(x,y) = \inf_{n\geq n_0} \Phi^n(x,y),$$

so that $\Phi(x,\cdot)$ is the infimum of continuous functions. Since, for any $\alpha \in \mathbb{R}$,

$$\{y \in \mathbb{R}^\infty \mid \Phi(x,y) < \alpha\} = \bigcup_{n\geq n_0} \{y \in \mathbb{R}^\infty \mid \Phi^n(x,y) < \alpha\},$$

where each $\{y \in \mathbb{R}^\infty \mid \Phi^n(x,y) < \alpha\}$ is an open set by the continuity of $\Phi^n(x,\cdot)$, it follows that $\{y \in \mathbb{R}^\infty \mid \Phi(x,y) < \alpha\}$ is also an open set contained in τ^∞. By definition, $\Phi(x,\cdot)$ is an upper semicontinuous function.

Joint Borel Measurability

We have shown that $\Phi : E \times \mathbb{R}^\infty \to \mathbb{R}$ is a function such that $\Phi(x,\cdot)$ is upper semicontinuous and thus Borel measurable function on $\mathbb{R}^\infty$ for any $x \in E$, and $\Phi(\cdot,y)$ is convex on E and thus continuous for any $y \in \mathbb{R}^\infty$. We now show that this implies that $\Phi(\cdot)$ is Borel measurable with respect to the product Borel σ-algebra $\mathcal{B}(E,\tau_E) \bigotimes \mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$.

Since E is a subset of the euclidean d-space, for any $n \in N_+$ we can construct a countable partition $\{B_k^n\}_{k\in N_+}$ of E and a corresponding sequence of points $\{x_k^n\}_{k\in N_+}$ such that $x_k^n \in B_k^n$ for any $k \in N_+$ such that

$$\operatorname{diam}(B_k^n) = \sup\{\left|x^1 - x^2\right| \mid x^1, x^2 \in B_k^n\} \leq \frac{1}{n}$$

for any $k \in N_+$[3]. Now we can define the function $\phi_n : E \times \mathbb{R}^\infty \to \mathbb{R}$ as

$$\phi_n(x,y) = \sum_{k=1}^{\infty} I_{B_k^n}(x)\Phi(x_k^n, y)$$

for any $(x,y) \in E \times \mathbb{R}^\infty$. Since each $\Phi(x_k^n,\cdot)$ is a $\mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$-measurable function and $I_{B_k^n}$ is $\mathcal{B}(E,\tau_E)$-measurable, the mapping

$$(x,y) \mapsto I_{B_k^n}(x)\Phi(x_k^n, y)$$

[3]One simple construction in the case $E = \mathbb{R}^d$ is to consider the collection of all half-open d-cells with edge length $\frac{1}{n\sqrt{d}}$, and let x_k^n be the (coordinate-wise) smallest vertex of each d-cell.

is $\mathcal{B}(E,\tau_E)\bigotimes\mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$-measurable. Furthermore, since $\phi_n(\cdot)$ is the pointwise limit of measurable functions, it follows that it is also $\mathcal{B}(E,\tau_E)\bigotimes\mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$-measurable.

It remains to show that $\phi_n(\cdot)$ converges pointwise to $\Phi(\cdot)$. To this end, fix $(x,y)\in E\times\mathbb{R}^\infty$. For any $\epsilon>0$, there exists some $\delta>0$ such that $|\Phi(x,y)-\Phi(x',y)|<\epsilon$ for any $x'\in E$ such that $|x-x'|<\delta$. Choose $N\in N_+$ so that $\frac{1}{N}<\delta$. For any $n\geq N$, there exists some $k\in N_+$ such that $x\in B_k^n$; the diameter of B_k^n is smaller than or equal to $\frac{1}{n}$, and $x_k^n\in B_k^n$, so it follows that

$$|x-x_k^n|\leq\frac{1}{n}<\delta,$$

which implies that

$$|\Phi(x,y)-\phi_n(x,y)|=|\Phi(x,y)-\Phi(x_k^n,y)|<\epsilon.$$

As such, for any $n\geq N$,

$$|\Phi(x,y)-\phi_n(x,y)|<\epsilon,$$

and since we can find such an $N\in N_+$ for any $\epsilon>0$, it follows that

$$\Phi(x,y)=\lim_{n\to\infty}\phi_n(x,y).$$

Therefore, $\Phi(\cdot)$, being the pointwise limit of $\mathcal{B}(E,\tau_E)\bigotimes\mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$-measurable functions, is itself $\mathcal{B}(E,\tau_E)\bigotimes\mathcal{B}(\mathbb{R}^\infty,\tau^\infty)$-measurable.

Q.E.D.

Our ultimate goal is to develop a theory for extremum estimation for convex objective functions. As usual, the first course of action is to verify when the extremum exists and if it can be formulated as some sort of measurable function. Continuing the theme of characterizing convex functions via their values on a countable dense subset, we want to find sufficient conditions such that the minimizer correspondence is non-empty valued and there exists a measurable selection of this correspondence that is a function from $\mathbb{R}^\infty$ into the domain E of the objective function.

To this end, we first define the value function $v:\mathbb{R}^\infty\to[-\infty,+\infty]$ as

$$v(y)=\inf_{x\in E}\Phi(x,y)$$

for any $y\in\mathbb{R}^\infty$. If $y\in F$, $v(y)$ is the minimum over E of the convex function corresponding to y. Since $\Phi(x,\cdot)$ is upper semicontinuous for any $x\in E$, it follows that, for any $\alpha\in\mathbb{R}$,

$$\{y\in\mathbb{R}^\infty\mid v(y)\geq\alpha\}=\bigcap_{x\in E}\{y\in\mathbb{R}^\infty\mid\Phi(x,y)\geq\alpha\},$$

where the right hand side is a closed subset of $\mathbb{R}^\infty$ because each $\{y \in \mathbb{R}^\infty \mid \Phi(x,y) \geq \alpha\}$ is closed by upper semicontinuity and closed sets are closed under arbitrary intersections. Thus,

$$\{y \in \mathbb{R}^\infty \mid v(y) < \alpha\}$$

is an open subset of $\mathbb{R}^\infty$ and $v(\cdot)$ is an upper semicontinuous function. It also follows that $v(\cdot)$ is measurable relative to the Borel σ-algebra on $\mathbb{R}^\infty$ and the standard Borel σ-algebra on the extended real line $[-\infty, +\infty]$.

Since the above argument does not depend on the fact that we take the infimum over the set E, it follows that, for any non-empty set $A \subset E$ (which is not necessarily measurable), the function $v_A : \mathbb{R}^\infty \to [-\infty, +\infty]$ defined as

$$v_A(y) = \inf_{x \in A} \Phi(x,y)$$

for any $y \in \mathbb{R}^\infty$ is upper semicontinuous, that is,

$$\{y \in \mathbb{R}^\infty \mid v_A(y) < \alpha\} \in \tau^\infty$$

for any $\alpha \in \mathbb{R}$.

Now define the minimizer correspondence $\Gamma : \mathbb{R}^\infty \to E$ as

$$\Gamma(y) = \{x \in E \mid \Phi(x,y) = v(y)\}$$

for any $y \in \mathbb{R}^\infty$. For any $y \in F$, $\Gamma(y)$ collects the set of points $x \in E$ that attain the infimum of the convex function corresponding to y. Note that $\Gamma(\cdot)$ may be empty valued; think of the exponential function defined on the real line $\mathbb{R}$.

The collection of all sequences $y \in \mathbb{R}^\infty$ that corresponds to a convex function for which a minimizer exists can now be defined using the correspondence $\Gamma(\cdot)$ as

$$F_0 = \{y \in \mathbb{R}^\infty \mid \Gamma(y) \neq \emptyset\}.$$

The following result shows that F_0 is contained in some Borel measurable subset of $\mathbb{R}^\infty$.

Theorem 3.34 (Measurability of Sequences with Non-empty Minimzer Correspondence)
Let E be a non-empty open convex subset of $\mathbb{R}^d$ equipped with the standard euclidean topology, and let $\mathbb{R}^\infty$ be the set of all real valued sequences with the product topology τ^∞, which is induced by the metric d defined in the previous section. Let $\Phi : E \times \mathbb{R}^\infty \to \mathbb{R}$ be the function defined earlier, and F the subset of $\mathbb{R}^\infty$ that collects all sequences that correspond to some convex function on E. Finally, for any non-empty subset A of E, let $v_A : \mathbb{R}^\infty \to [-\infty, +\infty]$ be the value function and $\Gamma(\cdot)$ the minimizer correspondence defined above.

Let F_0 be the set defined as

$$F_0 = \{y \in \mathbb{R}^\infty \mid \Gamma(y) \neq \emptyset\} \cap F.$$

Then, $F_0 \in \mathcal{B}(\mathbb{R}^\infty, \tau^\infty)$.

Proof) First define $\mathbb{B}$ as the collection of all closed balls centered on points in E_0 with rational radii that are fully contained in E:

$$\mathbb{B} = \{\bar{B}(x,q) \mid q \in \mathbb{Q}, \quad x \in E_0, \quad \bar{B}(x,q) \subset E\}.$$

Now define $\mathbb{K}$ as the collection of all finite unions of sets in $\mathbb{B}$:

$$\mathbb{K} = \Big\{ \bigcup_{i=1}^{k} B_i \mid B_1, \cdots, B_k \in \mathbb{Q}, \quad k \in N_+ \} \cup \{\emptyset\}.$$

Note that $\mathbb{B}$ is a countable collection of compact subsets of E due to the fact that E_0 and $\mathbb{Q}$ are both countable, which makes $\mathbb{K}$ a countable collection of compact subsets of E as well (finite unions of compact sets is compact). Furthermore, we can show that any compact subset K of E is actually contained in a set in $\mathbb{K}$. To see this, choose any $x \in K$; since $x \in E$, the openness of E and the denseness of E tells us that there exists some point $x_0(x) \in E_0$ and $q(x) \in \mathbb{Q}$ such that $x \in B(x_0(x), q(x)) \subset \bar{B}(x_0(x), q(x)) \subset E$. Thus, the collection $\{B(x_0(x), q(x))\}_{x \in K}$ forms an open cover of K. By compactness, there now exists a finite number of points $x^1, \cdots, x^k \in K$ such that

$$K \subset \bigcup_{i=1}^{k} B(x_0(x^i), q(x^i)) \subset \bigcup_{i=1}^{k} \bar{B}(x_0(x^i), q(x^i)).$$

However, each $\bar{B}(x_0(x^i), q(x^i))$ is a set in $\mathbb{B}$, so that we have our desired result.

Define

$$\tilde{F}_0 = \{y \in \mathbb{R}^\infty \mid \Gamma(y) \neq \emptyset\}.$$

We can establish the following equality:

$$\tilde{F}_0^c = \bigcap_{B \in \mathbb{B}} \{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\}.$$

Suppose that $y \in \tilde{F}_0^c$, so that $\Gamma(y) = \emptyset$. Then, $\Phi(x, y) > v(y)$ for any $x \in E$. For any $B \in \mathbb{B}$, since B is a compact set, it follows from the extreme value theorem that there exists some $x \in B$ such that

$$v_B(y) = \Phi(x, y) > v(y).$$

This holds for any $B \in \mathbb{B}$, so we have shown that

$$\tilde{F}_0^c \subset \bigcap_{B \in \mathbb{B}} \{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\}.$$

To show the reverse inclusion, choose any $y \in \mathbb{R}^\infty$ such that $v_B(y) > v(y)$ for any $B \in \mathbb{B}$, and suppose for the sake of contradiction that $\Gamma(y) \neq \emptyset$. Then, there exists some $x^* \in E$ such that

$$v(y) = \Phi(x^*, y).$$

Since $\{x^*\}$ is a compact subset of the open set E, and any metric space is a locally compact Hausdorff space, there exists an open subset V of $\mathbb{R}^d$ with compact closure such that $x^* \subset V \subset \overline{V} \subset E$. Since $\overline{V}$ is compact, it follows that there exists some $B \in \mathbb{B}$ such that $\overline{V} \subset B$, so that $x^* \in B$. Since x^* minimizes $\Phi(\cdot, y)$ on E, it also minimizes it on B, so that

$$v_B(y) = \Phi(x^*, y) = v(y).$$

This contradicts our initial assumption, so it must be the case that $\Gamma(y) = \emptyset$. We have thus established that

$$\bigcap_{B \in \mathbb{B}} \{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\} \subset \tilde{F}_0^c$$

as well.

Now fix any $B \in \mathbb{B}$, and note that

$$\{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\} = \bigcup_{q,r \in \mathbb{Q}, q<r} \{y \in \mathbb{R}^\infty \mid v_B(y) \geq r\} \cap \{y \in \mathbb{R}^\infty \mid q > v(y)\}$$

For each $r, q \in \mathbb{Q}$ such that $r > q$, the upper semicontinuity of $v_B(\cdot)$ and $v(\cdot)$ implies that the intersection $\{y \in \mathbb{R}^\infty \mid v_B(y) \geq r\} \cap \{y \in \mathbb{R}^\infty \mid q > v(y)\}$ is the intersection of a closed and open set and consequently Borel measurable. Since σ-algebras are closed

under countable unions, this allows us to conclude that

$$\{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\} \in \mathcal{B}(\mathbb{R}^\infty, \tau^\infty).$$

Because $\mathbb{B}$ is a countable collection of sets and σ-algebras are closed under countable intersections,

$$\tilde{F}_0^c = \bigcap_{B \in \mathbb{B}} \{y \in \mathbb{R}^\infty \mid v_B(y) > v(y)\} \in \mathcal{B}(\mathbb{R}^\infty, \tau^\infty).$$

Finally, because σ-algebras are closed under complementation, $F_0 = \tilde{F}_0 \cap F \in \mathcal{B}(\mathbb{R}^\infty, \tau^\infty)$.

Q.E.D.

Now that we have established the measurability of the set F_0 of real sequences corresponding to convex functions that attain their minimum, we can move onto proving the existence of a measurable selection of the minimizer correspondence on F_0. The next result uses the measurable selection theorem to establish the desired result.

Theorem 3.35 (Existence of Measurable Selection for Minimizer Correspondence) Let E be an open convex subset of $\mathbb{R}^d$ with countable dense subset $E_0 = \{x_i\}_{i \in N_+}$, and let $\mathbb{R}^\infty$ be the space of real valued sequences equipped with the product topology τ^∞, which is induced by the metric d defined in the previous section. Let

$$F_0 = \{y \in F \mid \Gamma(y) \neq \emptyset\}$$

be the Borel measurable set defined above. Let $\mathcal{F}$ be the trace of $\mathcal{B}(\mathcal{R}^\infty, \tau^\infty)$ on F_0, which is exactly the Borel σ-algebra generated by the subspace topology on F_0 induced by τ^∞, and similarly define $\mathcal{E}$ as the natural Borel σ-algebra on E.

Then, the minimizer correspondence restricted to F_0 admits a measurable selection, that is, there exists a function $g_* : F_0 \to E$ measurable relative to $\mathcal{F}$ and $\mathcal{E}$ such that

$$g_*(y) \in \operatorname*{argmin}_{x \in E} \Phi(x, y)$$

for any $y \in F_0$.

Proof) For notational simplicity, we let $\Gamma(\cdot)$ continue to denote the restriction of the minimizer correspondence to F_0.

Note that $\Gamma(\cdot)$ is a correspondence from a measurable space $(F_0, \mathcal{F})$ into the metric space E, which is complete and separable under the euclidean metric, and $\mathcal{E}$ is the Borel σ-algebra on E induced by the euclidean metric. Furthermore, $\Gamma(\cdot)$ is non-empty valued. We can also show that $\Gamma(\cdot)$ is closed valued relative to E: choose any $y \in F_0$ and

sequence $\{x_n\}_{n\in N_+}$ contained in $\Gamma(y)$ converging to some $x \in E$. Because $y \in F_0 \subset F$, there exists a unique convex function $f : E \to \mathbb{R}$ such that $f(x) = \Phi(x, y)$ for any $x \in E$. By definition,

$$f(x_n) = \Phi(x_n, y) = v(y)$$

for any $n \in N_+$, and since convex functions are continuous, taking $n \to \infty$ on both sides shows us that $f(x) = \Phi(x, y) = v(y)$, so that $x \in \Gamma(y)$ as well. This shows us that $\Gamma(y)$ is a closed set relative to E.

Therefore, to apply the Kuratkowski-Nyll-Nardzewski measurable selection theorem, we need only show that, for any subset A of E that is closed relative to E,

$$\{y \in F_0 \mid \Gamma(y) \cap A \neq \emptyset\}$$

is $\mathcal{F}$-measurable. We will actually show the stronger result that this set is measurable for any subset of E (it does not even need to be measurable).

The set above is empty if $A = \emptyset$, which is trivially measurable. Now suppose that A is a non-empty subset of E. We will first show that $v_A(\cdot)$ is real valued. To this end, choose any $y \in F_0$ and $x_0 \in A$. Putting $\alpha = \Phi(x_0, y)$, note that

$$v(y) = \inf_{x\in E} \Phi(x, y) \leq v_A(y) = \inf_{x\in A} \Phi(x, y) \leq \Phi(x_0, y) = \alpha.$$

Because $\Gamma(y) \neq \emptyset$, $v(y)$ is a real number (it is equal to the function $\Phi(\cdot, y)$ evaluated at some point in E) and therefore $v_A(y) \in \mathbb{R}$ as well.
The upper semicontinuity of $v(\cdot)$ and $v_A(\cdot)$ now implies that both are measurable relative to $\mathcal{F}$ and $\mathcal{B}(\mathbb{R})$. Since

$$\{y \in F_0 \mid \Gamma(y) \cap A \neq \emptyset\} = \{y \in F_0 \mid v(y) = v_A(y)\},$$

and because the right hand side is a set of values for which two measurable real valued functions are equal, the set on the left hand side is $\mathcal{F}$-measurable. Now the desired result follows from the measurable selection theorem.

Q.E.D.

3.6.3 The Basic Convergence Theorem for Convex Extremum Estimation

We are now ready to prove the core result for extremum estimation problems involving convex objective functions. It states that a sufficient condition for weak convergence of a minimum estimator is for the finite-dimensional distributions of the objective function to converge weakly. This is a significant simplification in the sense that it allows us to replace the uniform convergence in probability results required for the basic consistency and asymptotic normality theorems with pointwise convergence. Since it was precisely the uniform convergence in probability result that required the compactness of the parameter space, this theorem allows us to jettison the compactness assumption as well. Thus, by imposing the additional condition of convexity on the objective function, we are able to derive asymptotic results for the corresponding extremum estimators under much weaker assumptions elsewhere.

We work with the following setting. Let E be an open convex subset of $\mathbb{R}^d$ with countable dense subset $E_0 = \{\theta_i\}_{i \in N_+}$, and let $\mathbb{R}^\infty$ be the space of real valued sequences equipped with the product topology τ^∞, which is induced by the metric d defined earlier. We define F_0 as the measurable subset of $\mathbb{R}^\infty$ that collects sequences corresponding to convex functions that attain their minimum on E.

Let $\mathcal{F}$ be the trace of $\mathcal{B}(\mathcal{R}^\infty, \tau^\infty)$ on F_0, which is exactly the Borel σ-algebra generated by the subspace topology on F_0 induced by τ^∞, and similarly define $\mathcal{E}$ as the natural Borel σ-algebra on E. Finally, let $g_* : F_0 \to E$ be the function, found above, that is measurable relative to $\mathcal{F}$ and $\mathcal{E}$ such that

$$g_*(y) \in \operatorname{argmin}_{\theta \in E} \Phi(\theta, y)$$

for any $y \in \mathbb{R}^\infty$.

For the sample size T, the objective function is given as a real valued stochastic process $\{Q_T(\theta)\}_{\theta \in E}$ with paths forming convex functions that attain their minimum on E: for any $\omega \in \Omega$ the function $f : E \to \mathbb{R}$ defined as

$$f(\theta) = [Q_T(\theta)](\omega)$$

for any $\theta \in E$ is a convex function attaining its minimum on E. Because each $Q_T(\theta)$ is real-valued a random variable, it follows that the random sequence $Q_T : \Omega \to \mathbb{R}^\infty$ defined as

$$Q_T(\omega) = \{[Q_T(\theta)](\omega)\}_{\theta \in E_0}$$

for any $\omega \in \Omega$ is measurable relative to $\mathcal{H}$ and the product σ-algebra $\mathcal{B}(\mathbb{R})^{E_0}$; since $(\mathbb{R}, \tau_\mathbb{R})$ is a second countable topological space, E_0 a countable index set, and τ^∞ the product topology on $\mathbb{R}^\infty$, it follows that

$$\mathcal{B}(\mathbb{R})^{E_0} = \mathcal{B}(\mathbb{R}^\infty, \tau^\infty).$$

Since Q_T maps into F_0, we can now conclude that Q_T is measurable relative to $\mathcal{H}$ and $\mathcal{F}$, and

define the extremum estimator $\hat{\theta}_T$ corresponding to the above objective function as

$$\hat{\theta}_T = g_*(Q_T);$$

by implication,

$$\hat{\theta}_T(\omega) \in \text{argmin}_{\theta \in E}[Q_T(\theta)](\omega)$$

for any $\omega \in \Omega$. It follows from the measurability of $g_*(\cdot)$ and Q_T that $\hat{\theta}_T$ is a random variable taking values in E.

We want to establish the conditions under which weak convergence of the finite-dimensional distributions of $\{Q_T(\theta)\}_{\theta \in E}$ to some real valued stochastic process $\{Q_0(\theta)\}_{\theta \in E}$ with paths in F_0 implies the weak convergence of the corresponding minimizers. It turns out that we must impose an identification condition on the limiting objective function $Q_0(\theta)$, similar to the case of arbitrary continuous objective functions and compact parameter space. The formal condition is given as follows:

Identification Condition There exists a Borel measurable subset F^* of F_0 such that $\{Q_0(\theta)\}_{\theta \in E_0}$ almost surely takes values in F^* and, for any $y \in F^*$, the corresponding convex function has a unique minimizer.

The latter part of the identification condition is familiar; it simply requires that the weak limit of the sequence of minimizers is well-defined as the unique minimizer of the limiting objective function. The first part, which requires F^* to be Borel measurable, turns out to be necessary because we want to preserve the uniqueness of the minimizer as a property possessed by the stochastic process $\{Q_0(\theta)\}_{\theta \in E_0}$, so that a stochastic process with the same distribution as $\{Q_0(\theta)\}_{\theta \in E_0}$ also has paths that correspond to convex functions with unique minimizers.

The main result, as well as its proof, is stated below. It is a more detailed version of lemma A.1 in Knight (1989).

Theorem 3.36 (The Basic Convergence Theorem for Convex Objectives) For any $T \in \mathbb{N}$, let $\{Q_T(\theta)\}_{\theta \in E}$ be a stochastic process with paths taking values in F_0, and let $\{\hat{\theta}_T\}_{T \in \mathbb{N}}$ be the corresponding extremum estimators. Suppose that there exists some measurable subset F^* of F_0 such that $\{Q_0(\theta)\}_{x \in E_0}$ almost surely takes values in F^* and, for any $y \in F^*$, the corresponding convex function has a unique minimizer.

If, for any $N \in N_+$,

$$(Q_T(\theta_1), \cdots, Q_T(\theta_N)) \xrightarrow{d} (Q_0(\theta_1), \cdots, Q_0(\theta_N))$$

as $T \to \infty$, then

$$\hat{\theta}_T \xrightarrow{d} \hat{\theta}_0$$

as $T \to \infty$.

Proof) For any $T \in \mathbb{N}$, define the random sequence Q_T taking values in F_0 as

$$Q_T(\omega) = \{[Q_T(\theta_i)](\omega)\}_{i \in N_+}$$

for any $\omega \in \Omega$, and let μ_T be its distribution. By assumption,

$$\mu_T \circ \pi^{-1}_{\theta_1, \cdots, \theta_N} \to \mu_0 \circ \pi^{-1}_{\theta_1, \cdots, \theta_N}$$

weakly for any $N \in N_+$, that is, the finite dimensional distributions of μ_T converge weakly. As such, by the characterization of weak convergence on $\mathbb{R}^\infty$,

$$\mu_T \to \mu_0$$

weakly, or equivalently,

$$Q_T \xrightarrow{d} Q_0.$$

If $g_*(\cdot)$ were a continuous function, then the desired result would follow immediately from the continuous mapping theorem. However, in general $g_*(\cdot)$ is not continuous, only measurable, so that we need to take a different approach.

Note that (F_0, d) is a separable metric space. As such, by Skorokhod's representation theorem, there exists a probability space $(\Omega', \mathcal{H}', \mathbb{P}')$ and random variables $\{Q^*_T\}_{T \in \mathbb{N}}$ defined on that space measurable relative to $\mathcal{H}'$ and $\mathcal{F}$ such that

$$Q^*_T \sim Q_T \quad \text{for any } T \in \mathbb{N}, \quad \text{and}$$

$$Q^*_T \xrightarrow{a.s.} Q^*_0 \quad \text{as } T \to \infty.$$

Defining $\hat{\theta}^*_T = g_*(Q^*_T)$ on Ω' for any $T \in \mathbb{N}$, it follows that

$$\hat{\theta}^*_T \sim \hat{\theta}_T \quad \text{for any } T \in \mathbb{N}$$

as well. Let $\Omega'_1 \subset \Omega'$ be the $\mathbb{P}'$-almost sure set on which the almost sure convergence of $\{Q^*_T\}_{T \in N_+}$ occurs. Similarly, define $\Omega'_2 = (Q^*_0)^{-1}(F_0)$. Since $\mu_0(F^*) = 1$ by assumption, it follows that

$$\mathbb{P}'\left[\Omega'_2\right] = \mathbb{P}'\left[(Q^*_0)^{-1}(F_0)\right] = \mu_0(F^*) = 1,$$

so that $\Omega'_0 = \Omega'_1 \cap \Omega'_2$ is an almost sure set under $\mathbb{P}'$.

Now choose any $\omega \in \Omega'_0$. We want to show that $\hat{\theta}^*_T(\omega) \to \hat{\theta}^*_0(\omega)$. First, for any $T \in \mathbb{N}$,

define $f_T : E \to \mathbb{R}$ as

$$f_T(\theta) = \Phi(\theta, Q_T^*(\omega))$$

for any $\theta \in E$. By design, $f_T(\cdot)$ is a convex function on E such that

$$f_T(\hat{\theta}_T^*(\omega)) = \min_{\theta \in E} f_T(\theta).$$

Furthermore, because $Q_0^*(\omega) \in F^*$, $\hat{\theta}_0^*(\omega)$ is the unique minimizer of $f_0(\cdot)$.

Since each $Q_T^*(\omega)$ takes values in $\mathbb{R}^\infty$, it can be represented as a sequence $\{y_{T,i}\}_{i \in N_+}$ of real numbers; for any $i \in N_+$, we now have

$$f_T(\theta_i) = y_{T,i} \quad \text{for any } i \in N_+.$$

Since

$$\lim_{T \to \infty} d(Q_T^*(\omega), Q_0^*(\omega)) = 0$$

and convergence in d implies elementwise convergence,

$$\lim_{T \to \infty} y_{T,i} = y_{0,i}$$

for any $i \in N_+$, so that $f_T(\cdot)$ converges pointwise on E_0 to $f_0(\cdot)$. Since $\{f_T\}_{T \in \mathbb{N}}$ is a sequence of convex functions, it follows that $f_T(\cdot)$ converges pointwise on E to $f_0(\cdot)$, and that this convergence is uniform on any compact subset K of E.

By the openness of E, there exists some $\eta > 0$ such that $B(\hat{\theta}_0^*(\omega), \eta) \subset E$. For any $\epsilon \in (0, \eta)$, define the set

$$K_\epsilon = \{\theta \in E \mid \left|\theta - \hat{\theta}_0^*(\omega)\right| = \epsilon\}.$$

Then, K_ϵ is a non-empty set contained in E, and K_ϵ is closed and bounded, so it is a compact subset of E in light of the Heine-Borel theorem. Define

$$\delta = \inf_{\theta \in K_\epsilon} (f_0(\theta) - f_0(\hat{\theta}_0^*(\omega)));$$

because K_ϵ is a compact set and $f_0(\cdot) - f_0(\hat{\theta}_0^*(\omega))$ is continuous on E, there exists some $\theta_\epsilon^* \in K_\epsilon$ such that

$$\delta = f_0(\theta_\epsilon^*) - f_0(\hat{\theta}_0^*(\omega)).$$

Since $\hat{\theta}_0^*(\omega)$ is the unique minimizer of $f_0(\cdot)$ and $\hat{\theta}_0^*(\omega) \notin K_\epsilon$, so that $\theta_\epsilon^* \neq \hat{\theta}_0^*(\omega)$, it

follows that $\delta > 0$.

The uniform convergence of $\{f_T\}_{T\in N_+}$ to f_0 on compact sets implies that

$$\sup_{\theta\in K_\epsilon} |f_T(\theta) - f_0(\theta)| \to 0,$$

and the pointwise convergence of $\{f_T\}_{T\in N_+}$ to f_0 implies that

$$\lim_{T\to\infty} f_T(\hat{\theta}_0^*(\omega)) = f_0(\hat{\theta}_0^*(\omega)).$$

It follows that there exists some $N \in N_+$ such that

$$\sup_{\theta\in K_\epsilon} |f_T(\theta) - f_0(\theta)| < \frac{\delta}{4} \quad \text{and} \quad \left|f_T(\hat{\theta}_0^*(\omega)) - f_0(\hat{\theta}_0^*(\omega))\right| < \frac{\delta}{4}$$

for any $T \geq N$. In particular, for any $T \geq N$ and $\theta \in K_\epsilon$,

$$f_T(\theta) - f_0(\theta) > -\frac{\delta}{4}$$

$$f_0(\theta) \geq f_0(\hat{\theta}_0^*(\omega))) + \delta$$

$$f_0(\hat{\theta}_0^*(\omega))) > f_T(\hat{\theta}_0^*(\omega))) - \frac{\delta}{4},$$

where the second inequality follows from the definition of $\delta > 0$. Putting these inequalities together, we can see that, for any $T \geq N$ and $\theta \in K_\epsilon$,

$$f_T(\theta) > f_0(\theta) - \frac{\delta}{4}$$

$$\geq \frac{3}{4}\delta + f_0(\hat{\theta}_0^*(\omega))$$

$$> \frac{\delta}{2} + f_T(\hat{\theta}_0^*(\omega)).$$

This holds for any $\theta \in K_\epsilon$, so we have

$$\inf_{\theta\in K_\epsilon} (f_T(\theta) - f_T(\hat{\theta}_0^*(\omega))) \geq \frac{\delta}{2} > 0$$

for any $T \geq N$.

Suppose $\left|\hat{\theta}_T^*(\omega) - \hat{\theta}_0^*(\omega)\right| > \epsilon$ for some $T \geq N$. Then, defining

$$\lambda = \frac{\epsilon}{\left|\hat{\theta}_T^*(\omega) - \hat{\theta}_0^*(\omega)\right|} \in (0,1)$$

and

$$\theta_\lambda = \lambda\hat{\theta}_T^*(\omega) + (1-\lambda)\hat{\theta}_0^*(\omega),$$

$\theta_\lambda \in E$ by convexity and

$$\begin{aligned}\left|\theta_\lambda - \hat{\theta}_0^*(\omega)\right| &= \left|\lambda\hat{\theta}_T^*(\omega) - \lambda\hat{\theta}_0^*(\omega)\right| \\ &= \lambda\left|\hat{\theta}_T^*(\omega) - \hat{\theta}_0^*(\omega)\right| = \epsilon,\end{aligned}$$

so that $\theta_\lambda \in K_\epsilon$. Furthermore, by the convexity of $f_T(\cdot)$,

$$f_T(\theta_\lambda) \leq \lambda f_T(\hat{\theta}_T^*(\omega)) + (1-\lambda) f_T(\hat{\theta}_0^*(\omega)) \leq f_T(\hat{\theta}_0^*(\omega)),$$

where we used the fact that, as a minimizer of $f_T(\cdot)$, $\hat{\theta}_T^*(\omega)$ satisfies

$$f_T(\hat{\theta}_T^*(\omega)) \leq f_T(\hat{\theta}_0^*(\omega)).$$

By implication,

$$f_T(\theta_\lambda) - f_T(\hat{\theta}_0^*(\omega)) \leq 0,$$

but because $\theta_\lambda \in K_\epsilon$, we must have

$$f_T(\theta_\lambda) - f_T(\hat{\theta}_0^*(\omega)) \geq \frac{\delta}{2} > 0$$

as well. This is a contradiction, so it must be the case that

$$\left|\hat{\theta}_T^*(\omega) - \hat{\theta}_0^*(\omega)\right| \leq \epsilon$$

for any $T \geq N$. Note that N depends only on δ and ϵ, and δ is in turn determined by ϵ. Therefore, N is determined by ϵ, and since we can find such an N for any $\epsilon > 0$, by definition

$$\lim_{T\to\infty} \hat{\theta}_T^*(\omega) = \hat{\theta}_0^*(\omega).$$

Since this convergence holds for any $\omega \in \Omega_0'$, where $\mathbb{P}'(\Omega_0') = 1$, it follows that

$$\hat{\theta}_T^* \overset{a.s.}{\to} \hat{\theta}_0^*$$

in the probability space $(\Omega', \mathcal{H}', \mathbb{P}')$. Since almost sure convergence implies convergence

in probability, which in turn implies weak convergence,

$$\hat{\theta}_T^* \xrightarrow{d} \hat{\theta}_0^*.$$

Finally, since $\hat{\theta}_T^* \sim \hat{\theta}_T$ for any $T \in \mathbb{N}$,

$$\hat{\theta}_T \xrightarrow{d} \hat{\theta}_0.$$

Q.E.D.

3.7 Quantile Regression

So far, we have applied the basic convergence theorem to establish the consistency of extremum estimators under non-compact parameter spaces, and deduced asymptotic normality from the basic asymptotic normality theorem under smoothness assumptions. However, the basic convergence theorem can also be applied to obtain the asymptotic distributions of extremum estimators corresponding to non-smooth objective functions. Perhaps the most useful result in this class of models is quantile regression models. Below we establish the consistency and asymptotic normality of parameter estimators in quantile regression models, which holds even when the sample is weakly dependent.

3.7.1 Quantiles and the Check Function

Let Y be a real valued random variable with distribution function $F_Y(\cdot)$ on $\mathbb{R}$. Suppose $F_Y(\cdot)$ is continuous and strictly increasing on $\mathbb{R}$, so that $F(y) \in (0,1)$ for any $y \in \mathbb{R}$[4]. Then, there exists an inverse function $F_Y^{-1} : (0,1) \to \mathbb{R}$ of $F_Y(\cdot)$ that is also strictly increasing and continuous. Recall that $F_Y^{-1}(\cdot)$ is the quantile function of Y, so that $F_Y^{-1}(\tau)$ yields the unique point y in $\mathbb{R}$ such that the probability Y takes values less than or equal to y is τ. We often denote the value $F_Y^{-1}(\tau)$ by $Q_\tau(Y)$, and call it the τth quantile of Y.

There exists a convenient characterization of the quantile of a real random variable whose distribution function is continuous and strictly increasing on the real line. Given any $\tau \in [0,1]$, we define the check function $\rho_\tau : \mathbb{R} \to \mathbb{R}_+$ as

$$\rho_\tau(u) = (\tau - I_{\{u<0\}})u = \begin{cases} (1-\tau)|u| & \text{if } u < 0 \\ \tau|u| & \text{if } u \geq 0 \end{cases}.$$

We can see that $\rho_\tau(\cdot)$ is a convex function on the real line that attains its minimum 0 when $u = 0$, and which has slope $-(1-\tau)$ when $u < 0$ and τ when $u > 0$. This implies that $\rho_\tau(\cdot)$ is differentiable on all points on the real line except for $u = 0$, with derivatives given as

$$\frac{\partial \rho_\tau(u)}{\partial u} = \begin{cases} -(1-\tau) & \text{if } u < 0 \\ \tau & \text{if } u > 0 \end{cases}.$$

The following lemma reveals that the τth quantile of an absolutely continuous real random variable can be characterized in terms of the check function:

[4]By implication, the support of the random variable Y is the entire real line. At first glance this may seem too strong of an assumption which, in particular, does not allow us to extend the setting to non-negative valued random variables. But in practice this full support assumption is not too harmful, since we can simply log-transform positive valued random variables to turn them into variables with full support.

Lemma (Characterization of Quantiles)
Let Y be a real random variable with distribution μ_Y and strictly increasing distribution function $F_Y : \mathbb{R} \to (0,1)$. Suppose that Y has finite first moment, and that μ_Y is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Then, for any $\tau \in (0,1)$, $Q_\tau(Y)$ is the unique minimizer of the convex function

$$q \mapsto \mathbb{E}\left[\rho_\tau(Y-q)\right].$$

Proof) We can first show that $F_Y(\cdot)$ is continuous and strictly increasing on $\mathbb{R}$. The continuity of $F_Y(\cdot)$ follows from the absolute continuity of μ_Y with respect to the Lebesgue measure; if there exists a point $y^* \in \mathbb{R}$ such that $F_Y(\cdot)$ is discontinuous at y, then

$$\lim_{y \nearrow y^*} F_Y(y) < F_Y(y^*),$$

so that

$$\mu_Y(\{y^*\}) = F_Y(y^*) - \lim_{y \nearrow y^*} F_Y(y) > 0.$$

This contradicts the fact that $\{y^*\}$ has measure 0 under the Lebesgue measure and therefore under μ_Y as well. This shows us that the τth quantile $Q_\tau(Y)$ of Y is well-defined as the unique value such that $F_Y(Q_\tau(Y)) = \tau$.

For any $q \in \mathbb{R}$, because $\rho_\tau(Y-q) = (\tau - I_{(-\infty,q)} \circ Y)(Y-q)$ is a random variable that is bounded above by the integrable random variable $(\tau+1)(|Y|+|q|)$, it follows that the mapping

$$q \mapsto \mathbb{E}\left[\rho_\tau(Y-q)\right]$$

is a well-defined real valued function on $\mathbb{R}$. This mapping defines a convex function on $\mathbb{R}$: for any $\lambda \in [0,1]$ and $q_1, q_2 \in \mathbb{R}$, because

$$\begin{aligned}\rho_\tau(Y - (\lambda q_1 + (1-\lambda)q_2)) &= \rho_\tau(\lambda(Y-q_1) + (1-\lambda)(Y_2 - q_2)) \\ &\leq \lambda\rho_\tau(Y-q_1) + (1-\lambda)\rho_\tau(Y-q_2)\end{aligned}$$

everywhere on the sample space by the convexity of $\rho_\tau(\cdot)$, we have

$$\mathbb{E}\left[\rho_\tau(Y - (\lambda q_1 + (1-\lambda)q_2))\right] \leq \lambda\mathbb{E}\left[\rho_\tau(Y-q_1)\right] + (1-\lambda)\mathbb{E}\left[\rho_\tau(Y-q_2)\right]$$

as well.

Define the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as

$$f(y,q) = \rho_\tau(y-q)$$

for any $(y,q) \in \mathbb{R}^2$. For any $q \in \mathbb{R}$, we already established that $f(\cdot,q)$ is a μ_Y-integrable function.

Furthermore,

$$\frac{\partial f(y,q)}{\partial q} = \frac{\partial \rho_\tau(y-q)}{\partial q}$$

exists for any $y \in \mathbb{R}$ such that $y \neq q$. Defining

$$A_q = \mathbb{R} \setminus \{q\},$$

since $A_q^c = \{q\}$ has measure 0 under the Lebesgue measure, and μ_Y is absolutely continuous with respect to the Lebesgue measure, $\mu_Y(A_q^c) = 0$ as well. The resulting function $y \mapsto \frac{\partial f(y,q)}{\partial q}$ defined on A_q is given as

$$\frac{\partial f(y,q)}{\partial q} = \begin{cases} -(1-\tau) & \text{if } y < q \\ \tau & \text{if } y \geq q \end{cases},$$

so that, being piecewise constant, it is a Borel measurable function.

Finally, we can see that, for any $(y,q) \in \mathbb{R}^2$ such that the derivative $\frac{\partial f(y,q)}{\partial q}$ exists,

$$\left| \frac{f(y,q+h) - f(y,q)}{h} \right| = \left| \frac{\rho_\tau(y-q-h) - \rho_\tau(y-q)}{h} \right| \leq 1$$

for any $h \neq 0$; a simple graphical argument suffices to establish this inequality, since the mapping $z \mapsto \rho_\tau(z-h)$ is just the version of the wedge $\rho_\tau(\cdot)$ shifted by h units to the left.

Therefore, Leibniz' rule tells us that the mapping $q \mapsto \mathbb{E}[\rho_\tau(Y-q)]$ is differentiable on $\mathbb{R}$ with derivative given as

$$\begin{aligned} \frac{\partial \mathbb{E}[\rho_\tau(Y-q)]}{\partial q} &= \frac{\partial}{\partial q} \int_{\mathbb{R}} \rho_\tau(y-q) d\mu_Y(y) \\ &= \int_{\mathbb{R}} \left(\frac{\partial \rho_\tau(y-q)}{\partial q} \right) d\mu_Y(y) \\ &= -(1-\tau) \cdot \mu_Y((-\infty, q)) + \tau \cdot \mu_Y([q, +\infty)) \\ &= -(1-\tau) F_Y(q) + \tau(1 - F_Y(q)) \end{aligned}$$

for any $q \in \mathbb{R}$. This tells us that

$$\frac{\partial \mathbb{E}\left[\rho_\tau(Y-q)\right]}{\partial q} = 0 \quad \text{if and only if} \quad -(1-\tau)F_Y(q) + \tau(1-F_Y(q)) = 0,$$

or if and only if $F_Y(q) = \tau$. Under our assumptions, $Q_\tau(Y)$ is the only value that solves this equation, so $Q_\tau(Y)$ is the unique value that satisfies the first order condition for minimization. The convexity of $q \mapsto \mathbb{E}\left[\rho_\tau(Y-q)\right]$ now implies that $Q_\tau(Y)$ is its unique minimizer on $\mathbb{R}$.

Q.E.D.

This characterization allows us to obtain a very straightforward formula to compute quantile regression estimators via the analogy principle later on.

3.7.2 Motivating the Quantile Regression Estimator

We now extend the basic framework above to a more general setting of stochastic processes. Let $\{Z_t = (Y_t, X_t)\}_{t\in\mathbb{Z}}$ be a doubly infinite stochastic process such that Y_t is real valued and X_t is k-dimensional for each t. Let $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{Z}\}$ be some filtration, and suppose that each Y_t is $\mathcal{F}_t$-measurable, while each X_t is $\mathcal{F}_{t-1}$-measurable. This implies that $\{Z_t\}_{t\in\mathbb{Z}}$ is $\mathcal{F}$-adapted.

This setting can accommodate many different modes of dependence among sample observations Z_t. If $\{Z_t\}_{t\in\mathbb{Z}}$ is iid, then we can simply define $\mathcal{F}_t$ as the σ-algebra generated by Y_t and the entirety of the exogenous regressors $\{X_s\}_{s\in\mathbb{Z}}$. It can also describe the case where $\{Z_t\}_{t\in\mathbb{Z}}$ is stationary and X_t contains lags of Y_t; in this case, $\mathcal{F}$ would simply be the filtration generated by $\{Y_s\}_{s\in\mathbb{Z}}$.

Due to the non-linearity inherent in any model that deals with quantiles, we impose stronger conditions than the usual weak stationarity assumption on $\{Z_t\}_{t\in\mathbb{Z}}$. Specifically, we assume that, for any $t \in \mathbb{Z}$, there exists a regular version $Q_{t-1} : \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ of the conditional probability of Y_t given $\mathcal{F}_{t-1}$; that is, Q_{t-1} is a transition probability kernel from $(\Omega, \mathcal{F}_{t-1})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$Q_{t-1}(\cdot, A) = \mathbb{P}\left(Y_t \in A \mid \mathcal{F}_{t-1}\right)$$

for any $A \in \mathcal{B}(\mathbb{R})$. We also define $F_{t-1} : \Omega \times \mathbb{R} \to [0,1]$ as

$$F_{t-1}(\omega, y) = Q_{t-1}(\omega, (-\infty, y])$$

for any $(\omega, y) \in \Omega \times \mathbb{R}$. Since $Q_{t-1}(\omega, \cdot)$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for any $\omega \in \Omega$, $F_{t-1}(\cdot, y)$ can be interpreted as the conditional distribution function of Y_t given $\mathcal{F}_{t-1}$ evaluated at the point $y \in \mathbb{R}$.

As in the preceding section, we assume that, for each $\omega \in \Omega$, $F_{t-1}(\omega, \cdot)$ takes values in $(0,1)$

and that it is strictly increasing on $\mathbb{R}$. Furthermore, to make use of the check function characterization of quantiles, we assume that each $Q_{t-1}(\omega,\cdot)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$; we saw above how this implies that $F_{t-1}(\omega,\cdot)$ is strictly increasing and continuous on $\mathbb{R}$. Since $F_{t-1}(\cdot,y)$ is a $\mathcal{F}_{t-1}$-measurable function by the fact that Q_{t-1} is a transition probability kernel, it follows that $F_{t-1}(\cdot)$ is measurable relative to $\mathcal{F}_{t-1}\bigotimes\mathcal{B}(\mathbb{R})$.

It follows that we can define the mapping $Q_\tau(Y_t \mid \mathcal{F}_{t-1}) : \Omega \to \mathbb{R}$ so that, for any $\omega \in \Omega$, $[Q_\tau(Y_t \mid \mathcal{F}_{t-1})](\omega)$ is the unique solution to the equation

$$F_{t-1}(\omega, q) = \tau.$$

We call $Q_\tau(Y_t \mid \mathcal{F}_{t-1})$ the conditional τth quantile of Y_t given $\mathcal{F}_{t-1}$. The quantile characterization derived above implies that $[Q_\tau(Y_t \mid \mathcal{F}_{t-1})](\omega)$ is the unique minimizer of the mapping

$$q \mapsto \int_{\mathbb{R}} \rho_\tau(y-q) Q_{t-1}(\omega, dy)$$

for any $\omega \in \Omega$. Equivalently, we can say that $Q_\tau(Y_t \mid \mathcal{F}_{t-1})$ is the unique minimizer of the mapping

$$q \mapsto \mathbb{E}\left[\rho_\tau(Y_t - q) \mid \mathcal{F}_{t-1}\right].$$

In a linear regression model, we model the conditional expectation of Y_t given $\mathcal{F}_{t-1}$ as a linear function of the regressors X_t:

$$\mathbb{E}\left[Y_t \mid \mathcal{F}_{t-1}\right] = X_t'\beta$$

for some $\beta \in \mathbb{R}^k$. Similarly, in a quantile regression model, the τth conditional quantile of Y_t given $\mathcal{F}_{t-1}$ is modeled as a linear function of X_t:

$$Q_\tau(Y_t \mid \mathcal{F}_{t-1}) = X_t'\beta(\tau)$$

for some $\beta(\tau) \in \mathbb{R}^k$. Under this assumption, it follows that

$$\mathbb{E}\left[\rho_\tau(Y_t - X_t'\beta(\tau)) \mid \mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\rho_\tau(Y_t - W) \mid \mathcal{F}_{t-1}\right]$$

for any $\mathcal{F}_{t-1}$-measurable real random variable W, and in particular,

$$\mathbb{E}\left[\rho_\tau(Y_t - X_t'\beta(\tau)) \mid \mathcal{F}_{t-1}\right] \leq \mathbb{E}\left[\rho_\tau(Y_t - X_t'b) \mid \mathcal{F}_{t-1}\right]$$

for any $b \in \mathbb{R}^k$. Taking expectations on both sides implies that

$$\mathbb{E}\left[\rho_\tau(Y_t - X_t'\beta(\tau))\right] \leq \mathbb{E}\left[\rho_\tau(Y_t - X_t'b)\right],$$

so that

$$\beta(\tau) \in \operatorname{argmin}_{b \in \mathbb{R}^k} \mathbb{E}\left[\rho_\tau(Y_t - X_t' b)\right].$$

This is how the coefficients $\beta(\tau)$ in a quantile regression model are identified, and they are estimated by the analogy principle; given a sample of size T, the estimator $\hat{\beta}_T(\tau)$ of $\beta(\tau)$ is found as the solution to the minimization problem

$$\min_{b \in \mathbb{R}^k} \quad \frac{1}{T} \sum_{t=1}^{T} \rho_\tau(Y_t - X_t' b).$$

In the next section we state the concepts we have introduced so far as formal assumptions, and rigorously prove that the quantile regression estimator $\hat{\beta}_T(\tau)$ is consistent and asymptotically normal for $\beta(\tau)$ under our assumptions.

3.7.3 Asymptotic Properties of the Quantile Regression Estimator

We make the following assumptions:

Q1 Data Generating Process

Let $\{Z_t = (Y_t, X_t)\}_{t \in \mathbb{Z}}$ be a stochastic process, where Y_t is real valued and X_t is k-dimensional. Let $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{Z}\}$ be some filtration such that $Y_t \in \mathcal{F}_t$ and $X_t \in \mathcal{F}_{t-1}$ for any $t \in \mathbb{Z}$.

We assume that $\{Z_t\}_{t \in \mathbb{Z}}$ is strictly stationary and ergodic; in particular, for any measurable function $g : \mathbb{R}^{k+1} \to \mathbb{R}$,

$$\frac{1}{T} \sum_{t=1}^{T} g(Z_t) \xrightarrow{p} \mathbb{E}\left[g(Z_0)\right],$$

granted that the expectation on the right hand side exists, and we assume that there exists some positive definite $k \times k$ matrix Σ_{xx} such that

$$\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \xrightarrow{p} \Sigma_{xx} = \mathbb{E}\left[X_0 X_0'\right].$$

Q2 Conditional Distribution Function and Density

We assume that for any $t \in \mathbb{Z}$ there exists a transition probability kernel Q_{t-1} from $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$Q_{t-1}(\cdot, A) \circ X_t = \mathbb{P}\left(Y_t \in A \mid \mathcal{F}_{t-1}\right)$$

for any $A \in \mathcal{B}(\mathbb{R})$. In other words, we assume that X_t is a sufficient statistic for the conditional distribution of Y_t given $\mathcal{F}_{t-1}$. We define the conditional distribution function $F_{t-1} : \mathbb{R}^{k+1} \to [0,1]$ as

$$F_{t-1}(y \mid x) = Q_{t-1}(x, (-\infty, y])$$

for any $(x, y) \in \mathbb{R}^{k+1}$.

We assume that there exists a conditional density function $f_{t-1} : \mathbb{R}^{k+1} \to \mathbb{R}_{++}$ such that $f_{t-1}(\cdot)$ is Borel measurable and

$$F_{t-1}(y \mid x) = \int_{-\infty}^{y} f_{t-1}(z \mid x) dz.$$

for any $(x, y) \in \mathbb{R}^{k+1}$. By implication, each $F_{t-1}(\cdot \mid x)$ is absolutely continuous with respect to the Lebesgue measure, and

$$\frac{\partial F_{t-1}(y \mid x)}{\partial y} = f_{t-1}(y \mid x)$$

for any $(x, y) \in \mathbb{R}^{k+1}$ by the fundamental theorem of calculus. We assume also that $F_{t-1}(\cdot \mid x)$ is strictly increasing on $\mathbb{R}$ for any $x \in \mathbb{R}^k$.

Q3 Conditional Quantile Function

For any $\tau \in (0,1)$, we let $Q_\tau(Y_t \mid \mathcal{F}_{t-1})$ be the conditional τth quantile of Y_t given $\mathcal{F}_{t-1}$. We assume that there exists some $\beta(\tau) \in \mathbb{R}^k$ such that

$$Q_\tau(Y_t \mid \mathcal{F}_{t-1}) = X_t' \beta(\tau)$$

for any $t \in \mathbb{Z}$. By implication, $\beta(\tau)$ is a solution of the minimization problem

$$\min_{b \in \mathbb{R}^k} \quad \mathbb{E}\left[\rho_\tau(Y_0 - X_0' b)\right].$$

Q4 Local Uniform Equicontinuity Assumption

We assume that $f_{t-1}(\cdot)$ is uniformly equicontinuous locally around each quantile; that is, for any $\tau \in (0,1)$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_{t \in \mathbb{Z}} \left| f_{t-1}(y \mid x) - f_{t-1}(x'\beta(\tau) \mid x) \right| < \epsilon$$

for any $x \in \mathbb{R}^k$ and $y \in \mathbb{R}$ such that $|y - x'\beta(\tau)| < \delta$.

Q5 Miscellaneous Asymptotic Arguments

We assume that $\{Z_t\}_{t\in\mathbb{Z}}$ has bounded fourth moments, so that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|Z_t|^4 < +\infty.$$

Furthermore, we assume that, for any $\tau \in (0,1)$, there exists a positive definite $k \times k$ matrix $\Sigma_f(\tau)$ such that

$$\frac{1}{T}\sum_{t=1}^{T} f_{t-1}(X_t'\beta(\tau) \mid X_t) X_t X_t' \xrightarrow{p} \Sigma_f(\tau).$$

The strongest assumptions we make are the stationarity and ergodicity assumptions imposed on the process $\{Z_t\}_{t\in\mathbb{Z}}$, and the local uniform equicontinuity assumption on the conditional density function. The former is necessary to ensure that laws of large numbers hold even in the presence of large non-linearities; it can be replaced with standard moment boundedness conditions if the data is iid. The latter, meanwhile, is a technical condition that ensures the convexity of the limiting objective function. This assumption is more plausible if the density function is time-invariant, as they happen to be in many practical applications.

Below we show, using the basic convergence theorem, that the quantile regression estimator is consistent and asymptotically normal under our assumptions. The argument is exactly that of Koenker and Xiao (2004).

Lemma For any $u, v \in \mathbb{R}$ and $\tau \in (0,1)$,

$$\rho_\tau(u-v) - \rho_\tau(u) = -v\left(\tau - I_{\{u<0\}}\right) + \int_0^v \left(I_{\{u\le s\}} - I_{\{u\le 0\}}\right) ds.$$

Proof) For any $u, v \in \mathbb{R}$,

$$\begin{aligned}
\rho_T(u-v) - \rho_T(u) &= \left(\tau - I_{\{u<v\}}\right)(u-v) - \left(\tau - I_{\{u<0\}}\right)u \\
&= -v\left(\tau - I_{\{u<0\}}\right) + \left(I_{\{u<0\}} - I_{\{u<v\}}\right)(u-v) \\
&= -v\left(\tau - I_{\{u<0\}}\right) + \left(I_{\{v\le u<0\}} - I_{\{0\le u<v\}}\right)(u-v).
\end{aligned}$$

The second term can be expressed succinctly as an integral:

$$\int_0^v \left(I_{\{u\le s\}} - I_{\{u\le 0\}}\right) ds = \left(I_{\{v\le u<0\}} - I_{\{0\le u<v\}}\right)(u-v).$$

To see that this equation holds true, we consider the following distinct and exhaustive cases:

Case 1 $v \leq 0$, $v \leq u < 0$

In this case, the right hand side is equal to $u - v$. Meanwhile,

$$\begin{aligned}\int_0^v \left(I_{\{u \leq s\}} - I_{\{u \leq 0\}}\right) ds &= -\int_v^0 \left(I_{\{u \leq s\}} - 1\right) ds \\ &= -\left[(0-u) - (-v)\right] = u - v.\end{aligned}$$

Case 2 $v \leq 0$, $u < v \leq 0$

In this case, the right hand side is equal to 0. Meanwhile,

$$\begin{aligned}\int_0^v \left(I_{\{u \leq s\}} - I_{\{u \leq 0\}}\right) ds &= -\int_v^0 \left(I_{\{u \leq s\}} - 1\right) ds \\ &= -\left[(0-v) - (-v)\right] = 0.\end{aligned}$$

Case 3 $v > 0$, $0 \leq u < v$

In this case, the right hand side is equal to $v - u$. Meanwhile,

$$\int_0^v \left(I_{\{u \leq s\}} - I_{\{u \leq 0\}}\right) ds = \int_0^v I_{\{u \leq s\}} ds = v - u.$$

Case 4 $v > 0$, $0 < v \leq u$

In this case, the right hand side is equal to 0, while

$$\int_0^v \left(I_{\{u \leq s\}} - I_{\{u \leq 0\}}\right) ds = \int_0^v I_{\{u \leq s\}} ds = 0.$$

We have manually verified that

$$\rho_T(u-v) - \rho_T(u) = -v\left(\tau - I_{\{u<0\}}\right) + \int_0^v \left(I_{\{u \leq s\}} - I_{\{u \leq 0\}}\right) ds$$

for any $u, v \in \mathbb{R}$.

Q.E.D.

Theorem (Asymptotic Normality of Quantile Regression Estimator)
Maintain assumptions Q1 to Q6, and for any $\tau \in (0,1)$, let $\{\hat{\beta}_T(\tau)\}_{T\in N_+}$ be a sequence of quantile regression estimators, where $\hat{\beta}_T(\tau)$ is obtained as a solution to the following minimization problem:

$$\min_{b\in\mathbb{R}^k} \quad \frac{1}{T}\sum_{t=1}^{T}\rho_T(Y_t - X_t'b).$$

Then,

$$\sqrt{T}\left[\hat{\beta}_T(\tau) - \beta(\tau)\right] \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\ \tau(1-\tau)\Sigma_f(\tau)^{-1}\Sigma_{xx}\Sigma_f(\tau)^{-1}\right].$$

Proof) Define

$$u_t = Y_t - X_t'\beta(\tau).$$

We can then see that $\sqrt{T}\left[\hat{\beta}_T(\tau) - \beta(\tau)\right]$ solves the minimization problem

$$\min_{b\in\mathbb{R}^k} \quad \sum_{t=1}^{T}\left[\rho_T\left(u_t - \frac{1}{\sqrt{T}}b'X_t\right) - \rho_T\left(u_t\right)\right].$$

For any $T \in N_+$, define the stochastic process $\{V_T(b)\}_{b\in\mathbb{R}^k}$ as

$$V_T(b) = \sum_{t=1}^{T}\left[\rho_T\left(u_t - \frac{1}{\sqrt{T}}b'X_t\right) - \rho_T\left(u_t\right)\right]$$

for any $b \in \mathbb{R}^k$. The convexity of $\rho_T(\cdot)$ implies that $\{V_T(b)\}_{b\in\mathbb{R}^k}$ has convex paths, and since $\sqrt{T}\left[\hat{\beta}_T(\tau) - \beta(\tau)\right]$ exists as a solution to the minimization problem above, the paths of $\{V_T(b)\}_{b\in\mathbb{R}^k}$ attain their minimum on $\mathbb{R}^k$.

It follows that $\{V_T(b)\}_{b\in\mathbb{Q}^k}$ is a random sequence in $\mathbb{R}^\infty$ taking values in the Borel measurable subset F_0. By the basic convergence theorem, our proof will be done if we can just show that:

- The finite dimensional distributions of $\{V_T(b)\}_{b\in\mathbb{Q}^k}$ converge to those of some random sequence $\{V_0(b)\}_{b\in\mathbb{Q}^k}$.

- There exists some measurable subset F^* of F_0 such that $\{V_0(b)\}_{b\in\mathbb{Q}^k}$ takes values in F^* and, for any $\mathbf{y} \in F^*$, the corresponding convex function has a unique minimizer.

- The unique minimizer of $\{V_0(b)\}_{b\in\mathbb{Q}^k}$ has the following distribution:

$$\hat{\beta}_0(\tau) \sim \mathcal{N}\left[O_{k\times 1},\ \tau(1-\tau)\Sigma_{xx}\right].$$

We start by showing that there exists some random sequence $\{V_0(b)\}_{b\in\mathbb{Q}^k}$ such that, for any distinct points $b_1,\cdots,b_n\in\mathbb{Q}^k$,

$$(V_T(b_1),\cdots,V_T(b_n)) \xrightarrow{d} (V_0(b_1),\cdots,V_0(b_n)).$$

as $T\to\infty$.

Choose any $b\in\mathbb{Q}^k$. Recall from the above lemma that

$$\rho_T(u-v)-\rho_T(u) = -v\left(\tau - I_{\{u<0\}}\right) + \int_0^v \left(I_{\{u\leq s\}} - I_{\{u\leq 0\}}\right) ds$$

for any $(u,v)\in\mathbb{R}$. Applying this result yields

$$\begin{aligned} V_T(b) &= \sum_{t=1}^T \left[\rho_T\left(u_t - \frac{1}{\sqrt{T}}b'X_t\right) - \rho_T(u_t)\right] \\ &= -b'\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T X_t\left(\tau - I_{(-\infty,0)}\circ u_t\right)\right] + \sum_{t=1}^T \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\right) ds. \end{aligned}$$

We focus first on the term on the left.

Note that the process $\{X_t(\tau - I_{(-\infty,0)}\circ u_t)\}_{t\in\mathbb{Z}}$ constitutes an MDS with respect to the filtration $\mathcal{F}$; it is clearly $\mathcal{F}$-adapted, integrable and

$$\begin{aligned} \mathbb{E}\left[X_t(\tau - I_{(-\infty,0)}\circ u_t) \mid \mathcal{F}_{t-1}\right] &= X_t\left(\tau - \mathbb{P}\left(u_t < 0 \mid \mathcal{F}_{t-1}\right)\right) \\ &= X_t\left(\tau - \mathbb{P}\left(Y_t < X_t'\beta(\tau) \mid \mathcal{F}_{t-1}\right)\right) \\ &= X_t\left[\tau - F_{t-1}(\cdot, X_t'\beta(\tau))\right] \\ &= X_t\left[\tau - F_{t-1}(\cdot, Q_\tau(Y_t \mid \mathcal{F}_{t-1}))\right] = O_{k\times 1}, \end{aligned}$$

where the final equality follows from the definition of the conditional τth quantile. We can therefore apply the martingale difference CLT if we can establish the following:

i) **Boundedness of Higher Order Moments**
Note that, for any $t\in\mathbb{Z}$,

$$\left|X_t(\tau - I_{(-\infty,0)}\circ u_t)\right| \leq (\tau+1)|X_t|,$$

so that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left|X_t(\tau - I_{(-\infty,0)}\circ u_t)\right|^4 \leq (\tau+1)^4\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|X_t|^4\right) < +\infty,$$

where the final inequality follows from the assumption of bounded fourth mo-

ments. It follows that $\{X_t(\tau - I_{(-\infty,0)} \circ u_t)\}_{t\in\mathbb{Z}}$ is L^4-bounded.

ii) **Convergence of Sample Variance**
For any $t \in \mathbb{Z}$,

$$X_t X_t'(\tau - I_{(-\infty,0)} \circ u_t)^2 = X_t X_t' \left(\tau^2 - (2\tau - 1) \cdot I_{(-\infty,0)} \circ u_t\right),$$

so that

$$\begin{aligned}
\mathbb{E}\left[X_t X_t'(\tau - I_{(-\infty,0)} \circ u_t)^2\right] &= \mathbb{E}\left[X_t X_t' \left(\tau^2 - (2\tau-1)\mathbb{P}\left(u_t < 0 \mid \mathcal{F}_{t-1}\right)\right)\right] \\
&= \mathbb{E}\left[X_t X_t' \left(\tau^2 - (2\tau-1)\tau\right)\right] \\
&= \tau(1-\tau)\mathbb{E}\left[X_t X_t'\right] = \tau(1-\tau)\Sigma_{xx}.
\end{aligned}$$

Therefore, the ergodicity of the stationary process $\{Z_t\}_{t\in\mathbb{Z}}$ implies that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t'(\tau - I_{(-\infty,0)} \circ u_t)^2 \xrightarrow{p} \tau(1-\tau)\Sigma_{xx},$$

where the probability limit is positive definite by the assumed positive definiteness of Σ_{xx} and the fact that $\tau \in (0,1)$.

We can therefore immediately apply the MDS CLT to conclude that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t(\tau - I_{(-\infty,0)} \circ u_t) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\, \tau(1-\tau)\Sigma_{xx}\right].$$

Letting $\mathcal{Z} \sim \mathcal{N}\left[O_{k\times 1},\, \tau(1-\tau)\Sigma_{xx}\right]$, it now follows that

$$-b'\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t(\tau - I_{(-\infty,0)} \circ u_t)\right] \xrightarrow{d} -b'\mathcal{Z}.$$

Moving onto the second term, define the triangular array $\{\zeta_{Tt}(b)\}_{1\leq t\leq T, T\in N_+}$ as

$$\zeta_{Tt}(b) = \int_0^{\frac{1}{\sqrt{T}}b'X_t} \Big(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\Big)\, ds$$

for any $1\leq t\leq T$ and $T\in N_+$. By Fubini's theorem applied to the bounded integrand, we have

$$\begin{aligned}
\mathbb{E}\left[\zeta_{Tt}(b)\mid \mathcal{F}_{t-1}\right] &= \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left[\mathbb{P}\left(u_t\leq s\mid \mathcal{F}_{t-1}\right) - \mathbb{P}\left(u_t\leq 0\mid \mathcal{F}_{t-1}\right)\right] ds \\
&= \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left[F_{t-1}(s+X_t'\beta(\tau)\mid X_t) - F_{t-1}(X_t'\beta(\tau)\mid X_t)\right] ds \\
&= \int_0^{\frac{1}{\sqrt{T}}b'X_t} f_{t-1}(X_t'\beta(\tau)\mid X_t) s ds \\
&\quad + \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left[\frac{F_{t-1}(s+X_t'\beta(\tau)\mid X_t) - F_{t-1}(X_t'\beta(\tau)\mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau)\mid X_t)\right] s ds.
\end{aligned}$$

Note that, for any $s\in\mathbb{R}$,

$$\begin{aligned}
&\left|\frac{F_{t-1}(s+X_t'\beta(\tau)\mid X_t) - F_{t-1}(X_t'\beta(\tau)\mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau)\mid X_t)\right| \\
&\qquad = \frac{1}{s}\left|\int_{X_t'\beta(\tau)}^{s+X_t'\beta(\tau)} \Big[f_{t-1}(z\mid X_t) - f_{t-1}(X_t'\beta(\tau)\mid X_t)\Big] dz\right|.
\end{aligned}$$

Choose any $\epsilon>0$. By the local uniform equicontinuity assumption, there exists a $\delta>0$ such that

$$\left|f_{t-1}(z\mid x) - f_{t-1}(x'\beta(\tau)\mid x)\right| < \frac{\epsilon}{b'\Sigma_{xx}b+1}$$

for any $x\in\mathbb{R}^k$, $t\in\mathbb{Z}$ and any $z\in\mathbb{R}$ such that $|z-x'\beta(\tau)|\leq\delta$. By implication, as long as $0<|s|\leq\delta$,

$$\left|\frac{F_{t-1}(s+X_t'\beta(\tau)\mid X_t) - F_{t-1}(X_t'\beta(\tau)\mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau)\mid X_t)\right| \leq \frac{\epsilon}{b'\Sigma_{xx}b+1}.$$

Suppose that

$$\max_{1\leq t\leq T}\left|\frac{1}{\sqrt{T}}b'X_t\right|\leq\delta \quad\text{and}\quad \left|b'\left(\frac{1}{T}\sum_{t=1}^T X_tX_t'\right)b - b'\Sigma_{xx}b\right|\leq 1.$$

Then, since each $\left|\frac{1}{\sqrt{T}}b'X_t\right|\leq\delta$,

$$\left|\int_0^{\frac{1}{\sqrt{T}}b'X_t}\left[\frac{F_{t-1}(s+X_t'\beta(\tau)\mid X_t) - F_{t-1}(X_t'\beta(\tau)\mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau)\mid X_t)\right] s ds\right|$$

$$\leq \int_0^{\left|\frac{1}{\sqrt{T}}b'X_t\right|} \left| \frac{F_{t-1}(s+X_t'\beta(\tau) \mid X_t) - F_{t-1}(X_t'\beta(\tau) \mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau) \mid X_t) \right| sds$$

$$\leq \frac{\epsilon}{b'\Sigma_{xx}b+1} \int_0^{\left|\frac{1}{\sqrt{T}}b'X_t\right|} sds$$

$$= \frac{\epsilon}{b'\Sigma_{xx}b+1} \times \frac{1}{2T} b'X_tX_t'b$$

for $1 \leq t \leq T$, so that

$$\left| \sum_{t=1}^{T} \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left[\frac{F_{t-1}(s+X_t'\beta(\tau) \mid X_t) - F_{t-1}(X_t'\beta(\tau) \mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau) \mid X_t) \right] sds \right|$$

$$\leq \frac{\epsilon}{b'\Sigma_{xx}b+1} \times b' \left(\frac{1}{2T} \sum_{t=1}^{T} X_tX_t' \right) b \leq \epsilon.$$

Therefore,

$$\mathbb{P}\left(\left| \sum_{t=1}^{T} \int_0^{\frac{1}{\sqrt{T}}b'X_t} \left[\frac{F_{t-1}(s+X_t'\beta(\tau) \mid X_t) - F_{t-1}(X_t'\beta(\tau) \mid X_t)}{s} - f_{t-1}(X_t'\beta(\tau) \mid X_t) \right] sds \right| > \epsilon \right)$$

$$\leq \mathbb{P}\left(\max_{1\leq t\leq T} \left| \frac{1}{\sqrt{T}} b'X_t \right| > \delta \right) + \mathbb{P}\left(\left| b' \left(\frac{1}{T} \sum_{t=1}^{T} X_tX_t' \right) b - b'\Sigma_{xx}b \right| > 1 \right).$$

The rightmost term goes to 0 because the stationarity and ergodicity of $\{Z_t\}_{t\in\mathbb{Z}}$ implies

$$b' \left(\frac{1}{T} \sum_{t=1}^{T} X_tX_t' \right) b \xrightarrow{p} b'\Sigma_{xx}b.$$

Meanwhile, because

$$\mathbb{P}\left(\max_{1\leq t\leq T} \left| \frac{1}{\sqrt{T}} b'X_t \right| > \delta \right) \leq \mathbb{P}\left(\max_{1\leq t\leq T} |X_t| > \frac{\delta}{|b|+1} \cdot \sqrt{T} \right)$$

and

$$\left\{ \max_{1\leq t\leq T} |X_t| > \frac{\delta}{|b|+1} \cdot \sqrt{T} \right\} \subset \bigcup_{t=1}^{T} \left\{ |X_t| > \frac{\delta}{|b|+1} \cdot \sqrt{T} \right\},$$

we can see that

$$\mathbb{P}\left(\max_{1\leq t\leq T} \left| \frac{1}{\sqrt{T}} b'X_t \right| > \delta \right) \leq \sum_{t=1}^{T} \mathbb{P}\left(|X_t| > \frac{\delta}{|b|+1} \cdot \sqrt{T} \right)$$

$$\leq \sum_{t=1}^{T} \frac{(|b|+1)^4}{T^2\delta^4} \mathbb{E}|X_t|^4$$

$$\leq \frac{(|b|+1)^4}{T\delta^4}\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|X_t|^4\right).$$

By the assumption of finite fourth moments, taking $T \to \infty$ on both sides shows us that

$$\lim_{T\to\infty}\mathbb{P}\left(\max_{1\leq t\leq T}\left|\frac{1}{\sqrt{T}}b'X_t\right| > \delta\right) = 0.$$

Therefore,

$$\lim_{T\to\infty}\mathbb{P}\left(\left|\sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b'X_t}\left[\frac{F_{t-1}(s+X_t'\beta(\tau)\mid X_t)-F_{t-1}(X_t'\beta(\tau)\mid X_t)}{s}-f_{t-1}(X_t'\beta(\tau)\mid X_t)\right]sds\right| > \epsilon\right) = 0.$$

This holds for any $\epsilon > 0$, so

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\left[\zeta_{Tt}(b)\mid\mathcal{F}_{t-1}\right] &= \sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b'X_t} f_{t-1}(X_t'\beta(\tau)\mid X_t)sds + o_p(1)\\ &= b'\left(\frac{1}{2T}\sum_{t=1}^{T}f_{t-1}(X_t'\beta(\tau)\mid X_t)X_tX_t'\right)b.\end{aligned}$$

By assumption,

$$\sum_{t=1}^{T}\mathbb{E}\left[\zeta_{Tt}(b)\mid\mathcal{F}_{t-1}\right]\xrightarrow{p}\frac{1}{2}b'\Sigma_f(\tau)b.$$

We have shown that

$$\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b) \mid \mathcal{F}_{t-1}\right] \xrightarrow{p} \frac{1}{2} b' \Sigma_f(\tau) b.$$

It remains to show that

$$\sum_{t=1}^{T} \zeta_{Tt}(b) \xrightarrow{p} \frac{1}{2} b' \Sigma_f(\tau) b$$

as well; to this end, we show that

$$\sum_{t=1}^{T} \left(\zeta_{Tt}(b) - \mathbb{E}\left[\zeta_{Tt}(b) \mid \mathcal{F}_{t-1}\right]\right) \xrightarrow{p} 0.$$

Define the triangular array $\{W_{Tt}(b)\}_{1 \leq t \leq T, T \in N_+}$ as

$$W_{Tt}(b) = \zeta_{Tt}(b) - \mathbb{E}\left[\zeta_{Tt}(b) \mid \mathcal{F}_{t-1}\right].$$

Then, $\{W_{Tt}(b)\}_{1 \leq t \leq T, T \in N_+}$ is a martingale difference array with respect to the filtration $\mathcal{F}$: each $W_{Tt}(b)$ is clearly $\mathcal{F}_t$-measurable,

$$\mathbb{E}\left[W_{Tt}(b)\right] = \mathbb{E}\left[\zeta_{Tt}(b)\right] - \mathbb{E}\left[\zeta_{Tt}(b)\right] = 0$$

for each $1 \leq t \leq T$, and for $2 \leq t \leq T$,

$$\mathbb{E}\left[W_{Tt}(b) \mid \mathcal{F}_{t-1}\right] = 0.$$

It follows that $\{W_{T1}(b), \cdots, W_{TT}(b)\}$ is a collection of mean zero serially uncorrelated random variables for each $T \in N_+$: by implication,

$$\begin{aligned}
\mathbb{E}\left|\sum_{t=1}^{T} W_{Tt}(b)\right|^2 &= \sum_{t=1}^{T} \mathbb{E}|W_{Tt}(b)|^2 \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\left(\zeta_{Tt} - \mathbb{E}\left[\zeta_{Tt}(b) \mid \mathcal{F}_{t-1}\right]\right)^2 \mid \mathcal{F}_{t-1}\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\zeta_{Tt}(b) \mid \mathcal{F}_{t-1}\right]^2\right] \\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right]\right].
\end{aligned}$$

Note that, for any $1 \le t \le T$,

$$\begin{aligned}
\zeta_{Tt}(b)^2 &= \left(\int_0^{\frac{1}{\sqrt{T}} b' X_t} \left(I_{(-\infty,s]} \circ u_t - I_{(-\infty,0]} \circ u_t \right) ds \right)^2 \\
&\le \left| \frac{1}{\sqrt{T}} b' X_t \right| \int_0^{\frac{1}{\sqrt{T}} b' X_t} \left(I_{(-\infty,s]} \circ u_t - I_{(-\infty,0]} \circ u_t \right)^2 ds \\
&= \left| \frac{1}{\sqrt{T}} b' X_t \right| \int_0^{\frac{1}{\sqrt{T}} b' X_t} \left(I_{(-\infty,s]} \circ u_t + I_{(-\infty,0]} \circ u_t - 2 \cdot I_{(-\infty,s] \cap (-\infty,0]} \circ u_t \right) ds,
\end{aligned}$$

where the first inequality is justified by Hölder's inequality. Since

$$\begin{aligned}
&\mathbb{E}\left[I_{(-\infty,s]} \circ u_t + I_{(-\infty,0]} \circ u_t - 2 \cdot I_{(-\infty,s] \cap (-\infty,0]} \circ u_t \mid \mathcal{F}_{t-1} \right] \\
&\qquad = \begin{cases} \mathbb{P}(u_t \le s \mid \mathcal{F}_{t-1}) - \mathbb{P}(u_t \le 0 \mid \mathcal{F}_{t-1}) & \text{if } s \ge 0 \\ \mathbb{P}(u_t \le 0 \mid \mathcal{F}_{t-1}) - \mathbb{P}(u_t \le s \mid \mathcal{F}_{t-1}) & \text{if } s < 0 \end{cases} \\
&\qquad = \left| \mathbb{P}(u_t \le s \mid \mathcal{F}_{t-1}) - \mathbb{P}(u_t \le 0 \mid \mathcal{F}_{t-1}) \right|,
\end{aligned}$$

we can see that

$$\begin{aligned}
\mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1} \right] &\le \left| \frac{1}{\sqrt{T}} b' X_t \right| \int_0^{\frac{1}{\sqrt{T}} b' X_t} \left| F_{t-1}(X_t' \beta(\tau) + s \mid X_t) - F_{t-1}(X_t' \beta(\tau) \mid X_t) \right| ds \\
&\le \left| \frac{1}{\sqrt{T}} b' X_t \right| \int_0^{\frac{1}{\sqrt{T}} b' X_t} f_{t-1}(X_t' \beta(\tau) \mid X_t) s ds \\
&\quad + \left| \frac{1}{\sqrt{T}} b' X_t \right| \int_0^{\frac{1}{\sqrt{T}} b' X_t} \left| \frac{F_{t-1}(X_t' \beta(\tau) + s \mid X_t) - F_{t-1}(X_t' \beta(\tau) \mid X_t)}{s} - f_{t-1}(X_t' \beta(\tau) \mid X_t) \right| s ds.
\end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1} \right] &\le \left(\max_{1 \le t \le T} \left| \frac{1}{\sqrt{T}} b' X_t \right| \right) \\
&\quad \times \left[b' \left(\frac{1}{T} \sum_{t=1}^{T} f_{t-1}(X_t' \beta(\tau) \mid X_t) X_t X_t' \right) b + o_p(1) \right].
\end{aligned}$$

We know that

$$\begin{aligned}
b' \left(\frac{1}{T} \sum_{t=1}^{T} f_{t-1}(X_t' \beta(\tau) \mid X_t) X_t X_t' \right) b &= O_p(1), \\
\max_{1 \le t \le T} \left| \frac{1}{\sqrt{T}} b' X_t \right| &= o_p(1),
\end{aligned}$$

so

$$\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right] \xrightarrow{p} 0.$$

Denote

$$\zeta_T = \sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right]$$

for any $T \in N_+$, and note that

$$\zeta_{Tt}(b)^2 \leq 4\frac{1}{T} b' X_t X_t' b$$

for each $1 \leq t \leq T$ and therefore that

$$\zeta_T \leq b' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right) b.$$

It follows that

$$|\zeta_T| \leq |b| \left(\frac{1}{T}\sum_{t=1}^{T} |X_t|^2\right)$$

and

$$|\zeta_T|^2 \leq |b|^2 \frac{1}{T^2} \left(\sum_{t=1}^{T} |X_t|^2\right)^2 \leq \frac{|b|^2}{T} \sum_{t=1}^{T} |X_t|^4,$$

where we used the fact that $\left(\sum_{i=1}^{n} a_i\right)^2 \leq n \sum_{i=1}^{n} a_i^2$ for any non-negative reals $a_1, \cdots, a_n$. Therefore,

$$\mathbb{E}|\zeta_T|^2 \leq |b|^2 \sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^4 < +\infty,$$

where the final inequality follows from the assumption of bounded fourth moments.

Now, for any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right]\right] &= \mathbb{E}\left[\zeta_T\right] \\
&= \mathbb{E}\left[\zeta_T \cdot I_{\{\zeta_T > \epsilon\}}\right] + \mathbb{E}\left[\zeta_T \cdot I_{\{\zeta_T \leq \epsilon\}}\right] \\
&\leq \left(\mathbb{E}\left[\zeta_T^2\right]\right)^{\frac{1}{2}} \left[\mathbb{P}\left(\zeta_T > \epsilon\right)\right]^{\frac{1}{2}} + \epsilon \\
&\leq |b| \left(\sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^4\right)^{\frac{1}{2}} \left[\mathbb{P}\left(\zeta_T > \epsilon\right)\right]^{\frac{1}{2}} + \epsilon,
\end{aligned}$$

where the first inequality follows from Hölder's inequality. Since $\zeta_T \xrightarrow{p} 0$, it follows that

$$\limsup_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right]\right] \leq \epsilon,$$

and since this holds for any $\epsilon > 0$,

$$\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}\left[\zeta_{Tt}(b)^2 \mid \mathcal{F}_{t-1}\right]\right] = 0.$$

By implication,

$$\lim_{T\to\infty} \mathbb{E}\left|\sum_{t=1}^{T} W_{Tt}(b)\right|^2 = 0$$

and, since L^2 convergence implies convergence in probability,

$$\sum_{t=1}^{T} W_{Tt}(b) \xrightarrow{p} 0.$$

It follows that

$$\sum_{t=1}^{T} \zeta_{Tt}(b) \xrightarrow{p} \frac{1}{2} b' \Sigma_f(\tau) b.$$

Putting all the results together, Slutsky's theorem implies that

$$V_T(b) = -b'\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t\left(\tau - I_{(-\infty,0)}\circ u_t\right)\right] + \sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b'X_t}\left(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\right)ds$$

$$\xrightarrow{d} -b'\mathcal{Z} + \frac{1}{2}b'\Sigma_f(\tau)b,$$

where $\mathcal{Z} \sim \mathcal{N}[O_{k\times 1}, \tau(1-\tau)\Sigma_{xx}]$. While we have shown this result for a single $b \in \mathbb{Q}^k$, this can easily be generalized to an arbitrary collection $b_1, \cdots, b_n$ of finite distinct elements in $\mathbb{Q}^k$, since

$$\begin{pmatrix} V_T(b_1) \\ \vdots \\ V_T(b_n) \end{pmatrix} = -\begin{pmatrix} b_1' \\ \vdots \\ b_n' \end{pmatrix}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t\left(\tau - I_{(-\infty,0)}\circ u_t\right)\right]$$

$$+\begin{pmatrix} \sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b_1'X_t}\left(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\right)ds \\ \vdots \\ \sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b_n'X_t}\left(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\right)ds \end{pmatrix}.$$

We saw above that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t\left(\tau - I_{(-\infty,0)}\circ u_t\right) \xrightarrow{d} \mathcal{Z},$$

and that

$$\sum_{t=1}^{T}\int_0^{\frac{1}{\sqrt{T}}b_i'X_t}\left(I_{(-\infty,s]}\circ u_t - I_{(-\infty,0]}\circ u_t\right)ds \xrightarrow{p} \frac{1}{2}b_i'\Sigma_f(\tau)b_i$$

for any $1 \le i \le n$. It follows from Slutsky's theorem that

$$\begin{pmatrix} V_T(b_1) \\ \vdots \\ V_T(b_n) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} -b_1'\mathcal{Z} + \frac{1}{2}b_1'\Sigma_f(\tau)b_1 \\ \vdots \\ -b_n'\mathcal{Z} + \frac{1}{2}b_n'\Sigma_f(\tau)b_n \end{pmatrix}.$$

Defining

$$V_0(b) = -b'\mathcal{Z} + \frac{1}{2}b'\Sigma_f(\tau)b$$

for any $b \in \mathbb{R}^k$, it follows that the sequence of random sequences $\{\{V_T(b)\}_{b\in\mathbb{Q}^k}\}_{T\in N_+}$ converges weakly to the random sequence $\{V_0(b)\}_{b\in\mathbb{Q}^k}$, where $\{V_0(b)\}_{b\in\mathbb{Q}^k}$ takes values in F_0 because each $\{V_T(b)\}_{b\in\mathbb{Q}^k}$ does.

All we need to show now to apply the basic convergence theorem is that $\{V_0(b)\}_{b\in\mathbb{Q}^k}$

takes values in some Borel measurable subset F^* of $\mathbb{R}^\infty$ such that each $y \in F^*$ corresponds to a convex function with a unique minimizer. To this end, we define

$$F^* = \{y \in \mathbb{R}^\infty \mid \exists z \in \mathbb{R}^k \quad \text{s.t.} \quad y_i = -b_i' z + \frac{1}{2} b_i' \Sigma_f(\tau) b_i \quad \forall i\},$$

where we enumerate $\mathbb{Q}^k$ as $\{b_i\}_{i \in N_+}$. Clearly, $F^* \subset F_0$ because each $y \in F^*$ corresponds to a convex function on $\mathbb{R}^k$ defined as $b \mapsto -b'z + \frac{1}{2} b' \Sigma_f(\tau) b$ for some $z \in \mathbb{R}^k$, and this function has a unique minimizer in $\Sigma_f(\tau)^{-1} z$. We will be done if we can show that F^* is $\mathcal{B}(\mathbb{R}^\infty, \tau^\infty)$-measurable. We will actually show that F^* is a closed subset of $\mathbb{R}^\infty$.

To this end, assume without loss of generality that $\{b_i\}_{i \in N_+}$ is ordered so that the first k vectors are linearly independent. Now let $\{y^n\}_{n \in N_+}$ be a sequence in F^* that converges to some $y \in \mathbb{R}^\infty$. We need to show that $y \in F^*$. By definition, for each $n \in N_+$ there exists an $z_n \in \mathbb{R}^k$ such that

$$y_i^n = -b_i' z_n + \frac{1}{2} b_i' \Sigma_f(\tau) b_i \quad \forall i.$$

It follows that

$$\underbrace{\begin{pmatrix} b_1' \\ \vdots \\ b_k' \end{pmatrix}}_{B} z_n = \frac{1}{2} \begin{pmatrix} b_1' \Sigma_f(\tau) b_1 \\ \vdots \\ b_k' \Sigma_f(\tau) b_k \end{pmatrix} + \begin{pmatrix} y_1^n \\ \vdots \\ y_k^n \end{pmatrix}$$

for any $n \in N_+$. The $k \times k$ matrix B is non-singular by assumption, so it follows that

$$z_n = B^{-1} \left[\frac{1}{2} \begin{pmatrix} b_1' \Sigma_f(\tau) b_1 \\ \vdots \\ b_k' \Sigma_f(\tau) b_k \end{pmatrix} + \begin{pmatrix} y_1^n \\ \vdots \\ y_k^n \end{pmatrix} \right] \to z := B^{-1} \left[\frac{1}{2} \begin{pmatrix} b_1' \Sigma_f(\tau) b_1 \\ \vdots \\ b_k' \Sigma_f(\tau) b_k \end{pmatrix} + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right]$$

as $n \to \infty$. Now choose any $i \in N_+$: taking $n \to \infty$ on both sides of the equation

$$y_i^n = -b_i' z_n + \frac{1}{2} b_i' \Sigma_f(\tau) b_i$$

shows us that

$$y_i = -b_i' z + \frac{1}{2} b_i' \Sigma_f(\tau) b_i,$$

and as such that $y \in F^*$.

It now follows immediately from the basic convergence theorem that

$$\sqrt{T}\left[\hat{\beta}_T(\tau) - \beta(\tau)\right] \xrightarrow{d} \Sigma_f(\tau)^{-1}\mathcal{Z} \sim \mathcal{N}\left[O_{k\times 1},\ \tau(1-\tau)\Sigma_f(\tau)^{-1}\Sigma_{xx}\Sigma_f(\tau)^{-1}\right],$$

since $\Sigma_f(\tau)^{-1}\mathcal{Z}$ is the unique solution to the problem

$$\min_{b\in\mathbb{R}^k} \quad V_0(b) = -b'\mathcal{Z} + \frac{1}{2}b'\Sigma_f(\tau)b.$$

Q.E.D.

Chapter 4

The Asymptotics of Bayesian Estimation

In the previous chapter, we focused on the consistency and asymptotic normality of a wide class of extremum estimators, including GMM and ML estimators. The asymptotic results in the last chapter were approached from a frequentist perspective, with the population parameter assumed to be fixed but unknown. In a Bayesian approach, the parameter is treated as a random variable with its own distribution, and estimators are drawn from the posterior distribution of the parameter updated using the data.

In this chapter we show that this Bayesian approach not only possesses desirable finite sample properties (recall the complete class theorem), but also that it is consistent and asymptotically normal from a frequentist perspective as well. These are remarkable results showing that, no matter what prior we choose, the estimators drawn from the posterior distribution are well-behaved in the limit. From this perspective, the prior distributions on the parameters of interest can be viewed as finite sample corrections that help improve the finite sample properties of estimators without affecting their asymptotics.

4.1 Existence of the Posterior

Since parameters are treated as random variables in Bayesian estimation, we first lay out the basic setting in which the results below will be stated. In Bayesian analysis, all estimators, regardless of whether they are point or set estimators, are constructed using the posterior distribution of the parameter. As such, we need only focus on the posterior distribution instead of working with different settings depending on the type of the estimator under consideration, as we do in a frequentist setting.

As in standard frequentist decision theory, we start by specifying the statistical model, that is, the distribution of the data given the parameter. In addition, Bayesian models require the specification of a prior distribution, which summarizes the information the statistician holds on the parameter of interest. The model, which we call the likelihood, and the prior can be combined to yield the joint distribution over the data and the parameter. If we assume the

existence of a σ-finite measure with respect to which the likelihood admits a measurable density function, then constructing the posterior distribution of the parameter given the data based on their joint distribution proves straightforward. However, in the general case where no such dominating measure exists, the posterior distribution exists only for parameter spaces that are also Polish spaces, and the proof of existence relies on the disintegration theorem.

We start by investigating the case where there exists a dominating measure, and then move onto the more general case. The following objects are common to both approaches.

Suppose $(E, \mathcal{E})$ is some measurable space and $\{X_t\}_{t \in N_+}$ a stochastic process taking values in $(E, \mathcal{E})$. Let (Θ, ϑ) be the parameter space, and denote by $\bar{\Theta}$ the parameter as a random variable taking values on (Θ, ϑ). For each $T \in N_+$, the T-sample statistical model is given as a transition probability kernel μ_T from $(\Theta, \mathcal{A})$ into $(E, \mathcal{E})^T$ such that

$$\mu_T(\cdot, A) \circ \bar{\Theta} = \mathbb{P}\left(X^T \in A \mid \bar{\Theta}\right)$$

for any $A \in \mathcal{E}^T$. That is, $\mu_T(\theta, \cdot)$ is the distribution of $X^T = (X_1, \cdots, X_T)$ given that the parameter takes the value $\theta \in \Theta$.

We let P be the unconditional distribution of $\bar{\Theta}$. P summarizes the statistician's prior information about the parameter, and is accordingly referred to as the prior distribution. The joint distribution $J_{X^T \times \bar{\Theta}}$ of X^T and $\bar{\Theta}$ is determined as

$$J_{X^T \times \bar{\Theta}}(A \times B) = \int_B \mu_T(\theta, A) dP(\theta)$$

for any $A \in \mathcal{E}^T$ and $B \in \vartheta$.

The marginal distribution m_T of X^T is then defined as

$$M_T(A) = J_{X^T \times \bar{\Theta}}(A \times \Theta) = \int_\Theta \mu_T(\theta, A) dP(\theta) \quad \forall A \in \mathcal{E}^T.$$

4.1.1 The Posterior Distribution under Domination

Assume for the time being that there exists a σ-finite measure λ_T on $(E, \mathcal{E})^T$ and a jointly measurable function $f_T : E^T \times \Theta \to \mathbb{R}_+$ such that

$$\mu_T(\theta, A) = \int_A f_T(x \mid \theta) d\lambda_T(x) \quad \forall A \in \mathcal{E}^T.$$

In other words, $f_T(\cdot \mid \theta)$ is the likelihood of X^T given that the parameter takes the value $\theta \in \Theta$.

In this case, the marginal distribution of X^T can be written as

$$\begin{aligned} M_T(A) &= \int_\Theta \mu_T(\theta, A) dP(\theta) \\ &= \int_\Theta \int_A f_T(x \mid \theta) d\lambda_T(x) dP(\theta) \end{aligned}$$

$$= \int_A \left(\int_\Theta f_T(x \mid \theta) dP(\theta) \right) d\lambda_T(x)$$

for any $A \in \mathcal{E}^T$. Defining the function $m_T : E^T \to \mathbb{R}_+$ as

$$m_T(x) = \int_\Theta f_T(x \mid \theta) dP(\theta) \quad \forall x \in E^T,$$

m_T is a $\mathcal{E}^T$-measurable non-negative function such that

$$m_T \in \frac{\partial M_T}{\partial \lambda_T}.$$

We call m_T the T-sample marginal likelihood. Heuristically, $m_T(x)$ can be interpreted as the "likelihood" of the prior distribution P given that the data $x \in E^T$ is realized. As such, it is often used as a means of prior selection; more on this later. For the time being, note that

$$\int_{E^T} m_T(x) d\lambda_T(x) = 1,$$

so that $m_T \in \mathbb{R}$ λ_T-almost everywhere and thus M_T-almost everywhere.

Now we can directly construct the posterior distribution of $\bar{\Theta}$ given X^T as follows:

Theorem 4.1.1 (Posterior Distribution under Domination)**.** *Define* $\Pi_T : E \times \vartheta \to [0,1]$ *as*

$$\Pi_T(x, B) = \begin{cases} \frac{\int_B f_T(x|\theta) dP(\theta)}{m_T(x)} & \text{if } m_T(x) \in (0, +\infty) \\ P(B) & \text{otherwise} \end{cases}$$

for any $(x, B) \in E \times \vartheta$*. Then,* Π_T *is a transition probability kernel from* $(E, \mathcal{E})^T$ *into* (Θ, ϑ) *such that*

$$\Pi_T(\cdot, B) \circ X^T = \mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right) \quad \forall B \in \vartheta.$$

Proof. For any $x \in E^T$, $\Pi_T(x, \cdot)$ is a measure on (Θ, ϑ) such that $\Pi_T(x, \Theta) = 1$. Furthermore, for any $B \in \vartheta$,

$$\Pi_T(\cdot, B) = I_{\{m_T \in (0,+\infty)\}} \frac{\int_B f_T(\cdot \mid \theta) dP(\theta)}{m_T} + P(B) \cdot I_{\{m_T \notin (0,+\infty)\}}$$

is a non-negative $\mathcal{E}^T$-measurable function. By definition, Π_T is a transition probability kernel from $(E, \mathcal{E})^T$ into (Θ, ϑ).

We now want to show that

$$\Pi_T(\cdot, B) \circ X^T = \mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right) \quad \forall B \in \vartheta$$

To this end, fix $B \in \vartheta$ and choose any $H \in \sigma X^T$. Letting $H = (X^T)^{-1}(A)$ for some $A \in \mathcal{E}^T$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(\Pi_T(\cdot, B) \circ X^T\right) I_H\right] &= \int_A \Pi_T(x, B) dM_T(x) \\
&= \int_A \Pi_T(x, B) m_T(x) d\lambda_T(x) \\
&= \int_{A \cap \{m_T \in (0,+\infty)\}} \Pi_T(x, B) m_T(x) d\lambda_T(x) + \int_{A \cap \{m_T = 0\}} P(B) m_T(x) d\lambda_T(x) \\
&= \int_{A \cap \{m_T \in (0,+\infty)\}} \int_B f_T(x \mid \theta) dP(\theta) d\lambda_T(x) \\
&= \int_B \left(\int_{A \cap \{m_T \in (0,+\infty)\}} f_T(x \mid \theta) d\lambda_T(x) \right) dP(\theta) \\
&= \int_B \mu_T(\theta, A \cap \{m_T \in (0, +\infty)\}) dP(\theta).
\end{aligned}$$

Since $\lambda_T(\{m_T = +\infty\}) = 0$ and $\mu_T(\theta, \cdot) << \lambda_T$ for any $\theta \in \Theta$, it follows that

$$\mu_T(\theta, \{m_T = +\infty\}) = 0 \quad \forall \theta \in \Theta.$$

As such,

$$\mathbb{E}\left[\left(\Pi_T(\cdot, B) \circ X^T\right) I_H\right] = \int_B \mu_T(\theta, A \cap \{m_T > 0\}) dP(\theta).$$

Finally, because

$$0 \leq \int_B f_T(x \mid \theta) dP(\theta) \leq m_T(x) \quad \forall x \in E^T,$$

we can see that $\int_B f_T(x \mid \theta) dP(\theta) = 0$ whenever $m_T(x) = 0$ and thus

$$\begin{aligned}
0 \leq \int_B \mu_T(\theta, A \cap \{m_T = 0\}) dP(\theta) &= \int_B \int_{A \cap \{m_T = 0\}} f_T(x \mid \theta) d\lambda_T(x) dP(\theta) \\
&= \int_{A \cap \{m_T = 0\}} \left(\int_B f_T(x \mid \theta) dP(\theta) \right) d\lambda_T(x) \leq 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left(\Pi_T(\cdot, B) \circ X^T\right) I_H\right] &= \int_B \mu_T(\theta, A) dP(\theta) \\
&= J_{X^T \times \bar{\Theta}}(A \times B) = \mathbb{P}\left(X^T \in A,\, \bar{\Theta} \in B\right) = \mathbb{E}\left[\mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right) \cdot I_H\right].
\end{aligned}$$

Since both $\Pi_T(\cdot, B) \circ X^T$ and $\mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right)$ are σX^T-measurable non-negative valued func-

tions,

$$\Pi_T(\cdot, B) \circ X^T = \mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right).$$

□

Note the similarity of our construction of Π_T to the traditional Bayes' rule formula. In fact, when there exists a σ-finite measure on (Θ, ϑ) with respect to which P is dominated, then the usual Bayes' rule yields the posterior density of $\bar{\Theta}$ given X^T.

4.1.2 The Posterior Distribution in the General Case

Now we drop the assumption that a σ-finite measure such as λ_T above exists. The existence of the posterior distribution in this more general case hinges on an important result concerning the existence of conditional distributions for random variables taking values in certain measurable spaces. First, we state a deep measure theoretic result without proof:

Theorem 4.1.2 (Kuratowski's Theorem)**.** *Suppose $(X, \mathcal{X})$ is a Polish space, that is, there exists a metric d on X such that (X, d) is a complete and separable metric space and $\mathcal{X}$ is the Borel σ-algebra on X induced by d. If X is uncountable, then there exists a bijection $g : X \to \mathbb{R}$ such that $g \in \mathcal{X}/\mathcal{B}(\mathbb{R})$ and $g^{-1} \in \mathcal{B}(\mathbb{R})/\mathcal{X}$.*

This theorem implies, in particular, that any uncountable Polish space has the same cardinality as the real line $\mathbb{R}$. It is especially powerful because any euclidean space equipped with its Borel σ-algebra is a Polish space, as well as any of its compact subspaces.

Kuratowski's theorem can be used to derive the following result:

Theorem 4.1.3 (Existence of Conditional Distribution for Polish Spaces)**.** *Let $(F, \mathcal{F}, \mu)$ be some probability space and $(X, \mathcal{X})$ an uncountable Polish space. Suppose $\mathbf{x} : F \to X$ is measurable relative to $\mathcal{F}$ and $\mathcal{X}$. Then, for any sub σ-algebra $\mathcal{F}'$ of $\mathcal{F}$, there exists a transition probability kernel K from $(F, \mathcal{F}')$ into $(X, \mathcal{X})$ such that*

$$\int_H K(\cdot, A) d\mu = \int_H (I_A \circ \mathbf{x})\, d\mu$$

for any $A \in \mathcal{X}$ and $H \in \mathcal{F}'$.

Proof. We first show that the result holds for $(X, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Using the notation for generalized conditional expectations introduced in the section on the sufficiency principle, for each

$q \in \mathbb{Q}$ we define C_q as the conditional expectation of the non-negative $\mathcal{F}$-measurable function $I_{(-\infty,q]} \circ \mathbf{x}$ with respect to $\mathcal{F}'$:

$$C_q = \mu(I_{(-\infty,q]} \circ \mathbf{x} \mid \mathcal{F}').$$

Note that

$$0 \leq C_q \leq C_r \leq 1 \quad \mu\text{-a.e.}$$

for any $q, r \in \mathbb{Q}$ such that $q \leq r$ by the monotonicity of conditional expectations. Furthermore, defining the set

$$F_{q,r} = \{y \in F \mid 0 \leq C_q(y) \leq C_r(y) \leq 1\},$$

we can see that $F_{q,r} \in \mathcal{F}'$ because C_q and C_r are $\mathcal{F}'$-measurable non-negative functions.

Define the sequence $\{f_n\}_{n \in N_+}$ of non-negative $\mathcal{F}$-measurable functions as

$$f_n(y) = I_{(-\infty,n]}(\mathbf{x}(y'))$$

for any $n \in N_+$ and $y \in F$. Then,

$$f_n \uparrow 1 \quad \text{pointwise as } n \to \infty.$$

By the monotone convergence theorem for conditional expectations,

$$C_n = \mu(f_n \mid \mathcal{F}') \uparrow 1 \quad \text{as } n \to \infty \ \mu\text{-a.e.}$$

Similarly,

$$C_{-n} \downarrow 0 \quad \text{as } n \to \infty \ \mu\text{-a.e.}$$

by the dominated convergence theorem for conditional expectations. Defining the sets

$$F_{+\infty} = \{y \in F \mid \sup_{n \in N_+} C_n(y) = 1\}, \quad F_{-\infty} = \{y \in F \mid \inf_{n \in N_+} C_{-n}(y) = 0\},$$

$F_{+\infty}$ and $F_{-\infty}$ are also $\mathcal{F}'$-measurable sets that have μ-measure 0.

Now define

$$F_0 = \left(\bigcap_{q,r \in \mathbb{Q}, q \leq r} F_{q,r} \right) \cap F_{+\infty} \cap F_{-\infty}.$$

Then,

$$F_0 \in \mathcal{F}' \quad \text{and} \quad \mu(F_0) = 1.$$

For each $y \in F_0$, we define the function $D_y : \mathbb{R} \to [0,1]$ as

$$D_y(r) = \inf\{C_q(y) \mid q \in \mathbb{Q}, q > r\}.$$

D_y possesses the following properties:

i) **Increasing on $\mathbb{R}$**

For any $r_1, r_2 \in \mathbb{R}$ such that $r_1 < r_2$, there exists some $q \in \mathbb{Q}$ such that $r_1 < q < r_2$. For any $q' \in \mathbb{Q}$ such that $q' > r_2$, we also have $q' > q$, so that $C_{q'}(y) \geq C_q(y)$ and therefore

$$D_y(r_2) = \inf\{C_{q'}(y) \mid q' > r_2, q' \in \mathbb{Q}\} \geq C_q(y).$$

By definition, $C_q(y) \geq D_y(r_1)$, so we have $D_y(r_1) \leq D_y(r_2)$.

ii) **Right Continuous on $\mathbb{R}$**

For any $r \in \mathbb{R}$ and a sequence $\{r_n\}_{n \in N_+}$ in $\mathbb{R}$ such that $r_n \downarrow r$, we want to show that $D_y(r_n) \downarrow D_y(r)$ as $n \to \infty$. To this end, note first that $D_y(r) \leq D_y(r_n)$ for any $n \in N_+$, so that

$$\inf_{n \in N_+} D_y(r_n) \geq D_y(r).$$

Now suppose, for the sake of contradiction, that

$$\inf_{n \in N_+} D_y(r_n) > D_y(r).$$

Then, there exists some $q \in \mathbb{Q}$, $q > r$ such that

$$D_y(r) \leq C_q(y) < \inf_{n \in N_+} D_y(r_n).$$

By our choice of r_n, there exists a $N \in N_+$ such that $r \leq r_n < q$ for any $n \in N_+$, and this implies that

$$D_y(r_n) \leq C_q(y) \quad \forall n \in N_+,$$

and since $\inf_{n \in N_+} D_y(r_n) \leq D_y(r_{n'})$ for any $n' \in N_+$ by the fact that $D_y(\cdot)$ is an increasing function, we have

$$\inf_{n \in N_+} D_y(r_n) \leq C_q(y),$$

a contradiction. It must therefore be the case that

$$\inf_{n \in N_+} D_y(r_n) = D_y(r).$$

iii) **Limits as** $r \to \pm\infty$

Suppose that

$$\sup_{r\in\mathbb{R}} D_y(r) < 1.$$

Since

$$\sup_{n\in N_+} C_n(y) = 1$$

by the way we designed F_0, it follows that, there exists an $N \in N_+$ such that

$$\sup_{r\in\mathbb{R}} D_y(r) < C_n(y) \leq 1 \quad \forall n \geq N.$$

Since $C_n(y) = D_y(n)$ for any $n \in N_+$, this implies that

$$\sup_{r\in\mathbb{R}} D_y(r) < D_y(n) \quad \forall n \geq N,$$

a contradiction. As such,

$$\sup_{r\in\mathbb{R}} D_y(r) = 1.$$

Similarly, we can show that

$$\inf_{r\in\mathbb{R}} D_y(r) = 0.$$

We have shown that D_y is a well-defined distribution function on $\mathbb{R}$, so there exists a unique probability measure v_y on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that D_y is the distribution function corresponding to v_y:

$$v_y((-\infty, r]) = D_y(r) \quad \forall r \in \mathbb{R}.$$

Now define the mapping $K : F \times \mathcal{X} \to [0,1]$ as

$$K(y, A) = I_{F_0}(y) v_y(A) + I_{F_0^c}(y)\delta_0(A) \quad \forall (y, A) \in F \times \mathcal{X},$$

where δ_0 is the Dirac measure that assigns a point mass to 0. We can easily show that K is a transition probability kernel from $(F, \mathcal{F}')$ into $(X, \mathcal{X})$:

i) Fix any $y \in F$. Since v_y and δ_0 are both probability measures, it follows that $K(y, \cdot)$ is a probability measure on $(X, \mathcal{X})$.

ii) Fix any $A \in \mathcal{X}$. Then, for any $r \in \mathbb{R}$,

$$\begin{aligned} K(y,(-\infty,r]) &= I_{F_0}(y) D_y(r) + I_{F_0^c}(y)\delta_0((-\infty,r]) \\ &= I_{F_0}(y)\left(\inf\{C_q(y) \mid q \in \mathbb{Q}, q > r\}\right) + I_{F_0^c}(y)\delta_0((-\infty,r]), \end{aligned}$$

so that

$$K(\cdot,(-\infty,r]) = I_{F_0}\left(\inf_{q\in\mathbb{Q},q>r} C_q\right) + I_{F_0^c}\delta_0((-\infty,r]).$$

The indicator functions are trivially $\mathcal{F}'$-measurable non-negative functions, and the infimum of a countable collection of $\mathcal{F}'$-measurable non-negative functions is itself a $\mathcal{F}'$-measurable non-negative function, so $K(\cdot,(-\infty,r]) \in \mathcal{F}'_+$.

Now define the collection $\mathcal{D}$ of $\mathcal{X}$-measurable sets as

$$\mathcal{D} = \{A \in \mathcal{X} \mid K(\cdot,A) \in \mathcal{F}'_+\}.$$

We just showed that the π-system $\{(-\infty,r] \mid r \in \mathbb{R}\}$ that generates $\mathcal{B}(\mathbb{R})$ is contained in $\mathcal{D}$. Thus, to conclude that $K(\cdot,A) \in \mathcal{F}'_+$ for any $A \in \mathcal{X}$, we need only show that $\mathcal{D}$ is a λ-system on X. This is again clear:

- $X \in \mathcal{D}$ because $K(\cdot,X) = 1$.

- For any $A, B \in \mathcal{D}$ such that $A \subset B$,

$$\begin{aligned} K(y,B\setminus A) &= I_{F_0}(y) v_y(B\setminus A) + I_{F_0^c}(y)\delta_0(B\setminus A) \\ &= I_{F_0}(y)\left[v_y(B) - v_y(A)\right] + I_{F_0^c}(y)\left[\delta_0(B) - \delta_0(A)\right] \quad \text{(Finite Additivity)} \\ &= K(y,B) - K(y,A) \end{aligned}$$

  for any $y \in F$, so that

$$K(\cdot,B\setminus A) = K(\cdot,B) - K(\cdot,A) \in \mathcal{F}'_+.$$

- For any $\{A_n\}_{n\in N_+} \subset \mathcal{D}$ such that $A_n \nearrow A$,

$$\begin{aligned} K(y,A) &= I_{F_0}(y)\left(\lim_{n\to\infty} v_y(A_n)\right) + I_{F_0^c}(y)\left(\lim_{n\to\infty}\delta_0(A_n)\right) \quad \text{(Sequential Continuity)} \\ &= \lim_{n\to\infty}\left(I_{F_0}(y)v_y(A_n) + I_{F_0^c}(y)\delta_0(A_n)\right) = \lim_{n\to\infty} K(y,A_n) \end{aligned}$$

  for any $y \in F$, and since measurability is preserved across pointwise limits, $K(\cdot,A) \in \mathcal{F}'_+$.

This proves that $\mathcal{D}$ is a λ-system that contains a π-system generating $\mathcal{B}(\mathbb{R})$, and as such that

$$K(\cdot, A) \in \mathcal{F}'_+ \quad \forall A \in \mathcal{X} = \mathcal{B}(\mathbb{R}).$$

We have therefore shown that K is a transition probability kernel from $(F, \mathcal{F}')$ into $(X, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

It remains to show that the equation

$$\int_H K(\cdot, A) d\mu = \int_H (I_A \circ \mathbf{x}) d\mu$$

holds for any $A \in \mathcal{X}$ and $H \in \mathcal{F}'$. Again, we use a $\pi - \lambda$ construction. Fix $H \in \mathcal{F}'$ and define the collection

$$\mathcal{D} = \Big\{ A \in \mathcal{X} \mid \int_H K(\cdot, A) d\mu = \int_H (I_A \circ \mathbf{x}) d\mu \Big\}.$$

For each $q \in \mathbb{Q}$, we saw above that

$$K(\cdot, (-\infty, q]) = I_{F_0} C_q + I_{F_0^c} \delta_0((-\infty, r]),$$

so that

$$\int_H K(\cdot, (-\infty, q]) d\mu = \int_{H \cap F_0} C_q d\mu,$$

where we used the fact that $\mu(F_0^c) = 0$. For each $q \in \mathbb{Q}$,

$$\begin{aligned} \int_{H \cap F_0} C_q d\mu &= \int_{H \cap F_0} \mu(I_{(-\infty, q]} \circ \mathbf{x} \mid \mathcal{F}') d\mu \\ &= \int_{H \cap F_0} \left(I_{(-\infty, q]} \circ \mathbf{x} \right) d\mu \end{aligned}$$

by the definition of conditional expectations and the fact that $H \cap F_0 \in \mathcal{F}'$. Therefore,

$$\int_H K(\cdot, (-\infty, q]) d\mu = \int_H \left(I_{(-\infty, q]} \circ \mathbf{x} \right) d\mu,$$

which shows us that $(-\infty, q] \in \mathcal{D}$; the π-system $\{(-\infty, q] \mid q \in \mathbb{Q}\}$ that generates $\mathcal{B}(\mathbb{R})$ is contained in $\mathcal{D}$.

$\mathcal{D}$ is also a λ-system on X:

- $X \in \mathcal{D}$ because

$$\int_H K(\cdot, X) d\mu = \mu(H) = \int_H (I_X \circ \mathbf{x}) \, d\mu.$$

- For any $A, B \in \mathcal{D}$ such that $A \subset B$,

$$\begin{aligned}\int_H K(\cdot, B \setminus A) d\mu &= \int_H K(\cdot, B) d\mu - \int_H K(\cdot, A) d\mu \\ &= \int_H (I_B \circ \mathbf{x})\, d\mu - \int_H (I_A \circ \mathbf{x}) d\mu \\ &= \int_H \left(I_{B \setminus A} \circ \mathbf{x}\right) d\mu,\end{aligned}$$

  so that $B \setminus A \in \mathcal{D}$.

- For any $\{A_n\}_{n \in N_+} \subset \mathcal{D}$ such that $A_n \nearrow A$,

$$\begin{aligned}\int_H K(\cdot, A) d\mu &= \int_H \left(\lim_{n\to\infty} K(\cdot, A_n)\right) d\mu \\ &= \lim_{n\to\infty} \int_H K(\cdot, A_n) d\mu && \text{(MCT)} \\ &= \lim_{n\to\infty} \int_H (I_{A_n} \circ \mathbf{x})\, d\mu \\ &= \int_H \left(\sup_{n \in N_+} (I_{A_n} \circ \mathbf{x})\right) d\mu && \text{(MCT)} \\ &= \int_H (I_A \circ \mathbf{x})\, d\mu.\end{aligned}$$

  This proves that $A \in \mathcal{D}$.

Therefore, by the $\pi - \lambda$ theorem, $\mathcal{B}(\mathbb{R}) \subset \mathcal{D}$, that is,

$$\int_H K(\cdot, A) d\mu = \int_H (I_A \circ \mathbf{x}) d\mu \quad \forall A \in \mathcal{X} = \mathcal{B}(\mathbb{R}).$$

This completes the proof for the case $(X, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

It remains to address the general case where $(X, \mathcal{X})$ is a Polish space. By Kuratowski's theorem, there exists a bijection $g : X \to \mathbb{R}$ such that $g \in \mathcal{X}/\mathcal{B}(\mathbb{R})$ and $g^{-1} \in \mathcal{B}(\mathbb{R})/\mathcal{X}$. Define the new variable $\mathbf{z} : F \to \mathbb{R}$ as

$$\mathbf{z} = g \circ \mathbf{x}.$$

Then, by what we showed above, there exists a transition probability kernel L from $(F, \mathcal{F}')$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\int_H L(\cdot, B) d\mu = \int_H (I_B \circ \mathbf{z})\, d\mu \quad \forall H \in \mathcal{F}', B \in \mathcal{B}(\mathbb{R}).$$

Define $K : F \times \mathcal{X} \to [0,1]$ as

$$K(y,A) = L(y,g(A)) \quad \forall (y,A) \in F \times \mathcal{X},$$

where K is well-defined because $g(A) = (g^{-1})^{-1}(A) \in \mathcal{B}(\mathbb{R})$ by the definition of measurability. $K(\cdot,A)$ is $\mathcal{F}'$-measurable by the measurability of $L(\cdot,g(A))$ for any $A \in \mathcal{X}$, and for any $y \in F$,

$$K(y,\emptyset) = L(y,\emptyset) = 0$$

and, for any disjoint collection $\{A_n\}_{n\in N_+} \subset \mathcal{X}$,

$$\begin{aligned} K\left[y, \bigcup_n A_n\right] &= L\left[y, g\left(\bigcup_n A_n\right)\right] = L\left[y, \bigcup_n g(A_n)\right] \\ &= \sum_{n=1}^{\infty} L(y,g(A_n)) = \sum_{n=1}^{\infty} K(y,A_n). \end{aligned}$$

This proves that $K(y,\cdot)$ is a measure on $(X,\mathcal{X})$, and since

$$K(y,X) = L(y,g(X)) = L(y,\mathbb{R}) = 1$$

by the sujectivity of $g(\cdot)$, $K(y,\cdot)$ is a probability measure. This proves that K is a transition probability kernel from $(F,\mathcal{F}')$ into $(X,\mathcal{X})$.

Finally, for any $H \in \mathcal{F}'$ and $A \in \mathcal{X}$, letting $B = g(A) \in \mathcal{B}(\mathbb{R})$, we have

$$\begin{aligned} \int_H K(\cdot,A)d\mu &= \int_H L(\cdot,g(A))d\mu = \int_H \left(I_{g(A)} \circ \mathbf{z}\right) d\mu \\ &= \int_H \left(\left(I_A \circ g^{-1}\right) \circ \mathbf{z}\right) d\mu = \int_H (I_A \circ \mathbf{x}) d\mu. \end{aligned}$$

This completes the proof for the general case.

□

One of the uses of the preceding theorem is to establish the existence of a conditional distribution of a random variable Y taking values in an uncountable Polish space $(X,\mathcal{X})$ given some sub σ-algebra $\mathcal{F}$ of $\mathcal{H}$. Specifically, note that, becuase $Y : \Omega \to X$ is measurable relative to $\mathcal{H}$ and $\mathcal{X}$ by the definition of a random variable, the preceding theorem tells us that there exists some transition probability kernel K from $(\Omega,\mathcal{F})$ into $(X,\mathcal{X})$ such that

$$\mathbb{E}[K(\cdot,A) \cdot I_H] = \mathbb{E}[(I_A \circ Y) I_H] \quad \forall A \in \mathcal{X}, H \in \mathcal{F}.$$

Therefore, by definition

$$K(\cdot,A) = \mathbb{P}(Y \in A \mid \mathcal{F})$$

for any $A \in \mathcal{X}$. In particular, when $(E, \mathcal{E})$ is an uncountable Polish space, it implies the existence of the statistical model μ_T without the need to assume it.

The next result is the more relevant application of the theorem for our purposes. It tells us that any probability measure on a product space can be decomposed into a conditional distribution and a marginal distribution, granted that one of the constituent spaces is Polish.

Theorem 4.1.4 (Disintegration Theorem)**.** *Let $(D, \mathcal{D})$ be a measurable space, $(X, \mathcal{X})$ an uncountable Polish space, and π a probability measure on $(D, \mathcal{D}) \bigotimes (X, \mathcal{X})$. Then, there exists a transition probability kernel K from $(D, \mathcal{D})$ into $(X, \mathcal{X})$ and a probability measure μ on $(D, \mathcal{D})$ such that*

$$\pi(A \times B) = \int_A K(\cdot, B) d\mu$$

for any $A \in \mathcal{D}$ and $B \in \mathcal{X}$.

Proof. Note that $(D \times X, \mathcal{D} \bigotimes \mathcal{X}, \pi)$ is a probability space, and define the measurable functions $\mathbf{x}$ and $\mathbf{y}$ on $D \times X$ as

$$\mathbf{x}(y, x) = x, \quad \mathbf{y}(y, x) = y$$

for any $(y, x) \in D \times X$. Note that

$$\mathbf{y}^{-1}(A) = A \times X \quad \forall A \in \mathcal{D},$$

so that the σ-algebra generated by $\mathbf{y}$, which we denote by the sub σ-algebra $\mathcal{F}$ of $\mathcal{D} \bigotimes \mathcal{X}$, is given as the following collection of measurable rectangles:

$$\mathcal{F} = \{A \times X \mid A \in \mathcal{D}\}.$$

By implication, a non-negative function $f : D \times X \to \mathbb{R}_+$ is $\mathcal{F}$-measurable if and only if $f(y, x) = v(y)$ for some $\mathcal{D}$-measurable non-negative function v. Let μ be the marginal distribution induced by π as follows:

$$\mu(A) = \pi(A \times X) \quad \forall A \in \mathcal{D}.$$

Since $(X, \mathcal{X})$ is an uncountable Polish space and $\mathbf{x}$ is a measurable mapping from $D \times X$ into X, the preceding theorem now tells us that there exists some transition probability kernel L from $(D \times X, \mathcal{F})$ into $(X, \mathcal{X})$ such that

$$\int_H L(\cdot, B) d\pi = \int_H (I_B \circ \mathbf{x})\, d\pi \quad \forall H \in \mathcal{F}, B \in \mathcal{X}.$$

By the definition of $\mathcal{F}$ and the fact that $K(\cdot, B)$ is $\mathcal{F}$-measurable for each $B \in \mathcal{X}$, there exists

some function $v_B \in \mathcal{D}_+$ such that

$$L((y,x),B) = v_B(y) \quad \forall (y,x) \in D \times X.$$

Define $K : D \times \mathcal{X} \to [0,1]$ as

$$K(y,B) = v_B(y) \quad \forall (y,B) \in D \times \mathcal{X}.$$

Then, for each $B \in \mathcal{X}$, $K(\cdot,B) = v_B \in \mathcal{D}_+$, and for each $y \in D$,

$$K(y,\cdot) = L((y,x_0),\cdot)$$

is a proability measure on $(X,\mathcal{X})$, where $x_0 \in X$ is an arbitrary point in X. This shows us that K is a transition probability kernel from $(D,\mathcal{D})$ into $(X,\mathcal{X})$.

Now choose any $A \in \mathcal{D}$ and $B \in \mathcal{X}$. Since $I_B \circ \mathbf{x} = I_{D\times B}$, we can see that

$$\begin{aligned}\pi(A \times B) &= \int_{A\times X} (I_B \circ \mathbf{x})\, d\pi = \int_{A\times X} K(\cdot,B) d\pi \\ &= \int_{A\times X} v_B(y) d\pi = \int_A v_B d\mu = \int_A K(\cdot,B) d\mu.\end{aligned}$$

This completes the proof.

□

It is now easy to see how the disintegration theorem implies the existence of the posterior distribution.

Theorem 4.1.5 (Posterior Distribution in the General Case)**.** *Suppose that (Θ,ϑ) is an uncountable Polish space. Then, for each $T \in N_+$ there exists some transition probability kernel Π_T from $(E,\mathcal{E})^T$ into (Θ,ϑ) such that*

$$\int_B \mu_T(\theta,A) dP(\theta) = \int_A \Pi_T(x,B) dM_T(x)$$

for any $A \in \mathcal{E}^T$, $B \in \vartheta$, and

$$\Pi_T(\cdot,B) \circ X^T = \mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right) \quad \forall B \in \vartheta.$$

Proof. Since $J_{X^T\times\bar{\Theta}}$ is a probability measure on the product space $(E,\mathcal{E})^T \bigotimes (\Theta,\vartheta)$, and (Θ,ϑ) is an uncountable Polish space, by the disintegration theorem there exists some transition probability kernel Π_T from $(E,\mathcal{E})^T$ into (Θ,ϑ) and a probability measure λ on $(E,\mathcal{E})^T$ such that

$$J_{X^T\times\bar{\Theta}}(A \times B) = \int_A \Pi_T(\cdot,B) d\lambda$$

for any $A \in \mathcal{E}^T$ and $B \in \vartheta$. Now note that

$$M_T(A) = J_{X^T \times \bar{\Theta}}(A \times \Theta) = \int_A \Pi_T(\cdot, \Theta) d\lambda = \lambda(A)$$

for any $A \in \mathcal{E}^T$, so that $\lambda = M_T$. This proves that the first equation holds.

Now choose any $B \in \vartheta$. Then, for any $H \in \sigma X^T$, letting $H = (X^T)^{-1}(A)$ for some $A \in \mathcal{E}^T$,

$$\mathbb{E}\left[(\Pi_T(\cdot, B) \circ X^T) \cdot I_H\right] = \int_A \Pi_T(\cdot, B) dM_T = J_{X^T \times \bar{\Theta}}(A \times B)$$

$$= \mathbb{P}\left(X^T \in A,\, \bar{\Theta} \in B\right) = \mathbb{E}\left[\mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right) \cdot I_H\right].$$

By definition,

$$\Pi_T(\cdot, B) \circ X^T = \mathbb{P}\left(\bar{\Theta} \in B \mid X^T\right).$$

□

4.2 Consistency of Bayesian Estimators and Posterior Consistency

From a frequentist perspective, the minimal asymptotic criterion that an estimator must satisfy is consistency, that is, as the sample size grows, the estimator must zero in on the true parameter with increasing accuracy. In a Bayesian context, it is unclear how we would search for consistent estimators, since there is no "true parameter" in this case. The Bayesian notion of consistency instead requires that, as the sample size grows, the posterior probability that the parameter is contained in some set A converges to the truth, that is, whether the parameter actually is contained in A. Formally, this can be written as

$$\mathbb{P}\left(\bar{\Theta} \in A \mid X_1, \cdots, X_T\right) \to I_{\{\bar{\Theta} \in A\}}$$

as $T \to \infty$. This notion of consistency is referred to as posteiror consistency. That posterior consistency holds is our first major result. First proven by Doob in 1949, this theorem shows us that Bayesian estimation clears the lowest bar to be considered valid from a frequentist perspective.

Theorem 4.2.1 (Doob's Theorem)**.** *Suppose that Θ is a Borel measurable subset of $\mathbb{R}^m$ with corresponding Borel σ-algebra ϑ, and that there exists a sequence $\{h_T\}_{T \in N_+}$, where each h_T is a measurable function from E^T into Θ, such that $h_T(X^T) \xrightarrow{p} \bar{\Theta}$. Then,*

$$\mathbb{E}\left[g(\bar{\Theta}) \mid X_1, \cdots, X_T\right] \overset{a.s.}{\underset{L^1}{\to}} g(\bar{\Theta})$$

as $T \to \infty$ for any $g : \Theta \to \mathbb{R}$ such that g is ϑ-measurable and $g(\bar{\Theta})$ is integrable.

Proof. Let $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{N}\}$ be the filtration on $\mathbb{N}$ generated by the sequence $\{X_t\}_{t \in \mathbb{N}}$, where we take X_0 to be any degenerate random variable on $(E, \mathcal{E})$. As usual, define

$$\mathcal{F}_\infty = \bigvee_{t \in \mathbb{N}} \mathcal{F}_t,$$

so that $\mathcal{F}_\infty$ is the σ-algebra representing the information contained in the sample as the sample size goes to infinity.

First, we use the condition that $h_T(X^T) \xrightarrow{p} \bar{\Theta}$. By implication, there exists a subsequence of $\{h_T(X^T)\}_{T \in N_+}$ that converges almost surely to $\bar{\Theta}$; for the sake of notational simplicity, let this subsequence be $\{h_T(X^T)\}_{T \in N_+}$ itself. Define the subsets Ω_0 and Ω_1 of Ω as

$$\Omega_0 = \bigcap_{i=1}^{m} \left\{ \limsup_{T \to \infty} h_{i,T}(X^T) = \liminf_{T \to \infty} h_{i,T}(X^T) \right\}$$

$$\Omega_1 = \bigcap_{i=1}^{m} \left\{ \limsup_{T \to \infty} h_{i,T}(X^T) = \liminf_{T \to \infty} h_{i,T}(X^T) = \bar{\Theta}_i \right\},$$

where $h_{i,T}$ is the ith component of the function h_T and $\bar{\Theta}_i$ the ith component of the random vector $\bar{\Theta}$. We can immediately see that $\Omega_1 \subset \Omega_0$ and $\mathbb{P}(\Omega_1) = 1$ by assumption, so that Ω_0 is also an almost sure set.

In addition, since each $h_{i,T}(X^T)$ is $\mathcal{F}_T$-measurable and thus $\mathcal{F}_\infty$-measurable, the upper and lower limits $\limsup_{T\to\infty} h_{i,T}(X^T)$ and $\liminf_{T\to\infty} h_{i,T}(X^T)$ are both $\mathcal{F}_\infty$-measurable, so that Ω_0 is also $\mathcal{F}_\infty$-measurable.

Choosing any $\theta_0 \in \Theta$, for any $T \in N_+$ define the function $\tilde{H}_T : \Omega \to \Theta$ as

$$\tilde{H}_T = h_T(X^T) \cdot I_{\Omega_0}.$$

The $\mathcal{F}_\infty$-measurability of Ω_0 immediately implies that $\{\tilde{H}_T\}_{T\in N_+}$ is a sequence of functions that are measurable relative to $\mathcal{F}_\infty$ and ϑ. Furthermore, by design, $\{\tilde{H}_T\}_{T\in N_+}$ is convergent everywhere on Ω. Letting

$$\tilde{\Theta}(\omega) = \lim_{T\to\infty} \tilde{H}_T(\omega)$$

for any $\omega \in \Omega$, since pointwise limits preserve measurability, $\tilde{\Theta}$ is measurable relative to $\mathcal{F}_\infty$ and ϑ. Since

$$\tilde{\Theta}(\omega) = \lim_{T\to\infty} \tilde{H}_T(\omega) = \bar{\Theta}(\omega)$$

for any $\omega \in \Omega_1$, we can see that $\tilde{\Theta} = \bar{\Theta}$ almost surely.

Define

$$Z = g(\tilde{\Theta});$$

since $Z = g(\bar{\Theta})$ almost surely, and the latter is integrable, Z is an integrable real random variable. Define $\{Y_t\}_{t\in\mathbb{N}}$ as

$$Y_t = \mathbb{E}_t[Z] := \mathbb{E}[Z \mid \mathcal{F}_t]$$

for any $t \in \mathbb{N}$. Then, by the martingale convergence theorem and the characterization of uniformly integrable martingales (theorem 7.17 in the probability theory text), $\{Y_t\}_{t\in\mathbb{N}}$ is a uniformly integrable $\mathcal{F}$-martingale, and

$$Y_t \overset{a.s.}{\overset{L^1}{\to}} \mathbb{E}_\infty[Z] := \mathbb{E}[Z \mid \mathcal{F}_\infty]$$

as $t \to \infty$. By the $\mathcal{F}_\infty$-measurability of Z, it follows that

$$Y_t \overset{a.s.}{\overset{L^1}{\to}} Z$$

as $t \to \infty$. Finally, since

$$\mathbb{E}\left[g(\bar{\Theta}) \mid X^t\right] = \mathbb{E}\left[Z \mid X^t\right] = Y_t$$

almost surely for any $t \in \mathbb{N}$, and $Z = g(\bar{\Theta})$ almost surely, we have

$$\mathbb{E}\left[g(\bar{\Theta}) \mid X^t\right] \overset{a.s.}{\underset{L^1}{\to}} g(\bar{\Theta})$$

as $t \to \infty$.

□

Corollary 4.2.2 (Posterior Consistency)**.** *Suppose that Θ is a Borel measurable subset of $\mathbb{R}^m$ with corresponding Borel σ-algebra ϑ, and that there exists a sequence $\{h_T\}_{T \in N_+}$, where each h_T is a measurable function from E^T into Θ, such that $h_T(X^T) \xrightarrow{p} \bar{\Theta}$. Then,*

$$\mathbb{P}\left(\bar{\Theta} \in A \mid X_1, \cdots, X_T\right) \overset{a.s.}{\underset{L^1}{\to}} I_{\{\bar{\Theta} \in A\}}$$

as $T \to \infty$ for any $A \in \vartheta$.

Proof. For any $A \in \vartheta$, putting $g = I_A \in (\vartheta)_+$ and applying Doob's theorem yields the desired result.

□

There are two key conditions that allow the above theorem to hold. Specifically, the parameter space must be finite-dimensional, that is, a subset of euclidean space, and there must exist some sequence $\{h_T\}_{T \in N_+}$ of measurable functions such that $h_T(X^T) \xrightarrow{p} \bar{\Theta}$.

In the proof, the second condition is used to show that, up to almost sure equivalence, θ is contained in the σ-algebra generated by the sample as the sample size goes to infinity, that is, $\bar{\Theta}$ is known when the sample size goes to infinity. The finite dimensionality of the parameter space is more of a technical assumption. It ensures that the set on which pointwise convergence occurs for the sequence $\{h_T(X^T)\}_{T \in N_+}$ can be expressed in terms of its upper and lower limits and therefore is known when the sample $\{X_t\}_{t \in N_+}$ is known, instead of also requiring the information contained in the limit θ.

Of the two, the second condition is clearly the more meaningful, and as such, it serves as our identification condition for posterior consistency.

4.3 Berstein-von Mises Theorem

Here we formally show that, in the limit, Bayesian and ML methods converge to the same distribution. In other words, no matter what prior we impose to derive the posterior distribution, the information in the sample dominates that in the prior as the sample size goes to infinity. This provides a theoretical basis for using Bayesian methods to perform statistical inference in large samples.

4.3.1 Total Variation Distance

To evaluate how close Bayesian and ML mehods become in the limit, we must first furnish a method to determine the distance between these two distributions, and then show that it converges to 0 as the sample size goes to infinity. In other words, we require a metric on the space of probability measures. One possible metric we can think of is the Prohorov metric, under which convergence is equivalent to the weak convergence of probability measures.

There is another metric, called Total Variation (TV) distance, that is similar to the supremum norm on the space of continuous functions. TV distance is stronger than the Prohorov metric in the sense that convergence in TV distance ensures the pointwise convergence of probability measures on any measurable set, in contrast to the Prohorov metric, which, by the Portmanteau theorem, only ensures pointwise convergence for measurable sets whose boundary has measure 0 under the limit. Furthermore, while the Prohorov metric only induces the topology of weak convergence for probability measures defined on a separable metric space, the TV distance can be defined for probability measures on more general measurable spaces.

Formally, let $(E,\mathcal{E})$ be a measurable set and $\mathcal{M}_p(\mathcal{E})$ the set of all probability measures on $(E,\mathcal{E})$. The TV norm $\|\cdot\|_{TV} : \mathcal{M}_p(\mathcal{E}) \to [0,1]$ is defined as

$$\|\mu\|_{TV} = \sup_{A\in\mathcal{E}} \mu(A)$$

for any $\mu \in \mathcal{M}_p(\mathcal{E})$; note that this value is bounded below by 0 and above by 1 because μ is a probability measure. The TV distance is then the metric on $\mathcal{M}_p(\mathcal{E})$ induced by the TV norm[1], that is, the TV distance of $\mu, v \in \mathcal{M}_p(\mathcal{E})$ is defined as

$$\|\mu - v\|_{TV} = \sup_{A\in\mathcal{E}} |\mu(A) - v(A)| \in [0,1].$$

For probability measures on $(E,\mathcal{E})$ that are absolutely continuous with respect to some σ-finite measure λ on $(E,\mathcal{E})$, we can obtain the following characterization of the TV distance in terms of the L^1-norm:

[1]Technically speaking, the TV distance is not a metric induced by $\|\cdot\|_{TV}$ because $\mathcal{M}_p(\mathcal{E})$ is not a vector space.

Theorem 4.3.1 (L^1 Characterizations of TV Distance)**.** *Let $(E,\mathcal{E})$ be a measurable space, and $\mu, v \in \mathcal{M}_p(\mathcal{E})$ be measures that are absolutely continuous with respect to some σ-finite measure λ on $(E,\mathcal{E})$. Then, letting $\|\cdot\|_1$ be the L^1-norm on the L^1-space $L^1(\mathcal{E},\lambda)$,*

$$\|\mu - v\|_{TV} = \frac{1}{2}\left\|\frac{\partial\mu}{\partial\lambda} - \frac{\partial v}{\partial\lambda}\right\|_1 := \frac{1}{2}\int_E \left|\frac{\partial\mu}{\partial\lambda} - \frac{\partial v}{\partial\lambda}\right| d\lambda,$$

where $\frac{\partial\mu}{\partial\lambda}, \frac{\partial v}{\partial\lambda}$ are the Radon-Nikodym derivatives of μ, v with respect to λ. Furthermore,

$$\|\mu - v\|_{TV} = \sup_{A\ in\mathcal{E}} [\mu(A) - v(A)] = \int_E \left(\frac{\partial\mu}{\partial\lambda} - \frac{\partial v}{\partial\lambda}\right)^+ d\lambda.$$

Proof. By the Radon-Nikodym theorem, there exist Radon-Nikodym derivatives $f, g \in \mathcal{E}_+$ of μ, v with respect to λ. Since

$$\mu(E) = \int_E f d\lambda = 1,$$

f is contained in $L^1(\mathcal{E},\lambda)$, and likewise, $g \in L^1(\mathcal{E},\lambda)$. Define

$$B = \{x \in E \mid f(x) \geq g(x)\} \in \mathcal{E},$$

where B is measurable because it involves measurable functions f, g.

Choose any $A \in \mathcal{E}$. Then,

$$\begin{aligned}
\mu(A) - v(A) &= \int_A (f-g)\, d\lambda \\
&= \int_{A\cap B} (f-g)\, d\lambda + \int_{A\cap B^c} (f-g)\, d\lambda \\
&\leq \int_{A\cap B} (f-g)\, d\lambda && (f-g \leq 0 \text{ on } B^c) \\
&\leq \int_B (f-g)\, d\lambda && (0 \leq (f-g)\cdot I_{A\cap B} \leq (f-g)\cdot I_B \text{ by Choice of } B) \\
&= \mu(B) - v(B).
\end{aligned}$$

Therefore,

$$\mu(B) - v(B) = \sup_{A\in\mathcal{E}} (\mu(A) - v(A)).$$

By an identical process, we can see that

$$v(B^c) - \mu(B^c) = \sup_{A\in\mathcal{E}} (v(A) - \mu(A)).$$

Note that, since

$$(\mu(B)-v(B))-(v(B^c)-\mu(B^c))=(\mu(B)-v(B))-(\mu(B)-v(B))=0,$$

we have the equality

$$\sup_{A\in\mathcal{E}}(\mu(A)-v(A))=\sup_{A\in\mathcal{E}}(v(A)-\mu(A)).$$

Finally, we can see that

$$\begin{aligned}\frac{1}{2}\int_E|f-g|d\lambda&=\frac{1}{2}\int_B(f-g)d\lambda+\frac{1}{2}\int_{B^c}(g-f)d\lambda\\&=\frac{1}{2}(\mu(B)-v(B))+\frac{1}{2}(v(B^c)-\mu(B^c))\\&=\frac{1}{2}\sup_{A\in\mathcal{E}}(\mu(A)-v(A))+\frac{1}{2}\sup_{A\in\mathcal{E}}(v(A)-\mu(A))\\&\leq\sup_{A\in\mathcal{E}}|\mu(A)-v(A)|=\|\mu-v\|_{TV}.\end{aligned}$$

Suppose the inequality holds strictly, so that

$$\frac{1}{2}\sup_{A\in\mathcal{E}}(\mu(A)-v(A))+\frac{1}{2}\sup_{A\in\mathcal{E}}(v(A)-\mu(A))<\sup_{A\in\mathcal{E}}|\mu(A)-v(A)|.$$

By implication, there exists some $C\in\mathcal{E}$ such that

$$\frac{1}{2}\sup_{A\in\mathcal{E}}(\mu(A)-v(A))+\frac{1}{2}\sup_{A\in\mathcal{E}}(v(A)-\mu(A))<|\mu(C)-v(C)|.$$

Without loss of generality, we can assume that $\mu(C)\geq v(C)$. Then, the above inequality can be written as

$$\sup_{A\in\mathcal{E}}(\mu(A)-v(A))<\mu(C)-v(C),$$

which contradicts the definition of the supremum. Therefore, the inequality above must hold as an equality, which shows us that

$$\frac{1}{2}\int_E|f-g|d\lambda=\|\mu-v\|_{TV}.$$

In particular, this shows us that

$$\|\mu-v\|_{TV}=\sup_{A\in\mathcal{E}}(\mu(A)-v(A))=\mu(B)-v(B)=\int_B(f-g)d\lambda.$$

The integral on the right hand side is equivalent to taking the integral of the positive part

$(f-g)^+$ of $f-g$ over the entire space E, so

$$\|\mu - v\|_{TV} = \sup_{A\ in\mathcal{E}} [\mu(A) - v(A)] = \int_E (f-g)^+ d\lambda.$$

□

The above theorem requires μ, v to be absolutely continuous with respect to some σ-finite measure λ. If such a λ is not readily available, for instance in non-euclidean spaces, the next best option is to define $\lambda = \mu + v$. Being the sum of two finite measures, λ is Σ-finite and therefore a σ-finite measure. Furthermore, $\mu << \lambda$ and $v << \lambda$, since if $\lambda(A) = \mu(A) + v(A) = 0$ for some $A \in \mathcal{E}$, then $\mu(A) = v(A) = 0$. This shows us that the absolute continuity assumption above is not at all restrictive when we focus our attention on a pair of probability measures.

We can also obtain a useful characterization of the TV distance in terms of the coupling of two probability measures. Let $(E, \mathcal{E})$ be a measurable space and $(\Omega, \mathcal{H}, \mathbb{P})$ a probability space. Given any $\mu, v \in \mathcal{M}_p(\mathcal{E})$, a coupling of μ and v is a random vector (X, Y) taking values in $(E, \mathcal{E})^2$ such that the marginal distributions of X and Y are equal to μ and v.

This can also be equivalently formulated in terms of distributions. A coupling of μ and v is a joint distribution π on $(E, \mathcal{E})^2$ such that the marginals of π are μ and v, that is,

$$\pi \circ C_1^{-1} = \mu, \quad \pi \circ C_2^{-1} = v,$$

where C_1, C_2 are the coordinate mappings

$$C_1(x_1, x_2) = x_1, \quad C_2(x_1, x_2) = x_2$$

for any $(x_1, x_2) \in E^2$. Thus, we can see that

$$\pi(A \times E) = \mu(A), \quad \pi(E \times A) = v(A)$$

for any $A \in \mathcal{E}$ when π is a coupling of μ and v.

Because this formulation does not require the specification of an underlying probability space, it is what we work with here. Let $C(\mu, v)$ be the collection of all couplings of probability measures μ, v on $(E, \mathcal{E})$. The following result shows us that there exists a coupling characterization of the TV distance:

Theorem 4.3.2 (Coupling Characterization of TV Distance)**.** *Let $(E,\mathcal{E})$ be a measurable space, and $\mu,v\in\mathcal{M}_p(\mathcal{E})$ be measures that are absolutely continuous with respect to some σ-finite measure λ on $(E,\mathcal{E})$. Then, letting $\Delta=\{(x,x)\mid x\in E\}$ be the diagonal of $E\times E$,*

$$\|\mu-v\|_{TV}=\inf\{\pi(\Delta^c)\mid\pi\in C(\mu,v)\}.$$

Proof. Choose $\mu,v\in\mathcal{M}_p(\mathcal{E})$, and let π be any coupling of μ and v. Then, for any $A\in\mathcal{E}$,

$$\begin{aligned}\mu(A)-v(A)&=\pi(A\times E)-\pi(E\times A)\\&=\pi(A\times E)+\pi(E\times A^c)-1\\&\leq\pi\left[(A\times E)\cap(E\times A^c)\right]&\text{(Subadditivity for Complements)}\\&=\pi(A\times A^c)\leq\pi(\Delta^c),\end{aligned}$$

where the last inequality follows since, for any $(x,y)\in A\times A^c$, $x\neq y$ and therefore $(x,y)\notin\Delta$. This demonstrates that

$$\sup_{A\in\mathcal{E}}(\mu(A)-v(A))\leq\pi(\Delta^c),$$

and in light of the previous theorem,

$$\|\mu-v\|_{TV}\leq\pi(\Delta^c).$$

We now want to show that there exists a coupling π that attains equality, that is, an optimal coupling of μ and v. This can be achieved as follows. Define the function $\psi:E\to E\times E$ as $\psi(x)=(x,x)$ for any $x\in E$; ψ thus maps E into Δ. Note that

$$\psi^{-1}(A_1\times A_2)=A_1\cap A_2$$

for any $A_1,A_2\in\mathcal{E}$. To see this, note that, for any $x\in A_1\cap A_2$,

$$\psi(x)=(x,x)\in A_1\times A_2$$

because $x\in A_1$ and $x\in A_2$, which establishes that $A_1\cap A_2\subset\psi^{-1}(A_1\times A_2)$. Conversely, suppose $x\in\psi^{-1}(A_1\times A_2)$. Then, $\psi(x)=(x,x)\in A_1\times A_2$, so that $x\in A_1$ and $x\in A_2$, or that $x\in A_1\cap A_2$; this show us that $\psi^{-1}(A_1\times A_2)\subset A_1\cap A_2$. Since the collection of all measurable rectangles on $E\times E$ is a π-system generating $\mathcal{E}\bigotimes\mathcal{E}$, it follows that ψ is a measurable function.

Let $f,g\in\mathcal{E}_+$ be the Radon-Nikodym derivatives of μ and v with respect to λ. As in the preceding theorem, define

$$B=\{x\in E\mid f(x)\geq g(x)\}.$$

Define the function $h = \min(f,g) \in \mathcal{E}_+$, and let m be the indefinite integral of h defined as

$$m(A) = \int_A h d\lambda$$

for any $A \in \mathcal{E}$. Using m, we can define the measure M on $E \times E$ as $M = m \circ \psi^{-1}$; note that

$$M(\Delta) = m(\psi^{-1}(\Delta)) = m(E) = m(\psi^{-1}(E \times E)) = M(E \times E),$$

so that m places all its mass on Δ. If $m(E) = 1$, then $f = g = h$ λ-a.e., so that $\mu = v$ on $(E,\mathcal{E})$. In this case,

$$\|\mu - v\|_{TV} = 0 = M(\Delta^c),$$

so that M is the desired optimal coupling.

Suppose that $m(E) = p < 1$. In this case, to obtain the optimal coupling, we need to construct a few more probability measures. Define the measures m_1, m_2 on $(E,\mathcal{E})$ as

$$m_1(A) = \int_A (f-h) d\lambda = \int_{A \cap B} (f-g) d\lambda$$

and

$$m_2(A) = \int_A (g-h) d\lambda = \int_{A \cap B^c} (g-f) d\lambda$$

for any $A \in \mathcal{E}$, which are well-defined because $f-h$ and $g-h$ are non-negative measurable functions. Note that

$$m_1(E) = \int_E f d\lambda - \int_E h d\lambda = 1 - m(E) = 1 - p,$$

and likewise, $m_2(E) = 1-p$. Letting $(m_1 \times m_2)$ be the product measure on $(E,\mathcal{E})^2$ such that

$$(m_1 \times m_2)(A) = \int_E \int_E I_A(x,y) dm_1(x) dm_2(y)$$

for any $A \in \mathcal{E}^2$. Note that

$$\begin{aligned}(m_1 \times m_2)(\Delta) &= \int_E \int_E I_\Delta(x,y) dm_1(x) dm_2(y) \\ &= \int_E m_1(\{y\}) dm_2(y) = \int_{B^c} m_1(\{y\})(g(y) - f(y)) d\lambda(y).\end{aligned}$$

For any $y \in B^c$,

$$m_1(\{y\}) = \int_{\{y\} \cap B} (f(x) - g(x)) d\lambda(x) = 0,$$

so it follows that

$$(m_1 \times m_2)(\Delta) = 0.$$

This shows us that

$$\begin{aligned}(m_1 \times m_2)(\Delta^c) &= (m_1 \times m_2)(E \times E) - (m_1 \times m_2)(\Delta) \\ &= m_1(E)m_2(E) = (1-p)^2,\end{aligned}$$

so that $(m_1 \times m_2)$ places the entirety of its mass on Δ^c, in contrast to M, which places the entirety of its mass on Δ.

The desired coupling π is now defined as

$$\pi = \frac{m_1 \times m_2}{1-p} + M,$$

where $m_1 \times m_2$ is the product of the measures m_1 and m_2. Note that π is indeed a probability measure on $(E, \mathcal{E})^2$;

$$\pi(E \times E) = \frac{m_1(E)m_2(E)}{1-p} + M(E \times E) = 1 - p + p = 1.$$

Furthermore, for any $A \in \mathcal{E}$,

$$\begin{aligned}\pi(A \times E) &= \frac{m_1(A)m_2(E)}{1-p} + M(A \times E) \\ &= m_1(A) + m(\psi^{-1}(A \times E)) = m_1(A) + m(A) \\ &= \int_A (f-h)d\lambda + \int_A h d\lambda = \int_A f d\lambda = \mu(A).\end{aligned}$$

Similarly, $\pi(E \times A) = v(A)$, so that μ and v are indeed the marginals of π.

Finally, note that, by the preceding theorem,

$$\|\mu - v\|_{TV} = \frac{1}{2}\int_E |f-g| d\lambda.$$

Moreover, since $M(\Delta) = M(E \times E) = p$, we have $M(\Delta^c) = M(E \times E) - M(\Delta) = 0$ and

$$\pi(\Delta^c) = \frac{1}{1-p}(m_1 \times m_2)(\Delta^c) + M(\Delta^c) = \frac{1}{1-p}(m_1 \times m_2)(\Delta^c) = (1-p).$$

Therefore,

$$\begin{aligned}\|\mu - v\|_{TV} &= \frac{1}{2}\int_E |f-g| d\lambda \\ &= \frac{1}{2}\int_E (f + g - 2\min(f,g))\, d\lambda\end{aligned}$$

$$= 1 - \int_E h\lambda = 1 - p = \pi(\Delta^c),$$

which shows that π is the optimal coupling.

□

Corollary 4.3.3 (Copuling Inequality)**.** *Let $(E,\mathcal{E})$ be a measurable space, and $\mu, v \in \mathcal{M}_p(\mathcal{E})$ be measures that are absolutely continuous with respect to some σ-finite measure λ on $(E,\mathcal{E})$. Then, letting $\Delta = \{(x,x) \mid x \in E\}$ be the diagonal of $E \times E$, for any coupling π of μ and v, we have*

$$\|\mu - v\|_{TV} \leq \pi(\Delta^c).$$

Proof. This is just the first thing proved in the preceding theorem.

□

Corollary 4.3.4 (Coupling Characterization of TV Distance in Terms of Random Variables)**.** *Let $(E,\mathcal{E})$ be a measurable space, and $\mu, v \in \mathcal{M}_p(\mathcal{E})$ be measures that are absolutely continuous with respect to some σ-finite measure λ on $(E,\mathcal{E})$. Then, there exists a probability space $(\Omega,\mathcal{H},\mathbb{P})$ and a random vector (X,Y) taking values in $(E,\mathcal{E})^2$ such that the marginal distributions of X and Y are μ and v, and we have*

$$\|\mu - v\|_{TV} = \mathbb{P}(X \neq Y).$$

Proof. From the coupling characterization of the TV distance, we can tell that there exists an optimal coupling π of μ and v such that

$$\|\mu - v\|_{TV} = \pi(\Delta^c),$$

where $\Delta = \{(x,x) \mid x \in E\}$ is the diagonal of $E \times E$. Let $(\Omega,\mathcal{H},\mathbb{P})$ be a probability space and (X,Y) a random vector taking values in $(E,\mathcal{E})^2$ that has distribution π (we can construct such a probability space and the corresponding random vector by appealing to Ionescu-Tulcea's theorem; see the probability theory text). It follows that

$$\pi(\Delta^c) = \mathbb{P}((X,Y) \notin \Delta) = \mathbb{P}(X \neq Y),$$

so we have

$$\|\mu - v\|_{TV} = \mathbb{P}(X \neq Y).$$

□

4.3.2 Proof of the Theorem

Starting with some intuition, the Bernstein-von Mises theorem holds when the underlying log-likelihood admits a second order Taylor expansion around the maximum likelihood estimator. Since the first order term is equal to 0 by the definition of the maximum likelihood estimator, the resulting expression is identical to the kernel of the normal log-likelihood. Meanwhile, the information contained in the prior is drowned out by the information in the likelihood as the sample size grows larger, so that the posterior density ultimately converges to the normal density.

The required second order Taylor expansion holds when the log-likelihood is sufficiently well-behaved, for example when it is twice continuously differentiable and dominated by some integrable function. In many cases these assumptions are overly restrictive; we often want to consider likelihoods that are not even differentiable everywhere. It is for this reason that we turn to the usual local asymptotic normality (LAN) property, which essentially posits that the log-likelihood admits a second-order expansion around the neighborhood of the true parameter. The LAN property follows from weak conditions such as differentiability in quadratic mean (DQM) in the iid case and similar conditions for non-iid data, but for the time being we simply assume that the likelihood satisfies the LAN property without delving into more primitive conditions.

Alongside the LAN property, we also require that the true parameter is well identified. Specifically, we assume that the difference between the parameter $\bar{\Theta}$ and the true parameter θ_0 is bounded in probability when suitably scaled. This identification assumption is more stringent than the one assumed to prove posterior consistency, since it specifies the specific rate at which $\bar{\Theta}$ must move toward θ_0.

In this section we discuss each sufficient condition in turn and examine the role they play in the proof of the Bernstein-von Mises theorem. It turns out that the LAN property ensures that the theorem holds on each compact subset of $\mathbb{R}^k$; and that the identification condition ensures that the posterior distribution is well-approximated by its truncation to a compact subset. Formally, our assumptions are given as follows:

B1 Frequentist Foundations

Unlike the discussion of posterior consistency, we take an explicitly frequentist approach here. Specifically, we let Θ be a Borel subset of the euclidean k-space $\mathbb{R}^k$, and $\theta_0 \in \Theta$ the true parameter of interest. Furthermore, we assume that θ_0 is contained in the interior of Θ.

B2 Existence of Dominating Measure

For any $T \in N_+$, there exists a σ-finite measure λ_T on $(E, \mathcal{E})^T$ and a jointly measurable function $f_T : E^T \times \Theta \to \mathbb{R}_{++}$ such that

$$\mu_T(\theta, A) = \int_A f_T(x \mid \theta) d\lambda_T(x) \quad \forall A \in \mathcal{E}^T.$$

We also assume that $f_T(x \mid \cdot)$ is continuous at θ_0 for each $x \in E^T$.

We will denote

$$f_T(x \mid \theta) = p_{T,\theta}(x)$$

in line with the usual notation used in theories incorporating local asymptotic normality.

B3 Existence of Lebesgue Density for Prior Distribution

The prior distribution P is absolutely continuous with respect to the Lebesgue measure on Θ, and the density is denoted $q : \Theta \to \mathbb{R}_+$:

$$P(A) = \int_A q(\theta) d\theta \quad \forall A \in \vartheta.$$

Furthermore, $q(\theta_0) > 0$ and q is continuous at θ_0.

By implication, the posterior distribution admits a density π_T defined as

$$\pi_T(\theta \mid x) = \begin{cases} \frac{f_T(x|\theta)q(\theta)}{m_T(x)} & \text{if } m_T(x) \in (0, +\infty) \\ q(\theta) & \text{otherwise} \end{cases},$$

where $m_T(x)$ is the usual marginal likelihood defined as

$$m_T(x) = \int_\Theta f_T(x \mid \theta) q(\theta) d\theta \quad \forall x \in E^T.$$

B4-1 Uniform Local Asymptotic Normality in Probability

There exists a positive definite $k \times k$ matrix V_0, a non-singular $k \times k$ matrix δ_T such that $\delta_T \to O_{k\times k}$ as $T \to \infty$, and a measurable function $\Delta_{T,\theta_0} : E^T \to \mathbb{R}^k$ for each $T \in N_+$ such that

$$\Delta_{T,0}(X^T) \xrightarrow{d} \mathcal{N}[O_{k\times 1},\, V_0]$$

and

$$\sup_{h\in K} \left| \log \frac{p_{T,\theta_0+\delta_T h}}{p_{T,\theta_0}}(X^T) - h'\Delta_{T,0}(X^T) + \frac{1}{2} h' V_0 h \right| = o_p(1) \quad \text{as } T \to \infty$$

for any compact subset K of $\mathbb{R}^k$, where the probability statement is made under the true parameter θ_0.

B4-2 Uniform LAN in Mean Square

There exists a positive definite $k \times k$ matrix V_0, a non-singular $k \times k$ matrix δ_T such that $\delta_T \to O_{k\times k}$ as $T \to \infty$, and a measurable function $\Delta_{T,\theta_0} : E^T \to \mathbb{R}^k$ for each $T \in N_+$ such

that

$$\Delta_{T,0}(X^T) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\, V_0\right]$$

and

$$\lim_{T\to\infty} \sup_{h\in K} \int_E \left| \log \frac{p_{T,\theta_0+\delta_T h}}{p_{T,\theta_0}}(x) - h'\Delta_{T,0}(x) + \frac{1}{2}h'V_0 h \right|^2 \mu_T(\theta_0, dx) = 0$$

for any compact subset K of $\mathbb{R}^k$.

Furthermore, for any compact subset K of $\mathbb{R}^k$, there exists some $C > 0$ such that

$$\lim_{T\to\infty} \mu_T \left[\theta_0,\, \left\{ x \in E^T \mid \sup_{h\in K} \pi_T(\theta_0 + \delta_T h \mid x) \le C \int_K \pi_T(\theta_0 + \delta_T g \mid x) dg \right\} \right] = 1.$$

B5 Identification Condition

For any sequence $\{M_T\}_{T\in N_+}$ of scalars such that $M_T \uparrow +\infty$,

$$\lim_{T\to\infty} \int_{E^T} \Pi_T \left[x,\, \left\{ \theta \in \Theta \mid \left| \delta_T^{-1}(\theta - \theta_0) \right| > M_T \right\} \right] \mu_T(\theta_0, dx) = 0.$$

In this setting, we will examine the roles of assumptions B4 and B5 in turn. We first introduce some objects that will come in handy going forward.

Posterior of $\delta_T^{-1}(\theta - \theta_0)$ and its Restriction to Compact Sets

We first define $\tilde{\Pi}_T : E^T \times \mathcal{B}(\mathbb{R}^k) \to [0,1]$ as

$$\tilde{\Pi}_T(x, B) = \int_{\Theta\cap(\theta_0+\delta_T B)} \pi_T(\theta \mid x) d\theta = \Pi_T(x,\, \Theta \cap (\theta_0 + \delta_T B))$$

for any $(x, B) \in E^T \times \mathcal{B}(\mathbb{R}^k)$. We can immediately see that $\tilde{\Pi}_T$ is a transition probability kernel from $(E, \mathcal{E})^T$ into $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. In addition, for any $B \in \mathcal{B}(\mathbb{R}^k)$,

$$\begin{aligned} \tilde{\Pi}_T(\cdot, B) \circ X^T &= \mathbb{P}\left(\bar{\Theta} \in \theta_0 + \delta_T B \mid X^T \right) \\ &= \mathbb{P}\left(\delta_T^{-1}\left(\bar{\Theta} - \theta_0 \right) \in B \mid X^T \right), \end{aligned}$$

so that $\tilde{\Pi}_T$ is the conditional distribution of $\delta_T^{-1}\left(\bar{\Theta} - \theta_0\right)$ given X^T.

Given any compact subset K of $\mathbb{R}^k$, we define

$$\tilde{\Pi}_T^K(x,B) = \begin{cases} \frac{\tilde{\Pi}_T(x,B\cap K)}{\tilde{\Pi}_T(x,K)} & \text{if } \tilde{\Pi}_T(x,K) \in (0,+\infty) \\ P(B) & \text{otherwise} \end{cases}.$$

$\tilde{\Pi}_T^K$ is once again a transition probability kernel from $(E,\mathcal{E})^T$ into $(\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k))$, and this time $\tilde{\Pi}_T^K(x,\cdot)$ represents the truncation of the distribution $\tilde{\Pi}_T(x,\cdot)$ to the compact set K for those values of $x \in E^T$ for which $\tilde{\Pi}_T(x,K) \in (0,+\infty)$. Note also that, because θ_0 is an interior point of Θ, $\delta_T \to O_{k\times k}$ as $T \to \infty$ and K is bounded by the Heine-Borel theorem,

$$\begin{aligned} \tilde{\Pi}_T(x,B\cap K) &= \Pi_T(x,\theta_0+\delta_T(B\cap K)) \\ &= \int_{\theta_0+\delta_T(B\cap K)} \pi_T(\theta \mid x)d\theta \\ &= \int_{B\cap K} |\det(\delta_T)|\pi_T(\theta_0+\delta_T h \mid x)dh. \end{aligned}$$

It follows that, if $\tilde{\Pi}_T(x,K) \in (0,+\infty)$,

$$h \mapsto \frac{\pi_T(\theta_0+\delta_T h \mid x)I_K(h)}{\int_K \pi_T(\theta_0+\delta_T h' \mid x)dh'}$$

is the density of $\tilde{\Pi}_T^K(x,\cdot)$ with respect to the Lebsegue measure on $\mathbb{R}^k$ for large enough T.

Conditional Distribution of $\mathcal{N}\left[V_0^{-1}\Delta_{T,0}(X^T), V_0^{-1}\right]$ given X^T

Now define $\Phi_T : E^T \times \mathcal{B}(\mathbb{R}^k) \to [0,1]$ as

$$\Phi_T(x,B) = |V_0|^{\frac{1}{2}} \int_B \phi\left[V_0^{\frac{1}{2}\prime}\left(h - V_0^{-1}\Delta_{T,0}(x)\right)\right]dh,$$

where $\phi : \mathbb{R}^k \to \mathbb{R}_+$ is the k-dimensional standard normal density and $V_0^{\frac{1}{2}}$ is the PSD square root of V_0. Φ_T is also a transition probability kernel, with the Lebesgue density of $\Phi_T(x,\cdot)$ given by the mapping

$$h \mapsto |V_0|^{\frac{1}{2}}\phi\left[V_0^{\frac{1}{2}\prime}\left(h - V_0^{-1}\Delta_{T,0}(x)\right)\right].$$

As with $\tilde{\Pi}_T$, we can define the truncation

$$\Phi_T^K(x,B) = \frac{\Phi_T(x,B\cap K)}{\Phi_T(x,K)}$$

for any compact set K that has non-zero Lebesgue measure (which ensures that $\Phi_T(x,K) > 0$ as well). For each $x \in E^T$, the distribution $\Phi_T^K(x,\cdot)$ has density

$$h \mapsto \frac{\phi\left[V_0^{\frac{1}{2}\prime}\left(h - V_0^{-1}\Delta_{T,0}(x)\right)\right] I_K(h)}{\int_K \phi\left[V_0^{\frac{1}{2}\prime}\left(h - V_0^{-1}\Delta_{T,0}(x)\right)\right] dh'}.$$

For notational brevity, we denote

$$\phi_T(h \mid x) = \phi\left[V_0^{\frac{1}{2}\prime}\left(h - V_0^{-1}\Delta_{T,0}(x)\right)\right].$$

BvM Theorem on Compact Sets

Formally, our objective is to show that

$$\left\|\tilde{\Pi}_T(X^T,\cdot) - \Phi_T(X^T,\cdot)\right\|_{TV} = o_p(1) \quad \text{as } T \to \infty.$$

As a preliminary step, we first prove that, for any compact subset K of $\mathbb{R}^k$ with non-zero Lebesgue measure,

$$\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV} \mu_T(\theta_0, dx) \to 0,$$

which implies that

$$\left\|\tilde{\Pi}_T^K(X^T,\cdot) - \Phi_T^K(X^T,\cdot)\right\|_{TV} \left(I_{\{\tilde{\Pi}_T(\cdot,K)>0\}} \circ X^T\right) = o_p(1) \quad \text{as } T \to \infty.$$

It is this step that the LAN property justifies, as we show below.

Theorem 4.3.5. *Suppose assumptions B1 to B3 hold, and that either B4-1 or B4-2 hold. Choose any compact subset K of $\mathbb{R}^k$ that has non-zero Lebesgue measure. Then,*

$$\lim_{T\to\infty} \int_{\{\tilde{\Pi}_T(\cdot,K)>0\}} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV} \mu_T(\theta_0, dx) = 0.$$

Proof. Choose any compact subset K of $\mathbb{R}^k$ with non-zero Lebesgue measure and $\epsilon \in (0,1)$. By the continuity of $q(\cdot)$ at θ_0, the assumption $q(\theta_0) > 0$ and the fact that δ_T is $o(1)$, there must exist some $N_0 \in N_+$ such that

$$\left|\frac{q(\theta_0 + \delta_T h)}{q(\theta_0)} - 1\right| < \epsilon$$

for any $T \geq N_0$. Define

$$D_T(h \mid x) = \log \frac{p_{T,\theta_0+\delta_T h}}{p_{T,\theta_0}}(x) - h'\Delta_{n,0}(x) + \frac{1}{2}h'V_0 h$$

for any $h \in K$ and $x \in E^T$ for large enough T.

From the L^1-characteriztion of total variation distance, choose any compact set K and $x \in E^T$ such that $\tilde{\Pi}_T(x, K) \in (0, +\infty)$ and $f_T(x \mid \theta_0) \in (0, +\infty)$. Due to the assumed continuity of $\theta \mapsto f_T(x \mid \theta)$ and $q(\cdot)$ at θ_0, as well as $q(\theta_0) > 0$, this implies that

$$m_T(x) = \int_\Theta f_T(x \mid \theta) q(\theta) d\theta > 0.$$

Now we note that

$$\left\| \tilde{\Pi}_T^K(x, \cdot) - \Phi_T^K(x, \cdot) \right\|_{TV} = \int_{K \cap \{\pi_T(\theta_0+\delta_T \cdot \mid x) > 0\}} \left(1 - \frac{[\int_K \pi_T(\theta_0 + \delta_T g \mid x) dg]\, \phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x)\, [\int_K \phi_T(h' \mid x) dh']} \right)^+ \tilde{\Pi}_T^K(x, dh).$$

For any $h \in K \cap \{\pi_T(\theta_0 + \delta_T \cdot \mid x) > 0\}$, we have

$$\begin{aligned} \frac{[\int_K \pi_T(\theta_0 + \delta_T g \mid x) dg]\, \phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x)\, [\int_K \phi_T(h' \mid x) dh']} &= \left[\int_K \frac{\pi_T(\theta_0 + \delta_T g \mid x)}{\phi_T(g \mid x)} \phi_T(g \mid x) dg \right] \frac{\phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x)\, [\int_K \phi_T(h' \mid x) dh']} \\ &= \int_K \frac{\pi_T(\theta_0 + \delta_T g \mid x) \phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x) \phi_T(g \mid x)} \Phi_T^K(x, dg). \end{aligned}$$

By the convexity of the mapping $x \mapsto (1-x)^+$ and Jensen's inequality, we have

$$\left\| \tilde{\Pi}_T^K(x, \cdot) - \Phi_T^K(x, \cdot) \right\|_{TV} \leq \int_{K \cap \{\pi_T(\theta_0+\delta_T \cdot \mid x) > 0\}} \int_K \left(1 - \frac{\pi_T(\theta_0 + \delta_T g \mid x) \phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x) \phi_T(g \mid x)} \right)^+ \Phi_T^K(x, dg) \tilde{\Pi}_T^K(x, dh).$$

In addition, since $m_T(x) \in (0, +\infty)$ and $f_T(x \mid \theta_0) > 0$, for any $g, h \in K$ such that $\pi_T(\theta_0 + \delta_T h \mid x) > 0$ we have

$$\begin{aligned} \frac{\pi_T(\theta_0 + \delta_T g \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x)} &= \frac{f_T(x \mid \theta_0 + \delta_T g) q(\theta_0 + \delta_T g)}{f_T(x \mid \theta_0 + \delta_T h) q(\theta_0 + \delta_T h)} \\ &= \frac{f_T(x \mid \theta_0 + \delta_T g)}{f_T(x \mid \theta_0)} \left(\frac{f_T(x \mid \theta_0 + \delta_T h)}{f_T(x \mid \theta_0)} \right)^{-1} \frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \\ &= \exp\left((g-h)'\Delta_{T,0}(x) - \frac{1}{2}\left[g'V_0 g - h'V_0 h \right] + D_T(g \mid x) - D_T(h \mid x) \right) \frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \end{aligned}$$

$$\begin{aligned} \frac{\phi_T(h \mid x)}{\phi_T(g \mid x)} &= \exp\left(-\frac{1}{2} \left(h - V_0^{-1}\Delta_{T,0}(x) \right)' V_0 \left(h - V_0^{-1}\Delta_{T,0}(x) \right) \right) \\ &\quad \times \exp\left(\frac{1}{2} \left(g - V_0^{-1}\Delta_{T,0}(x) \right)' V_0 \left(g - V_0^{-1}\Delta_{T,0}(x) \right) \right), \end{aligned}$$

so that

$$\frac{\pi_T(\theta_0+\delta_T g \mid x)\phi_T(h \mid x)}{\pi_T(\theta_0+\delta_T h \mid x)\phi_T(g \mid x)} = \exp(D_T(g \mid x) - D_T(h \mid x))\frac{q(\theta_0+\delta_T g)}{q(\theta_0+\delta_T h)}.$$

This holds for any $g \in K$ and $h \in K$ such that $\pi_T(\theta_0+\delta_T h \mid x) > 0$, and

$$\frac{q(\theta_0+\delta_T g)}{q(\theta_0+\delta_T h)} = \frac{q(\theta_0+\delta_T g)}{q(\theta_0)} \times \frac{q(\theta_0)}{q(\theta_0+\delta_T h)} \geq \frac{1-\epsilon}{1+\epsilon}$$

for any $T \geq N_0$.

Now we can use either assumption B4-1 or B4-2 to justify the remaining steps.

Case 1: Assumption B4-1 holds

By uniform LAN in probability,

$$\sup_{h\in K}\left|D_T(h \mid X^T)\right| = o_p(1) \quad \text{as } T \to \infty,$$

or equivalently,

$$\lim_{T\to\infty} \mu_T\left[\theta_0, \quad \left\{x \in E^T \mid \sup_{h\in K}\left|D_T(h \mid X^T)\right| > \eta\right\}\right] = 0$$

for any $\eta \in (0,1)$.

Using what we have already shown above, we now have

$$\left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV} \leq \left(1 - \frac{1-\epsilon}{1+\epsilon}\exp\left(-2\sup_{h\in K}|D_T(h \mid x)|\right)\right)^+$$

for any $T \geq N_0$. We can see now that

$$\begin{aligned}
&\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV} \mu_T(\theta_0, dx) \\
&\qquad = \int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap\{f_T(\cdot|\theta_0)>0\}} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV} \mu_T(\theta_0, dx) \\
&\leq \int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap\{f_T(\cdot|\theta_0)>0\}} \left(1 - \frac{1-\epsilon}{1+\epsilon}\exp\left(-2\sup_{h\in K}|D_T(h \mid x)|\right)\right)^+ \mu_T(\theta_0, dx) \\
&\qquad \leq \eta + \mu_T\left[\theta_0, \left\{x \in E^T \mid 1 - \frac{1-\epsilon}{1+\epsilon}\exp\left(-2\sup_{h\in K}|D_T(h \mid x)|\right) > \eta\right\}\right] \\
&\qquad\qquad \leq \eta + \mu_T\left[\theta_0, \left\{x \in E^T \mid \sup_{h\in K}|D_T(h \mid x)| > -\frac{1}{2}\log\left(\frac{1+\epsilon}{1-\epsilon}(1-\eta)\right)\right\}\right]
\end{aligned}$$

for any $T \geq N_0$. By assumption, the latter term goes to 0 as $T \to \infty$, so that

$$\limsup_{T\to\infty} \int_{\{\tilde{\Pi}_T(\cdot,K)>0\}} \left\| \tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot) \right\|_{TV} \mu_T(\theta_0, dx) \leq \eta.$$

This holds for any $\eta \in (0,1)$, so we have our desired result.

Case 2: Assumption B4-2 holds

Now suppose that we know that

$$\sup_{h\in K} \int_{E^T} |D_T(h \mid x)|^2 \mu_T(\theta_0, dx) = o(1) \quad \text{as } T \to \infty$$

and that there exists some $C(K) > 0$ such that, if we define the measurable set $H_T(K) \in \mathcal{E}^T$ as

$$H_T(K) = \{x \in E^T \mid \sup_{h\in K} \pi_T(\theta_0 + \delta_T h \mid x) \leq C(K) \int_K \pi_T(\theta_0 + \delta_T h' \mid x) dh'\},$$

then

$$\mu_T[\theta_0, H_T(K)^c] = o(1) \quad \text{as } T \to \infty.$$

First, note that, because $\Delta_{T,0}(X^T) \xrightarrow{d} \mathcal{N}[O_{k\times 1},\ V_0]$, the sequence $\{\Delta_{T,0}(X^T)\}_{T\in N_+}$ is bounded in probability, that is, there exists some natural number $N_1 \geq N_0$ and some $M > 0$ such that

$$\mu_T\left[\theta_0, \{x \in E^T \mid |\Delta_{T,0}(x)| > M\}\right] < \epsilon \quad \forall T \geq N_1.$$

Define

$$\mathbf{M}_T = \{x \in E^T \mid |\Delta_{T,0}(x)| \leq M\}.$$

Noting how $(1-x)^+ \leq |\log(x)|$ for any $x > 0$, we can see that, for any $T \geq N_0$,

$$\begin{aligned}
\left(1 - \frac{\pi_T(\theta_0 + \delta_T g \mid x)\phi_T(h \mid x)}{\pi_T(\theta_0 + \delta_T h \mid x)\phi_T(g \mid x)}\right)^+ &= \left(1 - \exp(D_T(g \mid x) - D_T(h \mid x)) \frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)}\right)^+ \\
&\leq \left| D_T(g \mid x) - D_T(h \mid x) + \log \frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \right| \\
&\leq |D_T(g \mid x)| + |D_T(h \mid x)| + \left| \log \frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \right|
\end{aligned}$$

for any $g, h \in K$. We saw above that

$$\frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \geq \frac{1-\epsilon}{1+\epsilon}$$

for any $g, h \in K$, which implies that

$$\frac{q(\theta_0 + \delta_T g)}{q(\theta_0 + \delta_T h)} \leq \frac{1+\epsilon}{1-\epsilon}$$

for any $g, h \in K$. Therefore, for any $T \geq N_1$,

$$\begin{aligned}
&\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap H_T(K)\cap \mathbf{M}_T} \left\| \tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot) \right\|_{TV} \mu_T(\theta_0, dx) \\
&\qquad \leq \int_{\mathbf{M}_T} \int_K |D_T(g \mid x)| \Phi_T^K(x, dg) \mu_T(\theta_0, dx) \\
&\quad + \int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap H_T(K)} \int_K |D_T(h \mid x)| \tilde{\Pi}_T^K(x, dh) \mu_T(\theta_0, dx) + \log\left(\frac{1+\epsilon}{1-\epsilon}\right).
\end{aligned}$$

Starting with the first term, note that

$$\left| V_0^{\frac{1}{2}} \left(h - V_0^{-1} \Delta_{T,0}(x) \right) \right| \leq \operatorname{tr}(V_0) \left(|h| + \left\| V_0^{-1} \right\| \times |\Delta_{T,0}(x)| \right)$$

for any $x \in \mathbf{M}_T$ and $h \in K$. Since K is bounded by Heine-Borel and $|\Delta_{T,0}(x)| \leq M$ by the definition of $\mathbf{M}_T$, it follows that there exists some $B > 0$ such that

$$\left| V_0^{\frac{1}{2}} \left(h - V_0^{-1} \Delta_{T,0}(x) \right) \right| \leq B \quad \forall x \in \mathbf{M}_T, \quad h \in K.$$

Therefore,

$$\phi\left(V_0^{\frac{1}{2}} \left(h - V_0^{-1} \Delta_{T,0}(x) \right) \right) \geq \left(\frac{1}{2\pi} \right)^{\frac{k}{2}} e^{-\frac{B^2}{2}}$$

where $\iota_k \in \mathbb{R}^k$ is the k-dimensional vector of ones. Since the maximum possible value that the standard normal density $\phi(\cdot)$ can attain is $\left(\frac{1}{2\pi}\right)^{\frac{k}{2}}$, we can see that the density of $\Phi_T^K(x,\cdot)$ is bounded above by

$$\frac{e^{-\frac{B^2}{2}}}{\operatorname{vol}(K)},$$

where $\operatorname{vol}(K)$ is the (non-zero) volume (or Lebesgue measure) of K. Denoting

$$C_1(K, \epsilon) = \frac{e^{-\frac{B^2}{2}}}{\operatorname{vol}(K)},$$

we can now see that

$$\int_{\mathbf{M}_T}\int_K |D_T(g \mid x)|\Phi_T^K(x,dg)\mu_T(\theta_0,dx) \leq C_1(K,\epsilon)\int_K \left(\int_{E^T} |D_T(g \mid x)|\mu_T(\theta_0,dx)\right) dg$$

(Fubini's Theorem)

$$\leq C_1(K,\epsilon)\mathrm{vol}(K)\left(\sup_{g\in K}\int_{E^T} |D_T(g \mid x)|^2\mu_T(\theta_0,dx)\right)^{\frac{1}{2}}.$$

(Jensen's Inequality)

Meanwhile, recall that, for $x \in H_T(K)$ such that $\tilde{\Pi}_T(x,K) > 0$,

$$\frac{\pi_T(\theta_0+\delta_T h \mid x)}{\int_K \pi_T(\theta_0+\delta_T h' \mid x)dh'} \leq C(K) \quad \forall h \in K,$$

so that

$$\int_K |D_T(h \mid x)|\tilde{\Pi}_T^K(x,dh) = \int_K |D_T(h \mid x)|\frac{\pi_T(\theta_0+\delta_T h \mid x)}{\int_K \pi_T(\theta_0+\delta_T h' \mid x)dh'}dh$$

$$\leq C(K)\int_K |D_T(h \mid x)|dh.$$

This allows us to interchange the order of integration:

$$\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap H_T(K)}\int_K |D_T(h \mid x)|\tilde{\Pi}_T^K(x,dh)\mu_T(\theta_0,dx) \leq C(K)\int_K \left(\int_{E^T} |D_T(h \mid x)|\mu_T(\theta_0,dx)\right) dh$$

(Fubini's Theorem)

$$\leq C(K)\mathrm{vol}(K)\left(\sup_{h\in K}\int_{E^T} |D_T(h \mid x)|^2\mu_T(\theta_0,dx)\right)^{\frac{1}{2}}.$$

(Jensen's Inequality)

As such,

$$\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap H_T(K)\cap\mathbf{M}_T} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)$$

$$\leq (C(K)+C_1(K,\epsilon))\,\mathrm{vol}(K)\left(\sup_{h\in K}\int_{E^T} |D_T(h \mid x)|^2\mu_T(\theta_0,dx)\right)^{\frac{1}{2}} + \log\left(\frac{1+\epsilon}{1-\epsilon}\right).$$

Taking $T \to \infty$ on both sides yields

$$\limsup_{T\to\infty}\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap H_T(K)\cap\mathbf{M}_T} \left\|\tilde{\Pi}_T^K(x,\cdot) - \Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx) \leq \log\left(\frac{1+\epsilon}{1-\epsilon}\right)$$

by assumption B4-2.

Now note how

$$\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap(H_T(K)^c\cup\mathbf{M}_T^c)}\left\|\tilde{\Pi}_T^K(x,\cdot)-\Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)\leq 2\left(\mu_T\left[\theta_0,\ H_T(K)^c\right]+\mu_T\left[\theta_0,\ \mathbf{M}_T^c\right]\right),$$

so that, as $T\to\infty$,

$$\limsup_{Tto\infty}\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}\cap(H_T(K)^c\cup\mathbf{M}_T^c)}\left\|\tilde{\Pi}_T^K(x,\cdot)-\Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)\leq 2\epsilon.$$

Putting the two results together,

$$\limsup_{Tto\infty}\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}}\left\|\tilde{\Pi}_T^K(x,\cdot)-\Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)\leq 2\epsilon+\log\left(\frac{1+\epsilon}{1-\epsilon}\right).$$

This in turn holds for any $\epsilon>0$, so

$$\int_{\{\tilde{\Pi}_T(\cdot,K)>0\}}\left\|\tilde{\Pi}_T^K(x,\cdot)-\Phi_T^K(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)=o(1).$$

□

Corollary 4.3.6. *Suppose assumptions B1 to B3 hold, and that either B4-1 or B4-2 hold. Then, there exists a sequence $\{\mathcal{K}_T\}_{T\in N_+}\subset\mathbb{R}^k$ of closed balls centered at the origin with radii $\{M_T\}_{T\in N_+}$ such that $M_T\nearrow+\infty$ and*

$$\lim_{T\to\infty}\int_{\{\tilde{\Pi}_T(\cdot,\mathcal{K}_T)>0\}}\left\|\tilde{\Pi}_T^{\mathcal{K}_T}(x,\cdot)-\Phi_T^{\mathcal{K}_T}(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)=0.$$

Proof. Define the sequence $\{K_n\}_{n\in N_+}$ as

$$K_n=\bar{B}(O_{k\times 1},n)\quad\forall n\in N_+,$$

so that it constitutes a sequence of compact sets that increase to $\mathbb{R}^k$. Define

$$a_{nT}=\int_{\{\tilde{\Pi}_T(\cdot,K_n)>0\}}\left\|\tilde{\Pi}_T^{K_n}(x,\cdot)-\Phi_T^{K_n}(x,\cdot)\right\|_{TV}\mu_T(\theta_0,dx)$$

for any $n,T\in N_+$. By construction, $a_{nT}\to 0$ as $T\to\infty$ for each $n\in N_+$.

We will construct here a subsequence $\{n_T\}_{T\in N_+}$ of N_+ such that $a_{n_T,T}\to 0$ as $T\to\infty$. We start by noting that, for each $n\in N_+$, there exists some $N_0(n)\in N_+$ such that

$$0\leq a_{nT}\leq\frac{1}{n}\quad\forall T\geq N_0(n),$$

where we choose $\{N_0(n)\}_{n\in N_+}$ so that $N_0(n)\leq N_0(n+1)$ for any $n\in N_+$.

Let the subsequence $\{N_m\}_{m\in N_+}$ of N_+ be constructed as follows:

$$N_1 = N_0(1), \quad N_{m+1} = \max\left(N_0(m+1),\, N_m+1\right).$$

By construction, $N_m \geq N_0(m)$ for any $m \in N_+$. Now, we define $\{n_T\}_{T\in N_+}$ as

$$n_T = \begin{cases} 1 & \text{if } T < N_1 \\ m & \text{if } N_m \leq T < N_{m+1} \end{cases}.$$

Since $N_m \to +\infty$ as $m \to \infty$, it follows that $n_T \to +\infty$ as $T \to \infty$ as well; intuitively, $\{n_T\}_{T\in N_+}$ represents the staggered rate at which we progress through the sets in $\{K_n\}_{n\in N_+}$.

Now, for any $T \geq N_1$, letting $N_m \leq T < N_{m+1}$ for some $m \in N_+$, since $N_0(m) \leq N_m \leq T$, we have

$$0 \leq a_{n_T,T} = a_{mT} < \frac{1}{m} = \frac{1}{n_T}.$$

Taking $T \to \infty$ on both sides shows us that

$$\lim_{T\to\infty} a_{n_T,T} = 0.$$

Finally, define $\{\mathcal{K}_T\}_{T\in N_+}$ as

$$\mathcal{K}_T = K_{n_T},$$

so that $M_T = n_T$ for each $T \in N_+$ as well. We have just shown above that

$$\lim_{T\to\infty} \int_{\{\tilde{\Pi}_T(\cdot,\mathcal{K}_T)>0\}} \left\| \tilde{\Pi}_T^{\mathcal{K}_T}(x,\cdot) - \Phi_T^{\mathcal{K}_T}(x,\cdot) \right\|_{TV} \mu_T(\theta_0, dx) = 0,$$

and $M_T \to +\infty$ as $T \to \infty$ as desired.

□

The Full BvM Theorem

It remains to see how the identification condition helps complete the proof of the BvM theorem. So far, we have seen how the LAN property helps us show the total variation distance between the posterior and the appropriately centered normal distributions converge in probability to 0 when they are conditioned on an increasing sequence of closed balls centered at the origin. The identification condition reveals that the posterior assigns positive mass to these closed balls with probability approaching 1 as $T \to \infty$, so that we can arrive at the desired result. We first state and prove a preliminary lemma.

Lemma 4.3.1. *The following results hold true:*

i) Let $(X, \mathcal{X})$ be an arbitrary measurable space, and P some probability measure on $(X, \mathcal{X})$. For any set $K \in \mathcal{X}$ such that $P(K) > 0$, defining the truncation $P^K : \mathcal{X} \to [0,1]$ as

$$P^K(B) = \frac{P(B \cap K)}{P(K)} \quad \forall B \in \mathcal{X},$$

we have

$$\left\| P - P^K \right\|_{TV} \leq 2P(K^c).$$

ii) Let $V \in \mathbb{R}^{k \times k}$ be some positive definite matrix, and let $\mathcal{N}_{\mathbf{m},V}$ denote the probability measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ corresponding to the normal distribution with mean $\mathbf{m} \in \mathbb{R}^k$ and variance matrix V. For any fixed $L > 0$ and $\epsilon > 0$, there exists some $M > 0$ such that

$$\inf_{|\mathbf{m}| \leq L} \mathcal{N}_{\mathbf{m},V} \left[\bar{B}(O_{k\times 1}, R) \right] \geq 1 - \epsilon \quad \forall R \geq M,$$

where $\bar{B}(O_{k\times 1}, R)$ is the closed ball centered at the origin with radius R.

Proof. i) For any $B \in \mathcal{X}$,

$$\begin{aligned}
\left| P(B) - P^K(B) \right| &= \frac{1}{P(K)} |P(B)P(K) - P(B \cap K)| \\
&= \frac{1}{P(K)} |P(B \setminus K)P(K) + P(B \cap K)P(K) - P(B \cap K)| \\
&= \frac{1}{P(K)} |P(B \setminus K)P(K) - P(B \cap K)P(K^c)| \\
&\leq P(B \setminus K) + \frac{P(B \cap K)}{P(K)} P(K^c).
\end{aligned}$$

Since $B \subset X$, the monotonicity of measures tells us that

$$\left|P(B) - P^K(B)\right| \leq 2P(K^c).$$

This holds for any $B \in \mathcal{X}$, so

$$\left\|P - P^K\right\|_{TV} = \sup_{B \in \mathcal{X}} \left|P(B) - P^K(B)\right| \leq 2P(K^c).$$

ii) Because V is positive definite, so it admits an orthogonal eigendecomposition $V = \Lambda D \Lambda'$, where Λ is an orthogonal matrix and the diagonal entries $\lambda_1, \cdots, \lambda_k > 0$ of D are the eigenvalues of V. Note that

$$\bar{B}(O_{k\times 1}, L) \subset \Lambda'[-L, L]^k,$$

since $\bar{B}(O_{k\times 1}, L) \subset [-L, L]^k$ and balls around the origin are invariant to rotations.

Choose any $r > 0$, and consider the rotation $\Lambda'[-r, r]^k$ of the k-cell $[-r, r]^k$. For any $\mathbf{m} \in \Lambda'[-L, L]^k$, by a linear change of variables we can see that

$$\begin{aligned}
\mathcal{N}_{\mathbf{m},V}\left(\Lambda'[-r,r]^k\right) &= \sqrt{\prod_{i=1}^k \lambda_i^{-1}} \left[\int_{\Lambda'[-r,r]^k} \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2}(x-\mathbf{m})'\Lambda' D^{-1} \Lambda (x-\mathbf{m})\right) dx\right] \\
&= \sqrt{\prod_{i=1}^k \lambda_i^{-1}} \left[\int_{[-r,r]^k} \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2}(x-\Lambda\mathbf{m})' D^{-1} (x-\Lambda\mathbf{m})\right) dx\right] \\
&= \prod_{i=1}^k \left[\int_{-r}^{r} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{1}{2\lambda_i}(x_i - (\Lambda\mathbf{m})_i)^2\right) dx_i\right].
\end{aligned}$$

We have thus reduced a k-dimensional problem to a sequence of univariate problems. The problem

$$\min_{\mathbf{m} \in \Lambda'[-L,L]^k} \mathcal{N}_{\mathbf{m},V}\left(\Lambda'[-r,r]^k\right)$$

now reduces to

$$\min_{\mathbf{u} \in [-L,L]^k} \quad \prod_{i=1}^k \left[\int_{-r}^{r} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{1}{2\lambda_i}(x_i - \mathbf{u}_i)^2\right) dx_i\right].$$

It is clear from solving each univariate problem in turn that

$$\inf_{\mathbf{m} \in \Lambda'[-L,L]^k} \mathcal{N}_{\mathbf{m},V}\left(\Lambda'[-r,r]^k\right) = \mathcal{N}_{\mathbf{m}^*(L),V}\left(\Lambda'[-r,r]^k\right),$$

where we choose

$$\mathbf{m}^*(L) = \Lambda' \begin{pmatrix} L \\ \vdots \\ L \end{pmatrix}.$$

Furthermore, the fact $\bar{B}(O_{k\times 1}, L) \subset \Lambda'[-L, L]^k$ that we established at the outset allows us to see that

$$\inf_{|\mathbf{m}|\leq L} \mathcal{N}_{\mathbf{m},V}\left(\Lambda'[-r,r]^k\right) \geq \mathcal{N}_{\mathbf{m}^*(L),V}\left(\Lambda'[-r,r]^k\right).$$

$\mathbf{m}^*(L)$ is completely independent of the radius r, and we now that the singleton $\{\mathcal{N}_{\mathbf{m}^*(L),V}\}$ is a tight collection of probability measures. By definition, there exists a compact subset K of $\mathbb{R}^k$ such that

$$\mathcal{N}_{\mathbf{m}^*(L),V}(K) \geq 1 - \epsilon.$$

By the Heine-Borel theorem, K is bounded, so there exists some $r^* > 0$ such that

$$K \subset \bar{B}(O_{k\times 1}, r^*) \subset \Lambda'[-r^*, r^*]^k.$$

Thus,

$$\begin{aligned} 1 - \epsilon &\leq \mathcal{N}_{\mathbf{m}^*(L),V}\left(\Lambda'[-r^*, r^*]^k\right) \\ &\leq \inf_{|\mathbf{m}|\leq L} \mathcal{N}_{\mathbf{m},V}\left(\Lambda'[-r^*, r^*]^k\right). \end{aligned}$$

Finally, since the set $\Lambda'[r^*, r^*]^k$, being a rotation of a k-cell, is bounded, there exists some $M > 0$ such that

$$\Lambda'[-r^*, r^*]^k \subset \bar{B}(O_{k\times 1}, M).$$

It follows that

$$1 - \epsilon \leq \inf_{|\mathbf{m}|\leq L} \mathcal{N}_{\mathbf{m},V}\left(\bar{B}(O_{k\times 1}, R)\right) \quad \forall R \geq M,$$

which concludes the proof.

□

Theorem 4.3.7 (Bernstein-von Mises Theorem)**.** *Suppose assumptions B1 to B5 hold. Then,*

$$\left\|\tilde{\Pi}_T(X^T,\cdot) - \Phi_T(X^T,\cdot)\right\|_{TV} = o_p(1) \quad \text{as } T \to \infty,$$

where the probability statement is made under the true parameter θ_0.

Proof. Let $\{K_T\}_{T\in N_+}$ be the sequence of compact balls centered at the origin and $\{M_T\}_{T\in N_+}$ their radii found in the preceding corollary. To show the desired result, we will make use of the triangle inequality applied to the total variation norm: for any $T \in N_+$,

$$\begin{aligned}\left\|\tilde{\Pi}_T(X^T,\cdot) - \Phi_T(X^T,\cdot)\right\|_{TV} \leq& \left\|\tilde{\Pi}_T(X^T,\cdot) - \tilde{\Pi}_T^{\mathcal{K}_T}(X^T,\cdot)\right\|_{TV} \\ &+ \left\|\tilde{\Pi}_T^{\mathcal{K}_T}(X^T,\cdot) - \Phi_T^{\mathcal{K}_T}(X^T,\cdot)\right\|_{TV} \\ &+ \left\|\Phi_T^{\mathcal{K}_T}(X^T,\cdot) - \Phi_T(X^T,\cdot)\right\|_{TV}.\end{aligned}$$

We already showed that the middle term vanishes in probability as $T \to \infty$. It remains to show that the remaining two terms are $o_p(1)$.

Starting with the first term, we define

$$H_T = \{x \in E^T \mid \tilde{\Pi}_T(x, \mathcal{K}_T) > 0\}.$$

Then, from the preceding lemma

$$\left\|\tilde{\Pi}_T(x,\cdot) - \tilde{\Pi}_T^{\mathcal{K}_T}(x,\cdot)\right\|_{TV} \leq 2\tilde{\Pi}_T(x, \mathcal{K}_T^c)$$

for any $x \in H_T$, so that

$$(I_{H_T} \circ X^T)\left\|\tilde{\Pi}_T(X^T,\cdot) - \tilde{\Pi}_T^{\mathcal{K}_T}(X^T,\cdot)\right\|_{TV} \leq 2\tilde{\Pi}_T(X^T, \mathcal{K}_T^c).$$

If we can show that

$$\tilde{\Pi}_T(X^T, \mathcal{K}_T^c) = o_p(1), \quad \text{and} \quad I_{H_T^c} \circ X^T = o_p(1),$$

then we can conclude that

$$\left\|\tilde{\Pi}_T(X^T,\cdot) - \tilde{\Pi}_T^{\mathcal{K}_T}(X^T,\cdot)\right\|_{TV} = o_p(1).$$

By the identification condition,

$$\int_{E^T} \tilde{\Pi}_T(x,\, \mathcal{K}_T^c)\mu_T(\theta_0, dx) = \int_{E^T} \tilde{\Pi}_T\left[x,\, \left\{\theta \in \Theta \mid \left|\delta_T^{-1}(\theta - \theta_0)\right| > M_T\right\}\right] \mu_T(\theta_0, dx) \to 0$$

as $T \to \infty$ because $M_T \nearrow +\infty$. By implication,

$$\tilde{\Pi}_T(X^T,\, \mathcal{K}_T^c) \xrightarrow{L^1} 0,$$

which immediately tells us that $\tilde{\Pi}_T(X^T,\, \mathcal{K}_T^c) = o_p(1)$. By the definition of convergence in probability,

$$\lim_{T\to\infty} \mu_T \left[\theta_0,\, \left\{x \in E^T \mid \tilde{\Pi}_T(x, \mathcal{K}_T^c) > \frac{1}{2}\right\}\right] = 0.$$

The event $\left\{x \in E^T \mid \tilde{\Pi}_T(x, \mathcal{K}_T) = 0\right\}$ is contained in $\left\{x \in E^T \mid \tilde{\Pi}_T(x, \mathcal{K}_T^c) > \frac{1}{2}\right\}$, so by the monotoncity of measures,

$$\lim_{T\to\infty} \mu_T \left[\theta_0,\, \left\{x \in E^T \mid \tilde{\Pi}_T(x, \mathcal{K}_T) = 0\right\}\right] = 0,$$

which is equivalent to the claim that

$$I_{H_T} \circ X^T \xrightarrow{p} 1.$$

Thus, we have

$$I_{H_T^c} \circ X^T, \quad \tilde{\Pi}_T(X^T, \mathcal{K}_T^c) = o_p(1)$$

and it follows that

$$\left\| \tilde{\Pi}_T(x, \cdot) - \tilde{\Pi}_T^{\mathcal{K}_T}(x, \cdot) \right\|_{TV} = o_p(1).$$

It remains to show that

$$\left\| \Phi_T(X^T, \cdot) - \Phi_T^{\mathcal{K}_T}(X^T, \cdot) \right\|_{TV} = o_p(1).$$

To this end, note that, because

$$V_0^{-1} \Delta_{T,0}(X^T) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\, V_0^{-1}\right]$$

by assumption under the true parameter θ_0, the collection $\{\mu_T(\theta_0, \cdot)^{-1} \circ \left(V_0^{-1}\Delta_{T,0}\right)^{-1}\}_{T\in N_+}$ is a tight sequence of probability measures on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ by the necessity part of Prohorov's theorem. By definition, for any $\epsilon > 0$ there exists a compact subset K of $\mathbb{R}^k$ such that

$$\inf_{T\in N_+} \mu_T \left[\theta_0,\, \left\{x \in E^T \mid V_0^{-1}\Delta_{T,0}(x) \in K\right\}\right] \geq 1 - \epsilon.$$

The Heine-Borel theorem tells us that K is bounded, so we can find some $L > 0$ such that

$$\inf_{T \in N_+} \mu_T \left[\theta_0,\ \left\{ x \in E^T \mid \left| V_0^{-1} \Delta_{T,0}(x) \right| \leq L \right\} \right] \geq 1 - \epsilon.$$

Denote

$$A_T = \left\{ x \in E^T \mid \left| V_0^{-1} \Delta_{T,0}(x) \right| \leq L \right\} \quad \forall T \in N_+.$$

Now, by the preceding lemma, we can find some $M > 0$ such that

$$\inf_{|\mathbf{m}| \leq L} \mathcal{N}_{\mathbf{m}, V_0^{-1}} \left[\bar{B}(O_{k \times 1},\ R) \right] \geq 1 - \epsilon \quad \forall R \geq M.$$

Choose $N_0 \in N_+$ so that $M_T > M$ for any $T \geq N_0$. Then, for any $T \geq N_0$, we can see that

$$\begin{aligned} \int_{E^T} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) &= \int_{A_T} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) + \int_{A_T^c} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) \\ &\leq \mu_T(\theta_0, A_T^c) + \int_{\{|V_0^{-1} \Delta_{T,0}(\cdot)| \leq L\}} \mathcal{N}_{V_0^{-1} \Delta_{T,0}(x), V_0^{-1}} \left[\bar{B}(O_{k \times 1},\ M_T)^c \right] \mu_T(\theta_0, dx). \end{aligned}$$

Since

$$\begin{aligned} &\int_{\{|V_0^{-1} \Delta_{T,0}(\cdot)| \leq L\}} \mathcal{N}_{V_0^{-1} \Delta_{T,0}(x), V_0^{-1}} \left[\bar{B}(O_{k \times 1},\ M_T)^c \right] \mu_T(\theta_0, dx) \\ &\qquad \leq \sup_{|\mathbf{m}| \leq L} \mathcal{N}_{\mathbf{m}, V_0^{-1}} \left[\bar{B}(O_{k \times 1}, M_T)^c \right] \\ &\qquad = 1 - \inf_{|\mathbf{m}| \leq L} \mathcal{N}_{\mathbf{m}, V_0^{-1}} \left[\bar{B}(O_{k \times 1},\ M_T) \right] \leq \epsilon \end{aligned}$$

because $M_T > M$ for $T \geq N_0$, it follows that

$$\int_{E^T} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) \leq (1 - \mu_T(\theta_0, A_T)) + \epsilon \leq 2\epsilon$$

for any $T \geq N_0$. Taking $T \to \infty$ on both sides reveals that

$$\limsup_{T \to \infty} \int_{E^T} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) \leq 2\epsilon,$$

and since this holds for any $\epsilon > 0$, we can see that

$$\lim_{T \to \infty} \int_{E^T} \Phi_T(x, \mathcal{K}_T^c) \mu_T(\theta_0, dx) = 0,$$

or equivalently,

$$\Phi_T(X^T, \mathcal{K}_T^c) = o_p(1)$$

under the true parameter θ_0.

From this and the fact that

$$\left\|\Phi_T(x,\cdot) - \Phi_T^{\mathcal{K}_T}(x,\cdot)\right\|_{TV} \leq 2\Phi_T(x,\mathcal{K}_T^c)$$

for any $x \in E^T$ we have our desired result.

□

4.3.3 Primitive Identification Conditions

Instead of directly verifying the identification condition, econometricians usually try to show the existence of suitable test sequences that separate the true paramter from parameter values outside a small neighborhood of the true parameter. This works when we have iid data; however, if the data are serially dependent, then identification assumptions of the form found in Chernozhukov and Hong (2003) concerning the tail behavior and quadratic curvature of the log-likelihood seems more suitable. Below we adapt the assumptions of that paper so that assumption B5 can be established.

Under assumption B3, $f_T(x \mid \theta) \in (0, +\infty)$ for any $(\theta, x) \in \Theta \times E^T$ and $T \in N_+$. This allows us to define the log-likelihood function $l_T : \Theta \times E^T \to \mathbb{R}$ as

$$l_T(\theta \mid x) = \log f_T(x \mid \theta) \quad \forall (\theta, x) \in \Theta \times E^T$$

for any $T \in N_+$. We impose the following assumptions on the log-likelihood:

B5.1 Tail Condition

For any $\eta > 0$, there exists a $c > 0$ such that

$$\liminf_{T\to\infty} \mu_T \left[\theta_0, \left\{ x \in E^T \mid \sup_{|\theta - \theta_0| > \eta} \frac{1}{\left\| \delta_T^{-1} \right\|^2} \left(l_T(\theta \mid x) - l_T(\theta_0 \mid x) \right) < -c \right\} \right] = 1.$$

B5.2 Annulus Condition

For any $\epsilon > 0$, there exists an $\eta > 0$ and an $M > 0$ such that

$$\liminf_{T\to\infty} \mu_T \left[\theta_0, \left\{ x \in E^T \mid \sup_{M < |\delta_T^{-1}(\theta - \theta_0)| \leq \|\delta_T^{-1}\| \eta} \frac{l_T(\theta \mid x) - l_T(\theta_0 \mid x)}{\left| \delta_T^{-1}(\theta - \theta_0) \right|^2} < -\epsilon \right\} \right] \geq 1 - \epsilon.$$

B5.3 Miscellaneous Conditions

For any $c > 0$,

$$\lim_{T\to\infty} |\det(\delta_T)| e^{-c \left\| \delta_T^{-1} \right\|^2} = 0.$$

Furthermore, the essential supremum of the prior density $q(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^k$ is bounded: $\|q\|_\infty < +\infty$.

As before, we write $\pi_T(\theta \mid x)$ for the posterior density of θ given $x \in E^T$ defined as

$$\pi_T(\theta \mid x) = \begin{cases} \frac{f_T(x|\theta) q(\theta)}{m_T(x)} & \text{if } m_T(x) \in (0, +\infty) \\ q(\theta) & \text{otherwise} \end{cases}$$

for the marginal likelihood $m_T(\cdot) = \int_\Theta f_T(\cdot \mid \theta) q(\theta) d\theta$, and Π_T for the transition probability kernel defined as

$$\Pi_T(x, A) = \int_A \pi_T(\theta \mid x) d\theta \quad \forall (x, A) \in E^T \times \vartheta,$$

so that we can interpret $\Pi_T(x, \cdot)$ as the roughly the posterior distribution of $\bar{\Theta}$ given $X^T = x$.

We now prove the following theorem, which shows that the posterior distribution of θ has the desired rate of convergence δ_T.

Theorem 4.3.8 (Frequentist Consistency of the Posterior Distribution)**.** *Suppose assumptions B1 to B3 hold alongside assumptions B5.1 to B5.3, and that either assumption B4-1 or B4-2 hold. Then, for any sequence $\{M_T\}_{T \in N_+}$ of scalars such that $M_T \uparrow +\infty$,*

$$\lim_{T \to \infty} \int_{E^T} \Pi_T \left[x, \left\{ \theta \in \Theta \mid \left| \delta_T^{-1} (\theta - \theta_0) \right| > M_T \right\} \right] \mu_T(\theta_0, dx) = 0.$$

We proceed by proving a sequence of preliminary lemmas.

Lemma 4.3.2. *Define*

$$\mathcal{A}_T = \{x \in E^T \mid m_T(x) = 0\} \cup \{x \in E^T \mid m_T(x) = +\infty\}$$

for any $T \in N_+$. Then, $\lambda_T(\mathcal{A}_T) = 0$ for any $T \in N_+$, so that we can focus on $x \in E^T$ such that $m_T(x) \in (0, +\infty)$ when making probability statements.

Proof. Suppose that $m_T(x) = 0$, so that $f_T(x \mid \theta) q(\theta) = 0$ for Lebesgue-every $\theta \in \Theta$ by the vanishing property. Since $q(\theta_0) > 0$ and $q(\cdot)$ is continuous at θ_0, there exists an neighborhood $\mathcal{N}$ around θ_0 such that $q(\theta) > 0$ for any $\theta \in \mathcal{N}$. $\mathcal{N}$ has non-zero measure under the Lebesgue measure, so there must exist some $\theta \in \mathcal{N}$ for which $f_T(x \mid \theta) q(\theta) = 0$, and by implication $f_T(x \mid \theta) = 0$. This contradicts our assumption that $f_T(x \mid \theta) > 0$ for any θ, so it must be the case that

$$\{x \in E^T \mid m_T(x) = 0\} = \emptyset.$$

Meanwhile, if $\lambda_T(\mathcal{A}_T) > 0$, then

$$\begin{aligned} 1 = \int_\Theta q(\theta) d\theta &= \int_\Theta \int_{E^T} f_T(x \mid \theta) q(\theta) d\lambda_T(x) d\theta \\ &= \int_{E^T} \left(\int_\Theta f_T(x \mid \theta) q(\theta) d\theta \right) d\lambda_T(x) \\ &= \int_{E^T} m_T(x) d\lambda_T(x) \end{aligned}$$

$$\geq \int_{\mathcal{A}_T} m_T(x) d\lambda_T(x) = +\infty.$$

This is a contradiction, so it must be the case that $\lambda_T(\mathcal{A}_T) = 0$.

□

Now define

$$H_{T,m} = \{\theta \in \Theta \mid \left|\delta_T^{-1}(\theta - \theta_0)\right| > m\}$$

for any $T \in N_+$ and $m > 0$. The next two lemmas deal with integration over carefully chosen versions of $H_{T,m}$ and its complement.

Lemma 4.3.3. *For any $\epsilon > 0$, there exist $C_1(\epsilon), C_2(\epsilon) > 0$ such that*

$$\liminf_{T\to\infty} \mu_T \left[\theta_0,\ \left\{x \in E^T \mid \int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta \geq \frac{1}{2}|\det(\delta_T)| q(\theta_0) C_2(\epsilon)\right\}\right] \geq 1 - \epsilon.$$

Proof. Choose any $\epsilon > 0$. Then, because $\Delta_{T,0}(X^T) = O_p(1)$ under $\mu_T(\theta_0, \cdot)$, it follows that there exists some $M(\epsilon) > 0$ and $N_0 \in N_+$ such that

$$\mu_T\left[\theta_0,\ \mathcal{D}_{1,T}(\epsilon)\right] \geq 1 - \frac{\epsilon}{2} \quad \forall T \geq N_0$$

for $\mathcal{D}_{1,T}(\epsilon)$ defined as

$$\mathcal{D}_{1,T}(\epsilon) = \{x \in E^T \mid |\Delta_{T,0}(x)| \leq M(\epsilon)\}.$$

Since the mapping

$$c \mapsto \frac{1}{2}\left(1 - \|V_0\| c^2\right) - cM(\epsilon)$$

is continuous and converges to 0 as $c \to 0$, it follows that there exists some $C_1(\epsilon) \in (0,1)$ such that

$$\frac{1}{2}\left(1 - \|V_0\| C_1(\epsilon)^2\right) - C_1(\epsilon) M(\epsilon) > 0.$$

Define

$$C_2(\epsilon) = \text{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1})) \left[\frac{1}{2}\left(1 - \|V_0\| C_1(\epsilon)^2\right) - C_1(\epsilon) M(\epsilon)\right] > 0,$$

where $\bar{B}_{C_1(\epsilon)}(O_{k\times 1})$ is the closed ball of radius $C_1(\epsilon)$ around the origin and $\text{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1}))$ is its volume (or Lebesgue measure).

Let $x \in \mathcal{D}_{1,T}(\epsilon)$. For any $h \in \mathbb{R}^k$, define

$$D_T(h \mid x) = l_T(\theta_0 + \delta_T h \mid x) - l_T(\theta_0) - h'\Delta_{T,0}(x) + \frac{1}{2}h'V_0 h,$$

and notice how, because $e^a \geq 1 + a$ for any $a \in \mathbb{R}$,

$$\begin{aligned}
\int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta &\geq \int_{|\delta_T^{-1}(\theta - \theta_0)| \leq C_1(\epsilon)} (1 + l_T(\theta \mid x) - l_T(\theta_0 \mid x)) q(\theta) d\theta \\
&= |\det(\delta_T)| \int_{|h| \leq C_1(\epsilon)} (1 + l_T(\theta_0 + \delta_T h \mid x) - l_T(\theta_0 \mid x))\, q(\theta_0 + \delta_T h) dh \\
&\qquad \text{(Linear Change of Variables)} \\
&= |\det(\delta_T)| \int_{|h| \leq C_1(\epsilon)} q(\theta_0 + \delta_T h) dh \\
&\quad + |\det(\delta_T)| \int_{|h| \leq C_1(\epsilon)} D_T(h \mid x) q(\theta_0 + \delta_T h) dh \\
&\quad + |\det(\delta_T)| \int_{|h| \leq C_1(\epsilon)} \left(h'\Delta_{T,0}(x) - \frac{1}{2}h'V_0 h \right) q(\theta_0 + \delta_T h) dh.
\end{aligned}$$

Since $q(\cdot)$ is continuous at θ_0 and $q(\theta_0) > 0$, there exists some $\kappa > 0$ such that

$$|q(\theta) - q(\theta_0)| < \frac{q(\theta_0)}{2} \quad \forall |\theta - \theta_0| < \kappa.$$

By implication, for any $h \in \bar{B}_1(O_{k \times 1})$,

$$q(\theta_0 + \delta_T h) > \frac{q(\theta_0)}{2}$$

for any $T \in N_+$ such that $\sup_{|h| \leq 1} |\delta_T h| \leq \|\delta_T\| < \kappa$. $\|\delta_T\| \to 0$ as $T \to \infty$, so it follows that there exists some $N_1 \geq N_0$ such that

$$q(\theta_0 + \delta_T h) > \frac{q(\theta_0)}{2} \quad \forall |h| \leq C_1(\epsilon) < 1, \quad T \geq N_1.$$

As such, for any $T \geq N_1$,

$$\begin{aligned}
\int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta &\geq |\det(\delta_T)| \frac{q(\theta_0)}{2} \text{vol}(\bar{B}_{C_1(\epsilon)}(O_{k \times 1})) \\
&\quad - |\det(\delta_T)| \frac{q(\theta_0)}{2} \int_{|h| \leq C_1(\epsilon)} |D_T(h \mid x)| dh \\
&\quad - |\det(\delta_T)| \frac{q(\theta_0)}{2} \text{vol}(\bar{B}_{C_1(\epsilon)}(O_{k \times 1})) \left(C_1(\epsilon) |\Delta_{T,0}(x)| + \frac{1}{2} C_1(\epsilon)^2 \|V_0\| \right).
\end{aligned}$$

Define

$$R_T(x) = \int_{|h| \leq C_1(\epsilon)} |D_T(h \mid x)| dh.$$

Choose any $\eta > 0$. Under assumption B4-1,

$$\mu_T \left[\theta_0,\ \{x \in E^T \mid \sup_{h \in \bar{B}_{C_1(\epsilon)}(O_{k\times 1})} |D_T(h \mid x)| > \eta \mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1}))\} \right] = o(1)$$

so that

$$\mu_T \left[\theta_0,\ \{x \in E^T \mid |R_T(x)| > \eta\} \right] = o(1)$$

as well.

Meanwhile, under assumption B4-2,

$$\begin{aligned}
\int_{E^T} |R_T(x)| \mu_T(\theta_0, dx) &= \int_{|h| \le C_1(\epsilon)} \left(\int_{E^T} |D_T(h \mid x)| \mu_T(\theta_0, dx) \right) dh && \text{(Fubini's Theorem)} \\
&\le \int_{|h| \le C_1(\epsilon)} \left(\int_{E^T} |D_T(h \mid x)|^2 \mu_T(\theta_0, dx) \right)^{\frac{1}{2}} dh && \text{(Jensen's Inequality)} \\
&\le \mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1})) \left(\sup_{h \in \bar{B}_{C_1(\epsilon)}(O_{k\times 1})} \int_{E^T} |D_T(h \mid x)|^2 \mu_T(\theta_0, dx) \right) = o(1).
\end{aligned}$$

Therefore, by an application of Markov's inequality, we once again have

$$\mu_T \left[\theta_0,\ \{x \in E^T \mid |R_T(x)| > \eta\} \right] = o(1).$$

In any case, we can see that there exists some $N_2 \ge N_1$ such that, for any $T \ge N_1$,

$$\mu_T \left[\theta_0,\ \mathcal{D}_{2,T}(\epsilon) \right] \ge 1 - \frac{\epsilon}{2},$$

where

$$\mathcal{D}_{2,T}(\epsilon) = \{x \in E^T \mid |R_T(x)| \le \frac{1}{2} \mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1}))\}.$$

Define

$$\mathcal{D}_T(\epsilon) = \mathcal{D}_{1,T}(\epsilon) \cap \mathcal{D}_{2,T}(\epsilon).$$

Then, for any $x \in \mathcal{D}_T(\epsilon)$, we know that

$$\begin{aligned}
&\int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta \\
&\ge \frac{1}{2} |\det(\delta_T)| \frac{q(\theta_0)}{2} \mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1})) \left[1 - \frac{1}{\mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1}))} R_T(x) - C_1(\epsilon) |\Delta_{T,0}(x)| - \frac{1}{2} C_1(\epsilon)^2 \|V_0\| \right]
\end{aligned}$$

$$\geq \frac{1}{2}|\det(\delta_T)|\frac{q(\theta_0)}{2}\mathrm{vol}(\bar{B}_{C_1(\epsilon)}(O_{k\times 1}))\left[\frac{1}{2}-C_1(\epsilon)M(\epsilon)-\frac{1}{2}C_1(\epsilon)^2\|V_0\|\right]$$

$$= |\det(\delta_T)|\frac{q(\theta_0)}{2}C_2(\epsilon).$$

Furthermore, for any $T \geq N_2$,

$$\mu_T\left[\theta_0,\, \mathcal{D}_T(\epsilon)\right] \geq \mu_T\left[\theta_0,\, \mathcal{D}_{1,T}(\epsilon)\right] + \mu_T\left[\theta_0,\, \mathcal{D}_{2,T}(\epsilon)\right] - 1 \geq 1-\epsilon,$$

so that

$$\liminf_{T\to\infty} \mu_T\left[\theta_0,\, \mathcal{D}_T(\epsilon)\right] \geq 1-\epsilon.$$

Finally, the set

$$\left\{x \in E^T \mid \int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x)-l_T(\theta_0|x)}q(\theta)d\theta \geq \frac{1}{2}|\det(\delta_T)|q(\theta_0)C_2(\epsilon)\right\}$$

contains $\mathcal{D}_T(\epsilon)$, so the lemma follows immediately from the preceding result.

□

Lemma 4.3.4. *Choose any $\epsilon > 0$. Then, there exist constants $c(\epsilon), M(\epsilon) > 0$ such that*

$$\liminf_{T\to\infty} \mu_T\Bigg[\theta_0,\, \Big\{x \in E^T \mid \int_{H_{T,M}} e^{l_T(\theta|x)-l_T(\theta_0|x)}q(\theta)d\theta \\ \leq |\det(\delta_T)| \times \|q\|_\infty \left(\int_{\{M<|h|\}} e^{-\epsilon|h|^2}dh\right) + e^{-c(\epsilon)\left\|\delta_T^{-1}\right\|^2}\Big\}\Bigg] \geq 1-\epsilon$$

for any $M \geq M(\epsilon)$.

Proof. Choose any $\epsilon > 0$. By assumptions B5.1 a nd B5.2, there exist $\eta(\epsilon) > 0$, $M(\epsilon) > 0$ and $c(\epsilon) > 0$ such that, for any $M \geq M(\epsilon)$,

$$\liminf_{T\to\infty} \mu_T\left[\theta_0,\, \mathcal{B}_T(M,\epsilon) \cap \mathcal{C}_T(\epsilon)\right] \geq 1-\epsilon,$$

where

$$\mathcal{B}_T(M,\epsilon) = \left\{x \in E^T \mid \sup_{M<\left|\delta_T^{-1}(\theta-\theta_0)\right|<\left\|\delta_T^{-1}\right\|\eta(\epsilon)} \frac{l_T(\theta \mid x) - l_T(\theta_0 \mid x)}{\left|\delta_T^{-1}(\theta-\theta_0)\right|^2} < -\epsilon\right\}$$

$$\mathcal{C}_T(\epsilon) = \left\{x \in E^T \mid \sup_{|\theta-\theta_0|\geq\eta(\epsilon)} \frac{1}{\left\|\delta_T^{-1}\right\|^2}\left[l_T(\theta \mid x) - l_T(\theta_0 \mid x)\right] < -c(\epsilon)\right\}.$$

Note that

$$\int_{H_{T,M}} e^{l_T(\theta|x)-l_T(\theta_0|x)} q(\theta)d\theta$$

$$= \underbrace{\int_{\{M<\left|\delta_T^{-1}(\theta-\theta_0)\right|<\left\|\delta_T^{-1}\right\|\eta(\epsilon)\}} e^{l_T(\theta|x)-l_T(\theta_0|x)} q(\theta)d\theta}_{Annul_T(x)} + \underbrace{\int_{\{\left|\delta_T^{-1}(\theta-\theta_0)\right|\geq\left\|\delta_T^{-1}\right\|\eta(\epsilon)\}} e^{l_T(\theta|x)-l_T(\theta_0|x)} q(\theta)d\theta}_{Tail_T(x)}.$$

These are the annulus and tail terms, respectively.

By our choice of x, for any $\theta \in \Theta$ such that $M < \left|\delta_T^{-1}(\theta-\theta_0)\right| < \left\|\delta_T^{-1}\right\|\eta$, we have

$$l_T(\theta \mid x) - l_T(\theta_0 \mid x) \leq \frac{l_T(\theta \mid x) - l_T(\theta_0 \mid x)}{\left|\delta_T^{-1}(\theta-\theta_0)\right|^2} \times \left|\delta_T^{-1}(\theta-\theta_0)\right|^2 \leq -\epsilon\left|\delta_T^{-1}(\theta-\theta_0)\right|^2,$$

so that

$$\begin{aligned}
Annul_T(x) &\leq \int_{\{M<\left|\delta_T^{-1}(\theta-\theta_0)\right|<\left\|\delta_T^{-1}\right\|\eta\}} e^{-\epsilon\left|\delta_T^{-1}(\theta-\theta_0)\right|^2} q(\theta)d\theta \\
&= |\det(\delta_T)| \int_{\{M<|h|<\left\|\delta_T^{-1}\right\|\eta\}} e^{-\epsilon|h|^2} q(\theta_0+\delta_T h)dh \qquad \text{(Linear Change of Variables)} \\
&\leq |\det(\delta_T)| \times \|q\|_\infty \left(\int_{\{M<|h|<\left\|\delta_T^{-1}\right\|\eta\}} e^{-\epsilon|h|^2} dh\right) \\
&\leq |\det(\delta_T)| \times \|q\|_\infty \left(\int_{\{M<|h|\}} e^{-\epsilon|h|^2} dh\right).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
Tail_T(x) &\leq \int_{\{|\theta-\theta_0|\geq\eta\}} e^{l_T(\theta|x)-l_T(\theta_0|x)} q(\theta)d\theta \\
&\leq \exp\left(\sup_{|\theta-\theta_0|\geq\eta}\left[l_T(\theta \mid x) - l_T(\theta_0 \mid x)\right]\right)\left(\int_{H_{T,M}} q(\theta)d\theta\right) \\
&\leq \exp\left(-c\left\|\delta_T^{-1}\right\|^2\right).
\end{aligned}$$

Therefore,

$$\int_{H_{T,M}} e^{l_T(\theta|x)-l_T(\theta_0|x)} q(\theta)d\theta \leq |\det(\delta_T)| \times \|q\|_\infty \left(\int_{\{M<|h|\}} e^{-\epsilon|h|^2} dh\right) + e^{-c\left\|\delta_T^{-1}\right\|^2}.$$

The desired result then follows from the same containment argument as in the preceding lemma.

□

Lemma 4.3.5. *There exists a mapping $I : (0,+\infty)^2 \times \mathbb{R}_+$ such that, for any $\epsilon > 0$,*

$$\lim_{m\to\infty} I(m,\epsilon) = 0$$

and there exists some $M(\epsilon) > 0$ such that, for any $M \geq M(\epsilon)$,

$$\limsup_{T\to\infty} \int_{E^T} \Pi_T\left[x, H_{T,M}\right] \mu_T(\theta_0, dx) \leq I(M,\epsilon) + 2\epsilon.$$

Proof. Choose any $x \in \mathcal{A}_T^c$, so that $m_T(x) \in (0,+\infty)$ and

$$\Pi_T\left[x,\ A\right] = \frac{\int_A e^{l_T(\theta|x)} q(\theta) d\theta}{\int_\Theta e^{l_T(\theta|x)} q(\theta) d\theta}.$$

for any $A \in \vartheta$ using the definition of the posterior density. Let $C_1(\epsilon), C_2(\epsilon), c(\epsilon)$ and $M(\epsilon)$ be the ones chosen in the preceding lemma, so that, for any $M \geq M(\epsilon)$,

$$\liminf_{T\to\infty} \mu_T\left[\theta_0,\ \mathcal{B}_T(\epsilon) \cap \mathcal{C}_T(M,\epsilon)\right] \geq 1 - \epsilon$$

where

$$\mathcal{B}_T(\epsilon) = \left\{ x \in E^T \mid \int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta \geq \frac{1}{2} |\det(\delta_T)| q(\theta_0) C_2(\epsilon) \right\}$$

$$\begin{aligned}\mathcal{C}_T(M,\epsilon) = \Big\{ x \in E^T \mid & \int_{H_{T,M}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta \\ & \leq |\det(\delta_T)| \times \|q\|_\infty \left(\int_{\{M<|h|\}} e^{-\epsilon|h|^2} dh \right) + e^{-c(\epsilon)\left\|\delta_T^{-1}\right\|^2} \Big\}.\end{aligned}$$

For any $x \in \mathcal{A}_T^c \cap \mathcal{B}_T(\epsilon) \cap \mathcal{C}_T(M,\epsilon)$, it now follows that

$$\begin{aligned}\Pi_T\left[x,\ H_{T,M}\right] &= \frac{\int_{H_{T,M}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta}{\int_\Theta e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta} \\ &\leq \frac{\int_{H_{T,M}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta}{\int_{H^c_{T,C_1(\epsilon)}} e^{l_T(\theta|x) - l_T(\theta_0|x)} q(\theta) d\theta} \\ &\leq \frac{2\|q\|_\infty}{q(\theta_0) C_2(\epsilon)} \left[\int_{|h|>M} e^{-\epsilon|h|^2} dh + \frac{1}{|\det(\delta_T)|} e^{-c(\epsilon)\left\|\delta_T^{-1}\right\|^2} \right].\end{aligned}$$

By assumption,

$$\frac{1}{|\det(\delta_T)|} e^{-c(\epsilon)\left\|\delta_T^{-1}\right\|^2} \to 0$$

as $T \to \infty$, so it follows that, for large enough T,

$$\Pi_T\left[x,\ H_{T,M}\right] \leq \frac{2\|q\|_\infty}{q(\theta_0)C_2(\epsilon)} \int_{|h|>M} e^{-\epsilon|h|^2} dh + \epsilon.$$

Define the function $I : (0,+\infty)^2 \to \mathbb{R}_+$ as

$$I(m,\epsilon) = \frac{2\|q\|_\infty}{q(\theta_0)C_2(\epsilon)} \int_{|h|>m} e^{-\epsilon|h|^2} dh.$$

Since $\int_{|h|>m} e^{-\epsilon|h|^2} dh \to 0$ as $m \to +\infty$ for fixed $\epsilon > 0$, we can see that

$$\lim_{m\to\infty} I(m,\epsilon) = 0.$$

Furthermore, defining

$$\mathcal{S}_T(M,\epsilon) = \mathcal{A}_T^c \cap \mathcal{B}_T(\epsilon) \cap \mathcal{C}_T(M,\epsilon),$$

we know that

$$\limsup_{T\to\infty} \mu_T\left[\theta_0,\ \mathcal{S}_T(M,\epsilon)^c\right] \leq \epsilon,$$

so that

$$\begin{aligned}
\int_{E^T} \Pi_T\left[x,\ H_{T,M}\right] \mu_T(\theta_0, dx) &= \int_{\mathcal{S}_T(M,\epsilon)} \Pi_T\left[x,\ H_{T,M}\right] \mu_T(\theta_0, dx) + \int_{\mathcal{S}_T(M,\epsilon)^c} \Pi_T\left[x,\ H_{T,M}\right] \mu_T(\theta_0, dx) \\
&\leq I(M,\epsilon) + \epsilon + \mu_T\left[\theta_0,\ \mathcal{S}_T(M,\epsilon)^c\right] \\
&\leq I(M,\epsilon) + 2\epsilon
\end{aligned}$$

for large enough T. This completes the proof.

□

Lemma 4.3.6. *For any sequence $\{M_T\}_{T\in N_+}$ of positive reals that increase to $+\infty$,*

$$\lim_{T\to\infty} \int_{E^T} \Pi_T\left[x,\ H_{T,M_T}\right] \mu_T(\theta_0, dx) = 0.$$

Proof. Choose some $\epsilon > 0$. We know that, for fixed $\epsilon > 0$,

$$\lim_{m\to\infty} I(m,\epsilon) = 0,$$

and that there exists some $M(\epsilon)$ such that, for any $M \geq M(\epsilon)$,

$$\limsup_{T\to\infty} \int_{E^T} \Pi_T\left[x,\ H_{T,M}\right] \mu_T(\theta_0, dx) \leq I(M,\epsilon) + 2\epsilon.$$

Let $M^* \geq M(\epsilon)$ be chosen so that $I(M^*,\epsilon) < \epsilon$ and choose $N \in N_+$ so that $M_T \geq M^*$ for any $T \geq N$. Then, for any $T \geq N$, $H_{T,M_T} \subset H_{T,M^*}$ and we have

$$\begin{aligned}\limsup_{T\to\infty} \int_{E^T} \Pi_T\left[x,\ H_{T,M_T}\right] \mu_T(\theta_0, dx) &\leq \limsup_{T\to\infty} \int_{E^T} \Pi_T\left[x,\ H_{T,M^*}\right] \mu_T(\theta_0, dx) \\ &\leq I(M^*,\epsilon) + 2\epsilon \leq 3\epsilon.\end{aligned}$$

This in turn holds for any $\epsilon > 0$, so

$$\int_{E^T} \Pi_T\left[x,\ H_{T,M_T}\right] \mu_T(\theta_0, dx) \to 0$$

as $T \to \infty$.

□

Chapter 5

Linear Regression Models

Here we finally introduce the linear regression model with which most of the econometric literature is concerned. We first state the most basic version of the model with normally distributed errors and independent observations; then, we drop the normality assumption and derive large sample properties of the coefficient and error variance estimators. Afterward, we present generalizations, such as non-linear models and models with dependent data.

5.1 The Normal Univariate Linear Regression Model

We first study the benchmark linear model under the traditional Gauss-Markov assumptions from a frequentist perspective. Most other models are generalizations of this basic model.

5.1.1 Model Assumptions

Our model assumptions are as follows:

A1 : **The Sample**

There is a random sample $\{(Y_t, X_t)\}_{1 \leq t \leq T}$ of size T, where each Y_t takes values in $\mathbb{R}$ and X_t in $\mathbb{R}^k$. The errors are defined as

$$\epsilon_t = Y_t - X_t'\beta$$

for any $1 \leq t \leq T$ for some parameter $\beta \in \mathbb{R}^k$.

For notational simplicity, we arrange the data in matrix form as

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix}, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_T' \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix}.$$

A2 : **No Perfect Collinearity**

We assume that the $k \times k$ random matrix $X'X$ is nonsingular with probability 1. Note that this is equivalent to assuming that X has k linearly independent columns with probability 1, since $X'X$ and X have the same null spaces in $\mathbb{R}^k$.

A3 : **Exogeneity (Mean Independence)**

We assume that the errors are mean independent of the data $X_1, \cdots, X_T$, that is,

$$\mathbb{E}[\epsilon_i \mid X_1, \cdots, X_T] = 0$$

for $1 \leq i \leq T$. This, in particular, implies that $\mathbb{E}[\epsilon_i X_j] = \mathbf{0}$ for any $1 \leq i, j \leq T$.

A4 : **Normal and Spherical Errors**

Given the data $X_1, \cdots, X_T$, the errors are spherical and normally distributed, that is,

$$\epsilon_i \mid X_1, \cdots, X_T \sim \mathcal{N}\Big[\mathbf{0}, \sigma^2 \cdot I_T\Big]$$

for $1 \leq i \leq T$, where $\sigma^2 > 0$ is our error variance parameter.

One final assumption we make is that the distribution of X does not depend on the parametrs β and σ^2. Given these assumptions, we can see that the conditional distribution of $Y = X\beta + \epsilon$ given X is normal with mean $X\beta$ and variance $\sigma^2 \cdot I_T$:

$$Y \mid X \sim \mathcal{N}\Big[X\beta,\ \sigma^2 \cdot I_T\Big].$$

The joint likelihood of the data (Y, X) is therefore given as

$$\begin{aligned} f_{Y,X}(y, x \mid \beta, \sigma^2) &= f_{Y|X}(y \mid x, \beta, \sigma^2) \cdot f_X(x) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}(y - x\beta)'(y - x\beta)\right) \cdot f_X(x). \end{aligned}$$

It follows from the Neyman factorization theorem that the quantities $Y'Y, X'X$ and $X'Y$ are sufficient statistics for the parameters β, σ^2. Thus, any sample information about these parameters are contained in these three quantities; we will see soon that the estimators of β, σ^2 we consider are functions of these sufficient statistics.

5.1.2 MLEs of Model Parameters

The first order of business is to obtain estimators of β and σ^2. The log-likelihood function is given in this case as

$$l(\beta,\sigma^2 \mid Y,X) = -\frac{T}{2}\left(\log(2\pi)+\log\left(\sigma^2\right)\right) - \frac{1}{2\sigma^2}(Y-X\beta)'(Y-X\beta)+\log f_X(X).$$

Therefore, the MLEs $\hat{\beta}$ and $\hat{\sigma}^2$ satisfy the first order conditions

$$\frac{\partial l(\hat{\beta},\hat{\sigma}^2 \mid Y,X)}{\partial \beta} = \frac{1}{\hat{\sigma}^2}X'(Y-X\hat{\beta}) = \mathbf{0}$$

$$\frac{\partial l(\hat{\beta}(Y,X),\hat{\sigma}^2 \mid Y,X)}{\partial \sigma^2} = -\frac{T}{2}\frac{1}{\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}(Y-X\hat{\beta})'(Y-X\hat{\beta}) = 0.$$

The MLEs are given as:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\sigma}^2 = \frac{1}{T}\underbrace{(Y-X\hat{\beta})'(Y-X\hat{\beta})}_{SSR}.$$

The MLE of β is simply its ordinary least squares (OLS) estimator, while the MLE of σ^2 is the average SSR, or sum of squared residuals.

5.1.3 Matrices, Residual Makers and FWL

Before talking of the finite sample distributions of the MLEs, we first define some useful matrices and where they appear in the model. The projection matrix P_X and residual maker M_X of X are defined as the following $T \times T$ random matrices:

$$P_X = X(X'X)^{-1}X', \quad M_X = I_T - P_X.$$

Some simple computation shows us that P_X, M_X are symmetric and idempotent, that is, $P_X^2 = P_X$ and $M_X^2 = M_X$. Furthermore,

$$P_X \cdot X = X, \quad M_X \cdot X = O_{T\times T}.$$

The fitted values $\hat{Y}$ of Y, defined as $\hat{Y} = X\hat{\beta}$, can be written as

$$\hat{Y} = X(X'X)^{-1}X'Y = P_X \cdot Y.$$

Likewise, the residuals $\hat{\epsilon} = Y - X\hat{\beta}$ can be written as

$$\hat{\epsilon} = Y - X(X'X)^{-1}X'Y = (I_T - P_X)Y = M_X \cdot Y.$$

Since $Y = X\beta + \epsilon$, the residuals are equivalently given as

$$\hat{\epsilon} = M_X(X\beta + \epsilon) = M_X \cdot \epsilon.$$

Now we can express MLE $\hat{\sigma}^2$ of σ^2 as

$$\hat{\sigma}^2 = \frac{1}{T}\hat{\epsilon}'\hat{\epsilon} = \frac{1}{T}\epsilon' M_X \epsilon.$$

This is a quadratic form involving a symmetric idempotent matrix M_X and (conditionally) normally distributed random vectors ϵ. We study the distribution of these kinds of quadratic forms below.

The projection and residual makers also have uses aside from allowing us to obtain a concise representaitn of $\hat{\sigma}^2$. One major example is given below:

Theorem 4.1 (Frisch-Waugh-Lovell Theorem)
Suppose the data matrix X can be partitioned as $X = (\mathbb{X}_1, \mathbb{X}_2)$, where $\mathbb{X}_i$ is $T \times k_i$-dimensional for $i = 1, 2$ and $k_1 + k_2 = k$. Let β be partitioned similarly as $\beta = (\beta_1', \beta_2')'$, where β_i is k_i-dimensional for $i = 1, 2$. Then, the OLS estimator of β_1, that is, the first k_i elements of $\hat{\beta}$, are given as

$$\hat{\beta}_1 = (\mathbb{X}_1' M_2 \mathbb{X}_1)^{-1} \mathbb{X}_1' M_2 Y,$$

where M_2 is the residual maker corresponding to $\mathbb{X}_2$ defined as

$$M_2 = I_T - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'.$$

Proof) We first study the properties of the residual maker M_2 and the matrix product $\mathbb{X}_1' M_2 \mathbb{X}_1$. By assumption, the matrix

$$X'X = \begin{pmatrix} \mathbb{X}_1'\mathbb{X}_1 & \mathbb{X}_1'\mathbb{X}_2 \\ \mathbb{X}_2'\mathbb{X}_1 & \mathbb{X}_2'\mathbb{X}_2 \end{pmatrix}$$

is almost surely nonsingular. It follows that $\mathbb{X}_1'\mathbb{X}_1$ and $\mathbb{X}_2'\mathbb{X}_2$ are also almost surely nonsingular, since $\mathbb{X}_1$ and $\mathbb{X}_2$ have almost surely linearly indepenedent columns. This allows us to see that M_2 is well-defined. Furthermore, due to the nonsingularity of $\mathbb{X}_2'\mathbb{X}_2$, we have

$$0 \neq \det(X'X) = \det(\mathbb{X}_2'\mathbb{X}_2) \cdot \det\left(\mathbb{X}_1'\mathbb{X}_1 - \mathbb{X}_1'\mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_1\right)$$

$$= \det(\mathbb{X}_2'\mathbb{X}_2) \cdot \det(\mathbb{X}_1' M_2 \mathbb{X}_1),$$

almost surely, so that $\mathbb{X}_1' M_2 \mathbb{X}_1$ is almost surely nonsingular as well. As such, all the expressions in the statement of the theorem are well-defined.

Moving onto the actual claim, note that, because $(X'X)\hat{\beta} = X'Y$, we have

$$\begin{aligned}\begin{pmatrix} \mathbb{X}_1' Y \\ \mathbb{X}_2' Y \end{pmatrix} &= \begin{pmatrix} \mathbb{X}_1' \mathbb{X}_1 & \mathbb{X}_1' \mathbb{X}_2 \\ \mathbb{X}_2' \mathbb{X}_1 & \mathbb{X}_2' \mathbb{X}_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \\ &= \begin{pmatrix} (\mathbb{X}_1' \mathbb{X}_1)\hat{\beta}_1 + (\mathbb{X}_1' \mathbb{X}_2)\hat{\beta}_2 \\ (\mathbb{X}_2' \mathbb{X}_1)\hat{\beta}_1 + (\mathbb{X}_2' \mathbb{X}_2)\hat{\beta}_2 \end{pmatrix}.\end{aligned}$$

This yields two equations

$$(\mathbb{X}_1' \mathbb{X}_1)\hat{\beta}_1 + (\mathbb{X}_1' \mathbb{X}_2)\hat{\beta}_2 = \mathbb{X}_1' Y$$

$$(\mathbb{X}_2' \mathbb{X}_1)\hat{\beta}_1 + (\mathbb{X}_2' \mathbb{X}_2)\hat{\beta}_2 = \mathbb{X}_2' Y.$$

Solving the second equation for $\hat{\beta}_2$ shows us that

$$\hat{\beta}_2 = (\mathbb{X}_2' \mathbb{X}_2)^{-1} \left[\mathbb{X}_2' Y - (\mathbb{X}_2' \mathbb{X}_1)\hat{\beta}_1 \right].$$

Substituting this for $\hat{\beta}_2$ in the first equation yields

$$(\mathbb{X}_1' \mathbb{X}_1)\hat{\beta}_1 + (\mathbb{X}_1' \mathbb{X}_2)(\mathbb{X}_2' \mathbb{X}_2)^{-1} \left[\mathbb{X}_2' Y - (\mathbb{X}_2' \mathbb{X}_1)\hat{\beta}_1 \right] = \mathbb{X}_1' Y,$$

so that

$$\mathbb{X}_1' \left[I_T - \mathbb{X}_2 (\mathbb{X}_2' \mathbb{X}_2)^{-1} \mathbb{X}_2' \right] \mathbb{X}_1 \cdot \hat{\beta}_1 = \mathbb{X}_1' \left[I_T - \mathbb{X}_2 (\mathbb{X}_2' \mathbb{X}_2)^{-1} \mathbb{X}_2' \right] Y.$$

Therefore,

$$\hat{\beta}_1 = (\mathbb{X}_1' M_2 \mathbb{X}_1)^{-1} \mathbb{X}_1' M_2 Y,$$

which is the result we desired.

Q.E.D.

5.1.4 Finite Sample Distributions of MLEs

We can obtain the finite sample distributions of $\hat{\beta}$ and $\hat{\sigma}^2$ without much effort. Starting with $\hat{\beta}$, note that

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + \epsilon) = \beta + (X'X)^{-1}X'\epsilon,$$

so that

$$\hat{\beta} \mid X \sim \mathcal{N}\left[\beta,\, \sigma^2(X'X)^{-1}\right].$$

To obtain the finite sample distribution of $\hat{\sigma}^2$, we require the following lemmas:

Lemma 4.2 (Eigenvalues of Idempotent Matrices)
Let $A \in \mathbb{R}^{n\times n}$ be an idempotent matrix, that is, a matrix such that $A^2 = A$. Then, the eigenvalues of A are equal to either 1 or 0 and if, in addition, A is symmetric, then the rank of A is equal to its trace.

Proof) Suppose A is idempotent, and let $\lambda \in \mathbb{C}$ be an eigenvalue of A. By definition, there exists a non-zero vector $v \in \mathbb{C}^n$ such that $Av = \lambda v$. Pre-multiplying both sides by A shows us that

$$\lambda v = Av = A^2v = A(\lambda v) = \lambda(Av) = \lambda^2 \cdot v.$$

Therefore,

$$\lambda|v|^2 = \lambda(\overline{v}'v) = \lambda^2(\overline{v}'v) = \lambda^2|v|^2,$$

and since $|v| > 0$, we are left with the equation $\lambda = \lambda^2$. The only complex values that solve this equation are 0 or 1, so $\lambda = 1$ or 0.

Now suppose that A is also symmetric. Then, by the principal axis theorem A can be orthogonally diagonalized as $A = PDP'$, where P is an $n \times n$ orthogonal matrix and D is a diagonal matrix collecting the eigenvalues of A. The rank of A is then equal to the rank of D, which is exactly the number of eigenvalues of A that are equal to 1, or the trace of D. On the other hand,

$$\text{tr}(A) = \text{tr}(PDP') = \text{tr}(D \cdot PP') = \text{tr}(D),$$

so that the trace of A is exactly the rank of A.

Q.E.D.

Lemma 4.3 (Quadratic Forms with Normal Vectors and Idempotent Matrices) Let Z be an n-dimensional standard normally distributed random vector, and A an $n \times n$ symmetric and idempotent matrix. Then,

$$Z'AZ \sim \chi^2_{\mathrm{tr}(A)}.$$

Proof) Let $\mathrm{tr}(A) = k$ be the rank of A. By the principal axis theorem, A can be orthogonally diagonalized as $A = PDP'$, where P is an $n \times n$ orthogonal matrix and D is a diagonal matrix collecting the eigenvalues of A. Since A is idempotent, its eigenvalues are either 1 or 0, with the number of eigenvalues equal to 1 being the trace, or rank, of A. The rank of A is assumed to be k, so letting the eigenvalues equal to 1 be arranged first in D, we can see that

$$D = \begin{pmatrix} I_k & O_{k\times(n-k)} \\ O_{(n-k)\times k} & O_{(n-k)\times(n-k)} \end{pmatrix}.$$

Therefore, letting $\tilde{P} \in \mathbb{R}^{n\times k}$ collect the first k columns of P, we can write

$$A = PDP' = \tilde{P}\tilde{P}'.$$

Note that, since the columns of $\tilde{P}$ form an orthonormal set,

$$\tilde{P}'\tilde{P} = I_k.$$

It follows that the k-dimensional random vector $\tilde{Z} = \tilde{P}'Z$ follows a k-dimensional standard normal distribution:

$$\tilde{Z} \sim \tilde{P}' \cdot \mathcal{N}[\mathbf{0}, I_n] = \mathcal{N}[\mathbf{0}, I_k].$$

As such,

$$Z'AZ = Z'\tilde{P}\tilde{P}'Z = \tilde{Z}'\tilde{Z} \sim \chi^2_k.$$

Q.E.D.

Lemma 4.3 is crucial to the derivation of the finite sample distribution of $\hat{\sigma}^2$. Recall that

$$T \cdot \hat{\sigma}^2 = \epsilon' M_X \epsilon,$$

where M_X is symmetric and idempotent, and

$$\frac{\epsilon}{\sigma} \mid X \sim \mathcal{N}[\mathbf{0}, I_T].$$

We can now write

$$\frac{T \cdot \hat{\sigma}^2}{\sigma^2} = \left(\frac{\epsilon}{\sigma}\right)' M_X \left(\frac{\epsilon}{\sigma}\right)$$

and

$$\text{tr}(M_X) = T - \text{tr}(X(X'X)^{-1}X') = T - \text{tr}((X'X)^{-1}X'X) = T - k.$$

As such, by lemma 4.3, we have

$$\frac{T \cdot \hat{\sigma}^2}{\sigma^2} \mid X \sim \chi^2_{T-k}.$$

5.1.5 Testing Linear Hypotheses for β

Now that we have derived the finite sample distributions of $\hat{\beta}$ and $\hat{\sigma}^2$, we are now interested in forming test statistics using these estimators. For starters, we note that $\hat{\beta}$ and $\hat{\sigma}^2$ are conditionally independent given X. More specifically, $X'\epsilon$ and $M_X \cdot \epsilon$ are conditionally independent given X. This can be seen by noting that

$$\begin{pmatrix} X'\epsilon \\ M_X\epsilon \end{pmatrix} \mid X \sim \begin{pmatrix} X' \\ M_X \end{pmatrix} \cdot \mathcal{N}\left[\mathbf{0}, \sigma^2 \cdot I_T\right] = \mathcal{N}\left[\mathbf{0}, \sigma^2 \begin{pmatrix} X'X & O_{n\times n} \\ O_{n\times n} & M_X \end{pmatrix}\right].$$

In other words, $X'\epsilon$ and $M_X\epsilon$ are jointly normal given X with conditional covariance 0, and as such they are conditionally independent given X by the equivalence of independence and uncorrelatedness for jointly normally distributed random variables. $\hat{\beta}$ is a linear function of $X'\epsilon$, while $\hat{\sigma}^2$ is basically the scaled norm of $M_X\epsilon$, so it follows that $\hat{\beta}$ and $\hat{\sigma}^2$ are conditionally independent given X.

We can now form test statistics. First consider testing the hypothesis $H_0 : R\beta = q$ against $H_1 : R\beta \neq r$, where $R \in \mathbb{R}^{k\times r}$ has full rank $r \leq k$, and $q \in \mathbb{R}^r$. From the finite sample distributions we obtained above,

$$R\hat{\beta} - R\beta \mid X \sim \mathcal{N}\left[\mathbf{0},\ \sigma^2 \cdot R(X'X)^{-1}R'\right],$$

so that

$$\left(R\hat{\beta}-R\beta\right)'\left[\sigma^2\cdot R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta}-R\beta\right)\mid X\mid\chi^2_k.$$

This statistic is conditionally independent of the chi-squared random variable $\frac{T\cdot\hat{\sigma}^2}{\sigma^2}$ given X, so by definition the statistic

$$F=\frac{\frac{1}{k}\left(R\hat{\beta}-R\beta\right)'\left[\sigma^2\cdot R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta}-R\beta\right)}{\frac{1}{T-k}\cdot\frac{T\cdot\hat{\sigma}^2}{\sigma^2}}$$

follows an F distribution with degrees of freedom equal to k and $T-k$, given X. Since this distribution does not depend on X, F follows the same F distribution unconditionally as well. To clean up the expression for F, we define the scaled sample error variance

$$s^2=\frac{T\cdot\hat{\sigma}^2}{T-k}=\frac{1}{T-k}SSR.$$

Then,

$$\begin{aligned}F&=\frac{1}{k}\cdot\frac{\left(R\hat{\beta}-R\beta\right)'\left[\sigma^2\cdot R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta}-R\beta\right)}{\frac{s^2}{\sigma^2}}\\&=\frac{1}{k}\left(R\hat{\beta}-R\beta\right)'\left[s^2\cdot R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta}-R\beta\right)'\sim F_{k,T-k}.\end{aligned}$$

Thus, under the null H_0,

$$F=\frac{1}{k}\left(R\hat{\beta}-q\right)'\left[s^2\cdot R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta}-q\right)'\sim F_{k,T-k},$$

and we can use this to find the appropriate critical values given the significance level α.

5.1.6 Equivalent Test Statistics for Zero Restrictions

When the null hypothesis tests whether a certain subset of βs jointly equal 0, we can derive a familiar version of the F statistic in terms of the difference of restricted and unrestricted SSRs. As in the statement of the FWL theorem, let the data matrix X be partitioned as $X=(\mathbb{X}_1,\mathbb{X}_2)$, where $\mathbb{X}_i$ has dimension $T\times k_i$ for $i=1,2$, and let $\beta=(\beta_1',\beta_2')'$ be partitioned conformably.

Consider the null hypothesis $H_0:\beta_2=\mathbf{0}$. This is a special case of the linear restrictions that we considered above with $R=(O_{k_2\times k_1},I_{k_2})$ and $q=\mathbf{0}$. The F statistic is then given as

$$F=\frac{1}{k}\cdot\hat{\beta}_2'\left[s^2\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1}\right]^{-1}\hat{\beta}_2=\frac{\frac{1}{k}\hat{\beta}_2'\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)\hat{\beta}_2}{\frac{SSR}{T-k}},$$

where $M_1=I_T-\mathbb{X}_1(\mathbb{X}_1'\mathbb{X}_1)^{-1}\mathbb{X}_1'$ is the residual maker of the block $\mathbb{X}_1$ and $\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1}$ is the

lower $k_2 \times k_2$ block of the inverse matrix $(X'X)^{-1}$.

Let $\tilde{\beta}_1$ be the OLS estimator of β_1 under the null H_0, that is, the OLS estimator of β_1 obtained by regressing Y on $\mathbb{X}_1$ only. By definition,

$$\tilde{\beta}_1 = (\mathbb{X}_1'\mathbb{X}_1)^{-1}\mathbb{X}_1'Y,$$

and the corresponding restricted SSR is given as

$$SSR_r = (Y - \mathbb{X}_1\tilde{\beta}_1)'(Y - \mathbb{X}_1\tilde{\beta}_1) = Y'M_1Y.$$

Since

$$Y = \mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2 + \epsilon,$$

expanding the restricted SSR shows us that

$$\begin{aligned} SSR_r &= (\mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2 + \epsilon)' M_1 (\mathbb{X}_1\beta_1 + \mathbb{X}_2\beta_2 + \epsilon) \\ &= (\mathbb{X}_2\beta_2 + \epsilon)' M_1 (\mathbb{X}_2\beta_2 + \epsilon) \\ &= \beta_2'\mathbb{X}_2'M_1\mathbb{X}_2\beta_2 + 2\beta_2'\mathbb{X}_2'M_1\epsilon + \epsilon'M_1\epsilon. \end{aligned}$$

Using the formula for matrix inversion, we can see that

$$\begin{aligned} P_X &= \begin{pmatrix} \mathbb{X}_1 & \mathbb{X}_2 \end{pmatrix} \begin{pmatrix} \mathbb{X}_1'\mathbb{X}_1 & \mathbb{X}_1'\mathbb{X}_2 \\ \mathbb{X}_2'\mathbb{X}_1 & \mathbb{X}_2'\mathbb{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{X}_1' \\ \mathbb{X}_2' \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{X}_1 & \mathbb{X}_2 \end{pmatrix} \begin{pmatrix} (\mathbb{X}_1'\mathbb{X}_1)^{-1} + (\mathbb{X}_1'\mathbb{X}_1)^{-1}\mathbb{X}_1'\mathbb{X}_2\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1}\mathbb{X}_2'\mathbb{X}_1(\mathbb{X}_1'\mathbb{X}_1)^{-1} & -(\mathbb{X}_1'\mathbb{X}_1)^{-1}\mathbb{X}_1'\mathbb{X}_2\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1} \\ -\left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1}\mathbb{X}_2'\mathbb{X}_1(\mathbb{X}_1'\mathbb{X}_1)^{-1} & \left(\mathbb{X}_2'M_1\mathbb{X}_2\right)^{-1} \end{pmatrix} \begin{pmatrix} \mathbb{X}_1' \\ \mathbb{X}_2' \end{pmatrix} \\ &= P_1 + P_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'P_1 - P_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2' - \mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'P_1 + \mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2' \\ &= P_1 - P_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'M_1 + \mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'M_1 \\ &= P_1 + M_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'M_1. \end{aligned}$$

As such,

$$\begin{aligned} M_X &= I_T - P_X \\ &= M_1 - M_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'M_1, \end{aligned}$$

and it follows that the unrestricted SSR is

$$\begin{aligned} SSR &= \epsilon' M_X \epsilon \\ &= \epsilon'M_1\epsilon - \epsilon'M_1\mathbb{X}_2(\mathbb{X}_2'M_1\mathbb{X}_2)^{-1}\mathbb{X}_2'M_1\epsilon. \end{aligned}$$

Using this in tandem with the fact that

$$\hat{\beta}_2 = \beta_2 + (\mathbb{X}_2' M_1 \mathbb{X}_2)^{-1} \mathbb{X}_2' M_1 \epsilon,$$

we can see that

$$\begin{aligned}\hat{\beta}_2' \left(\mathbb{X}_2' M_1 \mathbb{X}_2\right) \hat{\beta}_2 &= \beta_2' \mathbb{X}_2' M_1 \mathbb{X}_2 \beta_2 + 2\epsilon' M_1 \mathbb{X}_2 \beta_2 + \epsilon' M_1 \mathbb{X}_2 (\mathbb{X}_2' M_1 \mathbb{X}_2)^{-1} \mathbb{X}_2' M_1 \epsilon \\ &= SSR_r - SSR.\end{aligned}$$

By implication, the F statistic can be written as

$$F = \frac{(SSR_r - SSR)/k}{SSR/(T-k)},$$

which is the alternative of the statistic taught in introductory econometrics courses.

5.1.7 Testing for Single Zero Restrictions

Another popular null hypothesis that we often test is the null $H_0 : \beta_i = 0$ for some $1 \leq i \leq k$, that is, the significance test for the ith independent variable. Note that

$$\hat{\beta}_i - \beta_i \mid X \sim \mathcal{N}\left[\mathbf{0}, \sigma^2 \cdot (\mathbb{X}_i' M_{-i} \mathbb{X}_i)^{-1}\right],$$

where $\mathbb{X}_i$ is the collection of the T sample observations of the ith variable, and M_{-i} is the residual maker associated with the remaining regressors. Thus, the Z statistic

$$Z_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 \cdot (\mathbb{X}_i' M_{-i} \mathbb{X}_i)^{-1}}}$$

is conditionally standard normal given X, and is conditionally independent of $V = \frac{(T-k) \cdot s^2}{\sigma^2}$ given X, which conditionally follows a chi squred distribution with $T-k$ degrees of freedom given X. Therefore,

$$T_i = \frac{Z_i}{\sqrt{V/(T-k)}} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 \cdot (\mathbb{X}_i' M_{-i} \mathbb{X}_i)^{-1}}} \mid X \sim t_{T-k}.$$

Since the distribution on the right hand side does not depend on X in any way, we can see that the unconditional distribution of T is also given as

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 \cdot (\mathbb{X}_i' M_{-i} \mathbb{X}_i)^{-1}}} \sim t_{T-k}.$$

We can use this t statistic to perform two-sided tests for the null $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$.

5.2 The Generalized Univariate Linear Regression Model

So far, we have studied the linear regression model under the traditional Gauss-Markov assumptions, which included the normality of the errors and homoskedasticity. These assumptions prove to be overly restrictive in practice, which leads us to study the properties of the OLS estiamtors of the model parameters even when these assumptions have been relaxed. It turns out that, even without the normality or spherical error assumptions, and indeed the assumption of identically distributed sample observations, the MLEs of β and σ^2 continue to possess desirable large sample properties. These properties are used to construct asymptotic test statistics, including White's famed heteroskedasticity-consistent covariance matrix estimator.

5.2.1 Model Assumptions

The assumptions of the model are now given in their more general forms below:

A1 : **The Sample**

There is a sample $\{(Y_t, X_t)\}_{t \in N_+}$, where each Y_t takes values in $\mathbb{R}$ and X_t in $\mathbb{R}^k$. The sample observations are independent but not identically distributed. The errors are defined as

$$\epsilon_t = Y_t - X_t'\beta$$

for any $t \in N_+$ for some parameter $\beta \in \mathbb{R}^k$.

A2 : **No Perfect Collinearity**

We assume that the $k \times k$ random matrix

$$\sum_{t=1}^{T} X_t X_t' = \begin{pmatrix} X_1 & \cdots & X_T \end{pmatrix} \begin{pmatrix} X_1' \\ \vdots \\ X_T' \end{pmatrix}$$

is nonsingular for any $T \geq k$ with probability 1.

A3 : **Exogeneity (Uncorrelatedness)**

We assume that the errors have mean 0, and that the tth error is uncorrelated with the tth regressors, that is, $\mathbb{E}[\epsilon_t X_t] = \mathbf{0}$. This represents a weaker version of the mean independence assumption from earlier.

A4 : **Asymptotic Behavior of the Data**

There exist a $k \times k$ positive definite matrices Q and Σ such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q, \quad \frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 X_t X_t' \xrightarrow{p} \Sigma.$$

A5 : **Boundedness of Moments**

There exists a $\delta > 0$ such that

$$\sup_{t \in N_+} \mathbb{E}|\epsilon_t X_t|^{2+\delta} < +\infty.$$

In addition, $\{X_t\}_{t \in N_+}$ has finite fourth moments:

$$\sup_{t \in N_+} \mathbb{E}|X_t|^4 < +\infty.$$

We comment on each of the new assumptions. For one, we relax the mean independence assumption in the normal model earlier to the requirement that the errors and regressors are only uncorrelated. The assumption that the sample averages $\frac{1}{T}\sum_{t=1}^{T} X_t X_t'$ and $\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 X_t X_t'$ converge in probability to positive definite matrices is standard, and holds automatically by Khinchin's WLLN if the data are i.i.d., and the means $\mathbb{E}[X_t X_t']$ and $\mathbb{E}[\epsilon_t^2 X_t X_t']$ are positive definite. In particular, if the data are i.i.d. and the errors are homoskedastic, that is, if $\mathbb{E}[\epsilon_t^2 \mid X_t] = \mathbb{E}[\epsilon_t^2] = \sigma^2$ for any $t \in N_+$, then

$$\mathbb{E}\left[\epsilon_t^2 X_t X_t'\right] = \sigma^2 \mathbb{E}[X_t X_t']$$

and therefore $\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 X_t X_t'$ converges in probability to $\sigma^2 Q$.

The boundedness assumption is also standard in the literature. The first assumption can be written alternatively as

$$\sup_{t \in N_+} \mathbb{E}\left|\epsilon_t^2 X_t' X_t\right|^{1+\frac{\delta}{2}} < +\infty,$$

which is exactly the version of the assumption that appears in White's 1980 paper. In the special case that the data are i.i.d., this implies that $\epsilon_t^2 X_t' X_t$ has finite moments of order slightly higher than 2. The assumption that the regressors have finite fourth moments is also found often in the literature; in the special case that the data is i.i.d. and the errors are homoskedastic in the sense above, the assumptions $\mathbb{E}|X_t|^4 < +\infty$ and $\mathbb{E}[\epsilon_t^4 \mid X_t] = \sigma_4 < +\infty$ are sufficient for $\mathbb{E}|\epsilon_t X_t|^4 < +\infty$, since

$$\mathbb{E}|\epsilon_t X_t|^4 = \mathbb{E}\left[\mathbb{E}\left[\epsilon_t^4 \mid X_t\right] \cdot |X_t|^4\right] = \sigma_4 \mathbb{E}|X_t|^4 < +\infty.$$

5.2.2 Asymptotic Properties of the OLS Estimator

Fortunately, under assumptions A1 to A5, the OLS estimator $\hat{\beta}_T$ possesses desirable large sample properties; less fortunately, it is unclear now what the estimator $\hat{\sigma}_T^2$ is estimating, since we dispensed with the homoskedastic environment that we worked with above. Recall that

$$\hat{\beta}_T = \left(\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\sum_{t=1}^{T} X_t(X_t'\beta + \epsilon_t)\right) = \beta + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t \epsilon_t\right).$$

This expression allows us to show that $\hat{\beta}_T$ is consistent and asymptotically normal:

Theorem 4.4 (Asymptotic Properties of OLS Estimator)
Under assumptions A1 to A5,

$$\hat{\beta}_T \xrightarrow{p} \beta$$

and

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, Q^{-1}\Sigma Q^{-1}\right].$$

Proof) First, for consistency, note that

$$\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \xrightarrow{p} Q^{-1}$$

by assumption A4 and the continuous mapping theorem. On the other hand, $\{\epsilon_t X_t\}_{t\in N_+}$ is a sequence of independent k-dimensional random vectors with common mean 0 (assumption A3) and bounded $2+\delta$-moments (assumption A5). By Chebyshev's WLLN, which holds for uncorrelated sequences of random variables with bounded second moments,

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t X_t \xrightarrow{p} \mathbf{0}.$$

Therefore, we can see that

$$\hat{\beta}_T = \beta + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t \epsilon_t\right) \xrightarrow{p} \beta + Q^{-1}\mathbf{0} = \beta$$

as $T \to \infty$.

For asymptotic normality, we can write

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t \epsilon_t\right).$$

Note that $\{\epsilon_t X_t\}_{t\in N_+}$ is an independent sequence with common mean 0, and therefore a martingale difference sequence with respect to the filtration that it generates. Furthermore, it is L^p-bounded for $p = 2+\delta > 2$, and

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 X_t X_t' \xrightarrow{p} \Sigma$$

by assumption, where Σ is positive definite. Therefore, by the martingale difference sequence CLT,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t \epsilon_t \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \Sigma\right].$$

By Slutsky's theorem, we can now see that

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] \xrightarrow{d} Q^{-1}\mathcal{N}\left[\mathbf{0}, \Sigma\right] = \mathcal{N}\left[\mathbf{0}, Q^{-1}\Sigma Q^{-1}\right].$$

Q.E.D.

5.2.3 Heteroskedasticity Consistent Covariance Matrix Estimation

For statistical inference to be possible at all, we must be able to consistently estimate the asymptotic variance $Q^{-1}\Sigma Q^{-1}$. Q is by definition the probability limit of the average data matrix, so our main concern will be Σ. By definition, this is the probability limit of the sample average

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 X_t X_t'.$$

Here, the only unknown is the error ϵ_t, so White, in his 1980 paper, suggests replacing ϵ_t with the residual $\hat{\epsilon}_{t,T} = Y_t - X_t'\hat{\beta}_T$. White's estimator for Σ is given as

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{t,T}^2 X_t X_t' = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - X_t'\hat{\beta}_T\right)^2 X_t X_t'.$$

Under our assumptions, White's estimator consistently estimates Σ:

Theorem 4.5 (Consistency of White Estimator)
Under assumptions A1 to A5,

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{t,T}^2 X_t X_t' \xrightarrow{p} \Sigma.$$

Proof) The proof is simple if we just expand $\hat{\Sigma}_T$ out as

$$\begin{aligned}
\hat{\Sigma}_T &= \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - X_t'\hat{\beta}_T\right)^2 X_t X_t' \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[X_t'(\beta - \hat{\beta}_T) + \epsilon_t\right]^2 X_t X_t' \\
&= \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left[(\beta - \hat{\beta}_T)' X_t X_t' (\beta - \hat{\beta}_T)\right] X_t X_t'}_{I} \\
&\quad + \underbrace{2\frac{1}{T}\sum_{t=1}^{T}\left[(\beta - \hat{\beta}_T)' X_t \epsilon_t\right] X_t X_t'}_{II} + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 X_t X_t'}_{III}.
\end{aligned}$$

Since term III goes to Σ in probability by assumption, we need only show that terms I and II are $o_p(1)$.

Focusing first on the term I, we start by searching for an upper bound to the trace

norm of I:

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_tX_t'(\beta-\hat{\beta}_T)\right]X_tX_t'\right\|^2 = \sum_{j=1}^{k}\sum_{m=1}^{k}\left|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_tX_t'(\beta-\hat{\beta}_T)\right]X_{tj}X_{tm}\right|^2$$

$$= \sum_{j=1}^{k}\sum_{m=1}^{k}\left|(\beta-\hat{\beta}_T)'\left[\frac{1}{T}\sum_{t=1}^{T}X_tX_t'X_{tj}X_{tm}\right](\beta-\hat{\beta}_T)\right|^2$$

$$\leq \left|\hat{\beta}_T-\beta\right|^4\cdot\sum_{j=1}^{k}\sum_{m=1}^{k}\left\|\frac{1}{T}\sum_{t=1}^{T}X_tX_t'X_{tj}X_{tm}\right\|^2$$

$$= \left|\hat{\beta}_T-\beta\right|^4\cdot\sum_{j=1}^{k}\sum_{m=1}^{k}\sum_{l=1}^{k}\sum_{p=1}^{k}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp}\right|^2,$$

where we used the definition of the trace norm to justify the first and last equalities, and the trace norm inequality $|Ax| \leq \|A\| \cdot |x|$ for any matrix A and conformable vector x to justify the inequality. For any fixed $1 \leq j,m,l,p \leq k$, repeated applications of Hölder's inequality show us that

$$\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp}\right| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_{tj}X_{tm}X_{tl}X_{tp}|$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}|X_{tj}|^4\right)^{\frac{1}{4}}\left(\mathbb{E}|X_{tm}|^4\right)^{\frac{1}{4}}\left(\mathbb{E}|X_{tl}|^4\right)^{\frac{1}{4}}\left(\mathbb{E}|X_{tp}|^4\right)^{\frac{1}{4}}.$$

Since

$$\mathbb{E}|X_{tj}|^4 \leq \mathbb{E}|X_t|^4$$

for any $1 \leq j \leq k$, this implies that

$$\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp}\right| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_t|^4.$$

Finally, since the fourth moments of X are bounded,

$$\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp}\right| \leq \sup_{t\in N_+}\mathbb{E}|X_t|^4 < +\infty$$

for any $T \in N_+$. This tells us that

$$\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp} = O_p(1) \quad \text{as } T\to\infty$$

for any fixed $1 \leq j,m,l,p \leq k$. Since the (finite) product and sum of $O_p(1)$ processes is also $O_p(1)$, and $\hat{\beta}_T - \beta = o_p(1)$ as $T \to \infty$ by the consistency of $\hat{\beta}_T$ for β, we can

conclude that

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_tX_t'(\beta-\hat{\beta}_T)\right]X_tX_t'\right\|^2 \leq \underbrace{\left|\hat{\beta}_T-\beta\right|^4}_{o_p(1)}\cdot\underbrace{\sum_{j=1}^{k}\sum_{m=1}^{k}\sum_{l=1}^{k}\sum_{p=1}^{k}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_{tl}X_{tp}\right|^2}_{O_p(1)} = o_p(1).$$

As such,

$$\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_tX_t'(\beta-\hat{\beta}_T)\right]X_tX_t' = o_p(1) \quad \text{as } T\to\infty.$$

Moving onto the term II, we similarly start by bounding the trace norm of II:

$$\begin{aligned}
\left\|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_t\epsilon_t\right]X_tX_t'\right\|^2 &= \sum_{j=1}^{k}\sum_{m=1}^{k}\left|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_t\epsilon_t\right]X_{tj}X_{tm}\right|^2 \\
&= \sum_{j=1}^{k}\sum_{m=1}^{k}\left|(\beta-\hat{\beta}_T)'\left[\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_t\epsilon_t\right]\right|^2 \\
&\leq \left|\hat{\beta}_T-\beta\right|^2\cdot\sum_{j=1}^{k}\sum_{m=1}^{k}\left|\frac{1}{T}\sum_{t=1}^{T}X_{tj}X_{tm}X_t\epsilon_t\right|^2 \\
&\leq \left|\hat{\beta}_T-\beta\right|^2\cdot\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\frac{1}{T}\sum_{t=1}^{T}|X_{tj}X_{tm}|\cdot|X_t\epsilon_t|\right)^2 \\
&\leq \left|\hat{\beta}_T-\beta\right|^2\cdot\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\frac{1}{T}\sum_{t=1}^{T}|X_t\epsilon_t|^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}(X_{tj}X_{tm})^2\right) \\
&= \left|\hat{\beta}_T-\beta\right|^2\cdot\left(\frac{1}{T}\sum_{t=1}^{T}|X_t\epsilon_t|^2\right)\left[\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\frac{1}{T}\sum_{t=1}^{T}(X_{tj}X_{tm})^2\right)\right],
\end{aligned}$$

where the first equality follows from the definition of the trace norm, the first inequality follows from the Cauchy-Schwarz inequality for the dot product on Euclidean space, the second inequality from the triangle inequality, and the last inequality from the Cauchy-Schwarz inequality for finite sums. Note that

$$\frac{1}{T}\sum_{t=1}^{T}|X_t\epsilon_t|^2 = \operatorname{tr}\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2X_tX_t'\right)\xrightarrow{p}\operatorname{tr}(\Sigma),$$

so that $\frac{1}{T}\sum_{t=1}^{T}|X_t\epsilon_t|^2 = O_p(1)$ (since convergence in probability implies convergence in distribution and sequences of random variables that converge in distribution are $O_p(1)$).

On the other hand, for any fixed $1 \leq j,m \leq k$,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}(X_{tj}X_{tm})^2\right| &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_{tj}X_{tm}|^2 \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}|X_{tj}|^4\right)^{\frac{1}{2}}\left(\mathbb{E}|X_{tm}|^4\right)^{\frac{1}{2}} \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_t| \leq \sup_{t\in N_+}\mathbb{E}|X_t|^4 < +\infty\end{aligned}$$

for any $T \in N_+$, so that $\frac{1}{T}\sum_{t=1}^{T}(X_{tj}X_{tm})^2 = O_p(1)$. This holds for any fixed $1 \leq j,m \leq k$, so we have

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_t\epsilon_t\right]X_tX_t'\right\|^2 \leq \underbrace{\left|\hat{\beta}_T-\beta\right|^2}_{o_p(1)}\cdot\underbrace{\left(\frac{1}{T}\sum_{t=1}^{T}|X_t\epsilon_t|^2\right)}_{O_p(1)}\underbrace{\left[\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\frac{1}{T}\sum_{t=1}^{T}(X_{tj}X_{tm})^2\right)\right]}_{O_p(1)} = o_p(1),$$

and therefore

$$\frac{1}{T}\sum_{t=1}^{T}\left[(\beta-\hat{\beta}_T)'X_t\epsilon_t\right]X_tX_t' = o_p(1) \quad \text{as } T\to\infty.$$

We have shown that terms I and II are both $o_p(1)$ as $T \to \infty$. Therefore,

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2X_tX_t' + o_p(1) \xrightarrow{p} \Sigma.$$

This completes the proof.

Q.E.D.

Thus, a consistent estimator of the asymptotic variance $Q^{-1}\Sigma Q^{-1}$ is given by the so-called "sandwich" estimator:

$$\hat{V}_T = \left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_{t,T}^2X_tX_t'\right]\left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)^{-1}.$$

This can be used to form tests for general non-linear null hypotheses.

5.2.4 Testing for General Non-linear Hypotheses for β

Consider a null hypothesis $H_0 : g(\beta) = q$, where $g : \mathbb{R}^k \to \mathbb{R}^r$ is a continuously differentiable function with derivative $Dg(\beta) \in \mathbb{R}^{r \times k}$ of full rank $r \leq k$, and $q \in \mathbb{R}^r$. Then, under the null H_0, by the Delta method

$$\sqrt{T}\left[g(\hat{\beta}_T) - q\right] = \sqrt{T}\left[g(\hat{\beta}_T) - g(\beta)\right] \xrightarrow{d} Z_g \sim \mathcal{N}\left[\mathbf{0}, Dg(\beta)Q^{-1}\Sigma Q^{-1}Dg(\beta)'\right].$$

By the continuity of the derivative $Dg : \mathbb{R}^k \to \mathbb{R}^{r \times k}$ and the continuous mapping theorem,

$$Dg(\hat{\beta}_T) \xrightarrow{p} Dg(\beta),$$

so that

$$Dg(\hat{\beta}_T)\hat{V}_T Dg(\hat{\beta}_T)' \xrightarrow{p} Dg(\beta)Q^{-1}\Sigma Q^{-1}Dg(\beta)',$$

where the limit is an $r \times r$ positive definite matrix. It follows that

$$\begin{aligned} W_T &= T(g(\hat{\beta}_T) - q)'\left[Dg(\hat{\beta}_T)\hat{V}_T Dg(\hat{\beta}_T)'\right]^{-1}(g(\hat{\beta}_T) - q) \\ &\xrightarrow{d} Z_g'\left(Dg(\beta)Q^{-1}\Sigma Q^{-1}Dg(\beta)'\right)^{-1} Z_g \sim \chi^2_r. \end{aligned}$$

We can therefore use this asymptotic Wald statistic

$$W_T = \left(g(\hat{\beta}_T) - q\right)'\left[Dg(\hat{\beta}_T) \cdot \left(\frac{1}{T}\hat{V}_T\right) Dg(\hat{\beta}_T)'\right]^{-1}\left(g(\hat{\beta}_T) - q\right)$$

to test for $H_0 : g(\beta) = q$ against $H_1 : g(\beta) \neq q$.

5.3 A Bayesian Approach to Univariate Normal Linear Regression

So far, we have discussed frequentist approaches, both under normality and under non-normality, derived finite or large sampe properties of the OLS estimators of β, and used these results to formulate exact or asymptotic tests for hypotheses on β. An alternate way to think of the model is by taking a Bayesian approach. The advantages of the Bayesian approach over the frequentist approach to a non-normal linear regression model is that we do not need to impose independence or stationarity conditions on the sample observations in the Bayesian approach, while these are required to derive large sample properties in the frequentist case[1]. The downside is that we do need to specify the distribution of the errors, so that the model becomes fully parametric, as opposed to the semi-parametric flavor of the generalized model.

5.3.1 Model Assumptions

We make the following assumptions, which may be better understood in this context as the model specification.

A1 : **Data Generating Process**

Let there be a sample $\{(Y_t, X_t)\}_{1 \leq t \leq T}$ of size T, where X_t may include lagged values of Y_t. We do not impose any restriction on the dependence structure on the data, and instead assume that the following distributional assumption holds:

$$Y \mid X_0, \beta, \sigma^2 \sim \mathcal{N}\left[X\beta,\ \sigma^2 \cdot I_n\right],$$

where X_0 collects the elements of $X_1, \cdots, X_T$ that do not appear in Y. This type of DGP includes the basline i.i.d. case with $X_0 = X$, but also encompasses more general models such as pth order autoregressive processes with $X_0 = (Y_0, \cdots, Y_{-p+1})$ collecting the initial values and $X_t = (Y_{t-1}, \cdots, Y_{t-p})'$ for $1 \leq t \leq T$.

Furthermore, we assume that X_0 is conditionally independent of the parameters β, σ^2, which are now random variables and not assumed fixed.

A2 : **Prior Specification**

We let the prior distributions of β and σ^2 be given as

$$\beta \mid \sigma^2 \sim \mathcal{N}\left[b_0, B_0\right],$$

[1] Of course, to speak of Bayesian asymptotics, we do also need to assume these restrictions on the dependence structure of the sample, but here we are only interested in finite sample Bayesian inference for now.

where the prior hyperparameters b_0, B_0 may or may not depend on σ^2, and

$$\sigma^2 \sim \mathcal{IG}\left[\frac{\alpha_0}{2}, \frac{\delta_0}{2}\right],$$

where $\mathcal{IG}$ represents the inverse gamma distribution. B_0 is a positive definite matrix given σ^2.

We also assume that the usual regularity assumptions for Bayesian inference, such as those introduced in the last chapter, hold true, so that we can use Bayes' rule to obtain the posterior density. Under the above DGP, the likelihood of Y, X_0 given the parameters β, σ^2 becomes

$$\begin{aligned} f_{Y,X_0}(Y, X_0 \mid \beta, \sigma^2) &= f_{Y|X_0}(Y \mid X_0, \beta, \sigma^2) \cdot f_{X_0}(X_0) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right) \cdot f_{X_0}(X_0). \end{aligned}$$

5.3.2 Full Conditional Distributions of β and σ^2

In the process of computing the posteriors of β and σ^2, it is often useful to first obtain the full conditional distributions of these parameters, namely, the conditional distribution of the parameters given all the data and the other parameter. Formally, the distributions $\beta \mid Y, X_0, \sigma^2$ and $\sigma^2 \mid Y, X_0, \beta$ are the full conditional distributions of β and σ^2. These can be used in an iterative process to sample from the posterior distribution of the parameters.

Full Conditional Distribution of β

Using Bayes' rule, we can obtain the kernel of the full conditional density of β as follows:

$$\begin{aligned} \pi(\beta \mid Y, X_0, \sigma^2) &\propto f_{Y,X_0}(Y, X_0 \mid \beta, \sigma^2) p(\beta \mid \sigma^2) \\ &\propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right)}_{\text{Likelihood}} \cdot \underbrace{\exp\left(-\frac{1}{2}(\beta - b_0)' B_0^{-1} (\beta - b_0)\right)}_{\text{Prior}} \\ &\propto \exp\left(-\frac{1}{2}\left[\beta'\left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)\beta - 2\beta'\left(\frac{1}{\sigma^2}X'Y + B_0^{-1}b_0\right)\right]\right). \end{aligned}$$

The last term ends up being proportional to the following expression:

$$\pi(\beta \mid Y, X_0, \sigma^2) \propto \exp\left(-\frac{1}{2}\left[\beta - \left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X'Y + B_0^{-1}b_0\right)\right]'\left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)\left[\beta - \left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X'Y + B_0^{-1}b_0\right)\right]\right).$$

This is the kernel of a normal distribution with mean

$$b_1(Y, X, \sigma^2) = \left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}X'Y + B_0^{-1}b_0\right)$$

and variance

$$B_1(Y, X, \sigma^2) = \left(\frac{1}{\sigma^2}X'X + B_0^{-1}\right)^{-1},$$

so it follows that

$$\beta \mid Y, X_0, \sigma^2 \sim \mathcal{N}\left[b_1(Y, X, \sigma^2), B_1(Y, X, \sigma^2)\right].$$

Full Conditional Distribution of σ^2

Again, the kernel of the full conditional density of σ^2 can be obtained using Bayes' rule as:

$$\begin{aligned}
\pi(\sigma^2 \mid Y, X_0, \beta) &\propto f_{Y,X_0}(Y, X_0 \mid \beta, \sigma^2) \cdot p(\beta \mid \sigma^2) p(\sigma^2 \mid \beta) \\
&= f_{Y,X_0}(Y, X_0 \mid \beta, \sigma^2) \cdot p(\sigma^2) \\
&\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right)}_{\text{Likelihood}} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right)}_{\text{Prior}} \\
&= \left(\frac{1}{\sigma^2}\right)^{\frac{T+\alpha_0}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\delta_0 + (Y - X\beta)'(Y - X\beta)\right]\right).
\end{aligned}$$

The right hand side is the kernel of an inverse gamma distribution, so

$$\sigma^2 \mid Y, X_0, \beta \sim \mathcal{IG}\left[\frac{T + \alpha_0}{2}, \frac{\delta_0 + (Y - X\beta)'(Y - X\beta)}{2}\right].$$

5.3.3 Sampling from the Posterior Distribution

In general, even though we can obtain a closed form expression for the full conditional distribution of β and σ^2, the posterior of β and σ^2, that is, the distributions $\beta \mid Y, X_0$ and $\sigma^2 \mid Y, X_0$, are unknown distributions. Therefore, to sample from $\beta, \sigma^2 \mid Y, X_0$, we must rely on Monte Carlo Markov Chain (MCMC) methods.

One useful MCMC method that works in this case, and indeed in any case where the full conditional distributions of the parameters are known, is Gibbs sampling. The algorithm proceeds as follows:

5.3.4 The Posterior Density for a Special Case

It turns out that for the prior

$$\beta \mid \sigma^2 \sim \mathcal{N}\left[\mathbf{0}, \sigma^2 B_0\right],$$

where B_0 is a positive definite matrix, the posterior density for β and σ^2 given Y, X_0 has an analytical form as a multivariate t distribution. Let $p(\beta \mid \sigma^2)$ and $p(\sigma^2)$ represent the prior densities of β and σ^2.

Bayes' rule tells us that the posterior density of (β, σ^2) given (Y, X) can be obtained as

$$\begin{aligned}
\pi(\beta, \sigma^2 \mid Y, X_0) &\propto f_{Y,X_0}(Y, X_0 \mid \beta, \sigma^2) p(\beta \mid \sigma^2) p(\sigma^2) \\
&\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{T}{2}} \exp\left(-\frac{1}{\sigma^2}(Y - X\beta)'(Y - X\beta)\right)}_{\text{Likelihood}} \cdot \underbrace{\left(\sigma^2\right)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}\beta' B_0^{-1}\beta\right)}_{\text{Prior for } \beta} \\
&\quad \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\alpha_0/2+1} \exp\left(-\frac{\delta_0}{2\sigma^2}\right)}_{\text{Prior for } \sigma^2} \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{T+k+\alpha_0}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\delta_0 + (\beta - b_1(Y,X))' B_1(Y,X)^{-1}(\beta - b_1(Y,X))\right]\right) \\
&\quad \times \exp\left(-\frac{1}{2\sigma^2}\left[Y'Y - b_1(Y,X)' B_1(Y,X)^{-1} b_1(Y,X)\right]\right) \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{T+k+\alpha_0}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\delta_1(Y,X) + (\beta - b_1(Y,X))' B_1(Y,X)^{-1}(\beta - b_1(Y,X))\right]\right),
\end{aligned}$$

where

$$\begin{aligned}
B_1(Y,X) &= (X'X + B_0^{-1})^{-1} \\
b_1(Y,X) &= (X'X + B_0^{-1})^{-1} X'Y, \\
\delta_1(Y,X) &= \delta_0 + Y'Y - b_1(Y,X)' B_1(Y,X)^{-1} b_1(Y,X).
\end{aligned}$$

Posterior of β

The posterior density of β given Y, X_0 can now be obtained as

$$\pi(\beta \mid Y, X_0) = \int_0^\infty \pi(\beta, \sigma^2 \mid Y, X_0) d\sigma^2$$

$$\propto \int_0^\infty \left(\frac{1}{\sigma^2}\right)^{\frac{T+k+\alpha_0}{2}+1} \exp\left(-\frac{1}{2\sigma^2}\left[\delta_1(Y,X)+(\beta-b_1(Y,X))'B_1(Y,X)^{-1}(\beta-b_1(Y,X))\right]\right)d\sigma^2.$$

The integrand is the kernel of a inverse gamma distribution with parameters $\frac{T+k+\alpha_0}{2}$ and $\frac{\delta_1(Y,X)+(\beta-b_1(Y,X))'B_1(Y,X)^{-1}(\beta-b_1(Y,X))}{2}$, so that

$$\begin{aligned}
\pi(\beta \mid Y, X_0) &\propto \left[\frac{(\delta_1(Y,X)+(\beta-b_1(Y,X))'B_1(Y,X)^{-1}(\beta-b_1(Y,X)))^{(T+k+\alpha_0)/2}}{2^{(T+k+\alpha_0)/2}\Gamma((T+k+\alpha_0)/2)}\right]^{-1} \\
&\propto \left[\delta_1(Y,X)+(\beta-b_1(Y,X))'B_1(Y,X)^{-1}(\beta-b_1(Y,X))\right]^{-\frac{T+\alpha_0}{2}-\frac{k}{2}} \\
&= \left[1+\frac{1}{T+\alpha_0}(\beta-b_1(Y,X))'\left(\frac{\delta_1(Y,X)}{T+\alpha_0}B_1(Y,X)\right)^{-1}(\beta-b_1(Y,X))\right]^{-\frac{T+\alpha_0}{2}-\frac{k}{2}}.
\end{aligned}$$

This is the kernel of a k-dimensional multivariate t distribution with location vector $b_1(Y,X)$, scale matrix $\frac{\delta_1(Y,X)}{T+\alpha_0}B_1(Y,X)$ and $T+\alpha_0$ degrees of freedom, so that

$$\beta \mid Y, X_0 \sim t_{T+\alpha_0}\left[b_1(Y,X),\ \frac{\delta_1(Y,X)}{T+\alpha_0}B_1(Y,X)\right].$$

The posterior mean of β given Y, X_0 is surprisingly easy to find, since the full conditional posterior mean of β does not depend on σ^2 if the prior mean b_0 is 0:

$$\begin{aligned}
\mathbb{E}[\beta \mid Y, X_0] &= \mathbb{E}\left[\mathbb{E}\left[\beta \mid Y, X_0, \sigma^2\right] \mid Y, X_0\right] \\
&= (X'X + B_0^{-1})^{-1}X'Y = b_1(Y,X).
\end{aligned}$$

Posterior of σ^2

The posterior density of σ^2 given Y, X_0 can be found similarly to that of β:

$$\begin{aligned}
\pi(\sigma^2 \mid Y, X_0) &= \int_{\mathbb{R}^k} \pi(\beta, \sigma^2 \mid Y, X_0)d\beta \\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{T_k+\alpha_0}{2}+1} \exp\left(-\frac{\delta_1(Y,X)}{2\sigma^2}\right) \\
&\quad \times \int_{\mathbb{R}^k} \exp\left(-\frac{1}{2\sigma^2}(\beta-b_1(Y,X))'B_1(Y,X)^{-1}(\beta-b_1(Y,X))\right)d\beta \\
&= \left(\frac{1}{\sigma^2}\right)^{\frac{T+k+\alpha_0}{2}+1} \exp\left(-\frac{\delta_1(Y,X)}{2\sigma^2}\right) \cdot (2\pi)^{\frac{k}{2}}\sqrt{\det(\sigma^2 B_1(Y,X))}
\end{aligned}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{T+\alpha_0}{2}+1} \exp\left(-\frac{\delta_1(Y,X)}{2\sigma^2}\right),$$

where we use the formula for the normalizing constant of a multivariate normal distribution to evaluate the integral. The rightmost term is the kernel of an inverse gamma distribution, so

$$\sigma^2 \mid Y, X_0 \sim \mathcal{IG}\left[\frac{T+\alpha_0}{2},\ \frac{\delta_1(Y,X)}{2}\right].$$

The posterior mean of σ^2 given Y, X_0 can now be obtained via the formula for the mean of an inverse gamma random variable:

$$\mathbb{E}\left[\sigma^2 \mid Y, X_0\right] = \frac{\delta_1(Y,X)}{T+\alpha_0-2}.$$

Chapter 6

Endogeneity and GMM

In the preceding model, we dealt with models in which the regressors were exogenous, that is, uncorrelated with the errors. Heuristically, the assumption that the regressors are exogenous is tantamount to equivalent to assuming that we can "keep all other things equal" when studying the effect of a certain independent variable on the dependent variable. This, however, requires that all relevant variables be included in the vector X_i of the regressors, which is generally not possible. There are a few reasons why we cannot control for all relevant variables in general:

- **Omitted Variables**
  It might not be possible to include some relevant variables that we want to control for in X_i, for instance if these variables are unobservable. The classical example of a model with omitted variables is a Mincer-type regression in which we want to investigate the effect of eductation on income. In this case, it is important that we control for each individual's ability, but ability, being unobservable, cannot be included as a regressor in X_i.

- **Measurement Errors**
  Some variables included in X_i might be measured with error, which can also skew our estimators of the partial effects in which we are interested. For instance, returning to the example above, we might want to use IQ as a proxy for ability; but because IQ measures individuals' abilities with some error, we are still unable to fully control for ability.

- **Simultaneity**
  If the dependent variable and independent variable are determined simultaneously, then we cannot properly estimate the effect of the latter on the former for a variety of reasons, one of which is that changes in the dependent variable might also give rise to changes in the independent variable (this is called a problem of reverse causality). The most famous example of simultaneity in economics is the classical problem of estimating supply and demand, which are determined simultaneously at market equilibrium.

6.1 An Illustrative Example

To see what happens when the exogeneity assumption breaks down, consider an i.i.d. sample $\{(Y_t, X_t)\}_{t \in N_+}$, where $\epsilon_t = Y_t - X_t'\beta$ is a mean zero error term for any $t \in N_+$ and some fixed parameter $\beta \in \mathbb{R}^k$. As before, let

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q$$

for some positive definite $k \times k$ matrix Q, so that $\frac{1}{T}\sum_{t=1}^{T} X_t X_t'$ is positive definite with probability approaching 1 as $T \to \infty$.

Suppose in this case that the exogeneity assumption $\mathbb{E}[\epsilon_t X_t] = \mathbf{0}$ breaks down; specifically, assume that

$$\mathbb{E}[\epsilon_t X_t] = \gamma$$

for some $\gamma \neq \mathbf{0}$. The OLS estimator in this case continues to be defined as

$$\hat{\beta}_T = \left(\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\sum_{t=1}^{T} X_t Y_t\right),$$

and using the fact that $Y_t = X_t'\beta + \epsilon_t$ for any $t \in N_+$, it can be expanded as

$$\hat{\beta}_T = \beta + \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t \epsilon_t\right).$$

By the continuous mapping theorem,

$$\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \xrightarrow{p} Q^{-1},$$

and by Khinchin's WLLN,

$$\frac{1}{T}\sum_{t=1}^{T} X_t \epsilon_t \xrightarrow{p} \mathbb{E}[\epsilon_t X_t] = \gamma.$$

Therefore, we can see that

$$\hat{\beta}_T - \beta \xrightarrow{p} Q^{-1}\gamma,$$

where $Q^{-1}\gamma \neq \mathbf{0}$ because γ is non-zero. This means that the OLS estimator converges to a vector other than the true parameter β, or in other words, that the OLS estimator is inconsistent, or asymptotically biased, for β.

Investigating the asymptotic bias $Q^{-1}\gamma$ in more depth, note that it can be written as

$$\mathbb{E}\left[X_t X_t'\right]^{-1} \mathbb{E}\left[X_t \epsilon_t\right].$$

If we assume that X_t has mean 0, the asymptotic bias becomes

$$\mathrm{Var}\,[X_t]^{-1}\,\mathrm{Cov}\,[X_t,\,\epsilon_t]\,.$$

Heuristically, when X_t is correlated with ϵ_t, then we cannot "keep ϵ_t fixed" when studying changes in X_t, so that the OLS estimator $\hat{\beta}_T$ ends up capturing not only the direct effect of X_t on Y_t, but also the indirect effect that X_t has on Y_t through changes in ϵ_t. It is this indirect effect that is reflected in the asymptotic bias term above.

So given that the regressors are endogenous, that is, that they violate the exogeneity assumption, how should we estimate β? By far the most popular way of dealing with endogeneity is to use intrumental variables, that is, variables Z_t that are uncorrelated with the error term ϵ_t but related to the original regressors X_t. The intuition for using instrumental variables is that, since the variables in Z_t are chosen to be uncorrelated with ϵ_t, we might be able to capture at least part of the direct effect of X_t on Y_t if we investigate the effect that X_t has on Y_t through Z_t. In other words, if we restrict our attention to changes in Y_t in response to changes in X_t that also shift Z_t, then the fact that Z_t is uncorrelated with ϵ_t might allow us to isolate the changes in Y_t that X_t is directly responsible for, while keeping ϵ_t fixed.

Below, we formalize this intuition by presenting an efficient GMM estimator of β using the instruments Z_t. We study its basic asymptotic properties, and then conclude the chapter by presenting a couple of specification tests that we can perform to help us decide whether we are using the right instruments or whether to use OLS or GMM.

6.2 GMM Estimation of β

We present here the instrumental variable approach to linear regression, which is the most popular means of overcoming the endogeneity problem.

6.2.1 Model Assumptions

We make the following assumptions:

A1 : **The Sample**

There is a sample $\{(Y_t, X_t, Z_t)\}_{t\in N_+}$, where each Y_t takes values in $\mathbb{R}$, X_t in $\mathbb{R}^k$, and Z_t in $\mathbb{R}^L$. The sample observations are independent but not identically distributed. The errors are defined as

$$\epsilon_t = Y_t - X_t'\beta$$

for any $t \in N_+$ for some parameter $\beta \in \mathbb{R}^k$. We assume that the errors have mean 0.

A2 : **No Perfect Collinearity**

We assume that there exist a $k \times k$ positive definite matrix Q_{xx} and an $L \times L$ positive definite matrix Q_{zz} such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q_{xx}, \quad \frac{1}{T}\sum_{t=1}^{T} Z_t Z_t' \xrightarrow{p} Q_{zz}.$$

By implication, $\frac{1}{T}\sum_{t=1}^{T} X_t X_t'$ and $\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$ are positive definite with probability approaching 1 as $T \to \infty$.

A3 : **Exogeneity of Instruments**

We assume the instruments Z_t are uncorrelated with the errors, that is,

$$\mathbb{E}\left[\epsilon_t Z_t\right] = \mathbf{0}$$

for any $t \in N_+$.

A4 : **Relevance of Instruments**

We assume that there exists an $L \times k$ matrix Q_{zx} of full rank k such that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t X_t' \xrightarrow{p} Q_{zx}.$$

This implies that $k \leq L$ and that $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ has full rank k with probability approaching 1 as $T \to \infty$.

A5 : **Boundedness of Moments**

There exists a $\delta > 0$ such that

$$\sup_{t \in N_+} \mathbb{E}|\epsilon_t Z_t|^{2+\delta} < +\infty.$$

In addition, $\{X_t\}_{t \in N_+}$ and $\{Z_t\}_{t \in N_+}$ have finite fourth moments:

$$\sup_{t \in N_+} \mathbb{E}|X_t|^4 < +\infty, \quad \sup_{t \in N_+} \mathbb{E}|Z_t|^4 < +\infty.$$

A GMM estimator of β is an estimator of β based on the moment condition

$$\mathbb{E}[\epsilon_t Z_t] = \mathbb{E}\left[Z_t(Y_t - X_t'\beta)\right] = \mathbf{0}.$$

We can then think of a sequence of weighting matrices that yields the smallest asympotic variance; the GMM estiamtor of β under this optimal sequence of weighting matrices is called the efficient GMM estimator. As before, all statements below hold with probabiltiy approaching 1 as $T \to \infty$.

6.2.2 Setting up the GMM Problem

Here we set up the GMM objective function under an appropriately defined sequence of weighting matrices.

Let $\Theta = \mathbb{R}^k$ be the parameter space, and let the Borel subset $E \subset \mathbb{R}^{1+k+L}$ be the data space. We can then define the function $g : E \times \Theta \to \mathbb{R}^L$ as

$$g((y, x, z), \beta) = z(y - x'\beta)$$

for any $(y, x, z) \in E$, so that the moment condition can be written as

$$\mathbb{E}[g((Y_t, X_t, Z_t), \beta)] = \mathbf{0}.$$

The correpsonding sample moment condition is

$$\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) = \mathbf{0}.$$

This represents a system of L equations for k unknowns; since there are more equations than unknowns, there are potentially multiple solutions to the above system of equations, and we must instead estimate β via the minimization of a quadratic form involving the sample moments. To this end, we first consider an aribtrary sequence of weighting matrices.

Let $\{W_T\}_{T \in N_+}$ be a sequence of $L \times L$ positive definite valued random matrices such that

$$W_T \xrightarrow{p} W,$$

where W is a positive definite $L \times L$ matrix.

We can now consider the sequence of real stochastic processes $\{\{\hat{Q}_T(b)\}_{b \in \Theta}\}_{T \in N_+}$ such that

$$\hat{Q}_T(b) = \left[\frac{1}{T}\sum_{t=1}^{T} g((Y_t, X_t, Z_t), b)\right]' W_T \left[\frac{1}{T}\sum_{t=1}^{T} g((Y_t, X_t, Z_t), \beta)\right].$$

for any $T \in N_+$ and $b \in \Theta$. The sequence of GMM estimators $\{\hat{\beta}_T\}_{T \in N_+}$ of β is found by solving the minimization problem

$$\min_{b \in \Theta} \quad \hat{Q}_T(b) = \left[\frac{1}{T} \sum_{t=1}^{T} Z_t (Y_t - X_t' b) \right]' W_T \left[\frac{1}{T} \sum_{t=1}^{T} Z_t (Y_t - X_t' b) \right]$$

for each $T \in N_+$. Note how, unlike what we assumed to establish the asymptotic properties of GMM estimators under general (possibly non-linear) moment conditions, the objective functions here are formulated not as continuous functions of the data and parameters, but instead as random functions defined on Θ. This reflects the fact that the sequence $\{W_T\}_{T \in N_+}$ of weighting matrices may not directly depend on the data.

6.2.3 The Existence and Asymptotic Properties of GMM Estimators

As we saw above, the GMM estimator of β given a sample of size T under the weighting matrix W_T is found as the solution to the following problem:

$$\min_{b \in \Theta} \quad \hat{Q}_T(b) = \left[\frac{1}{T} \sum_{t=1}^{T} Z_t (Y_t - X_t' b) \right]' W_T \left[\frac{1}{T} \sum_{t=1}^{T} Z_t (Y_t - X_t' b) \right].$$

Note that

$$\frac{\partial \hat{Q}_T(b)}{\partial b} = -2 \cdot \left[\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right]' W_T \left[\frac{1}{T} \sum_{t=1}^{T} Z_t (Y_t - X_t' b) \right]$$

$$\frac{\partial^2 \hat{Q}_T(b)}{\partial b \partial b'} = 2 \cdot \left[\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right]' W_T \left[\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right].$$

Since $\frac{1}{T} \sum_{t=1}^{T} Z_t X_t'$ has rank $k \leq L$ and W_T is positive definite valued,

$$\left[\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right]' W_T \left[\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right]$$

is also positive semidefinite with full rank, that is, positive definite. Therefore, the objective function is strictly convex, and since the parameter space Θ is also open, any estimator of β that solves the first order condition

$$\frac{\partial \hat{Q}_T(b)}{\partial b} = \mathbf{0}$$

will be the unique minimizer of the objective function and therefore the desired GMM estimator of β. Solving the above equation yields our estimator

$$\hat{\beta}_T = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right).$$

Using the fact that $Y_t = X_t'\beta + \epsilon_t$ for any $t \in N_+$, we can now expand $\hat{\beta}_T$ in a manner similar to what we did for the OLS estimator:

$$\begin{aligned}
\hat{\beta}_T &= \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t (X_t'\beta + \epsilon_t) \right) \\
&= \beta + \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T (\mathbb{X}_1, \cdots, \mathbb{X}_T) \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t \epsilon_t \right).
\end{aligned}$$

The consistency and asymptotic normality of the GMM estimator for β now follows easily from this expansion, as shown below:

Theorem 6.1 (Consistency and Asymptotic Normality of the GMM Estimator) Suppose assumptions A1 to A5 hold, and let $\{\hat{\beta}_T\}_{T \in N_+}$ be the sequence of GMM estimators obtained under the sequence $\{W_T\}_{T \in N_+}$ of weighting matrices. Then,

$$\hat{\beta}_T \xrightarrow{p} \beta,$$

and

$$\sqrt{T} \left[\hat{\beta}_T - \beta \right] \xrightarrow{d} \mathcal{N}[\mathbf{0}, V]$$

for the $k \times k$ positive definite matrix V defined as

$$V = (Q_{zx}' W Q_{zx})^{-1} Q_{zx}' W \Sigma W Q_{zx} (Q_{zx}' W Q_{zx})^{-1}.$$

Proof) From the expansion of $\hat{\beta}_T$ established above, we can see that

$$\hat{\beta}_T - \beta = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t \epsilon_t \right).$$

By assumption,

$$W_T \xrightarrow{p} W, \quad \frac{1}{T} \sum_{t=1}^{t} Z_t X_t' \xrightarrow{p} Q_{zx}$$

for some positive definite $L \times L$ matrix W and $L \times k$ matrix Q_{zx} of rank k. Therefore, $Q'_{zx} W Q_{zx}$ has full rank k, so that it is invertible. By the continuity of the matrix inversion operation under the trace norm, we now have

$$\left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \xrightarrow{p} (Q'_{zx} W Q_{zx})^{-1}$$

from the continuous mapping theorem.

Similarly, since $\{Z_t \epsilon_t\}_{t \in N_+}$ is an independent (and thus pairwise uncorrelated) stochastic process with common mean 0 and bounded second moments, by Chebyshev's WLLN

$$\frac{1}{T} \sum_{t=1}^{T} Z_t \epsilon_t \xrightarrow{p} \mathbf{0}.$$

From the continuous mapping theorem again, we have

$$\hat{\beta}_T - \beta \xrightarrow{p} (Q'_{zx} W Q_{zx})^{-1} Q'_{zx} W \mathbf{0} = \mathbf{0},$$

which establishes the desired consistency result.

For asymptotic normality, we rely on the expression

$$\sqrt{T} \left[\hat{\beta}_T - \beta \right] = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' W_T \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t \epsilon_t \right).$$

Since $\{Z_t \epsilon_t\}_{t \in N_+}$ is a sequence of independent random variables with mean 0, it is automatically a martingale difference sequence with respect to the filtration it generates. In addition, $\{|Z_t \epsilon_t|\}_{t \in N_+}$ is $L^{2+\delta}$-bounded, and

$$\frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 Z_t Z_t' \xrightarrow{p} \Sigma$$

for some $L \times L$ positive definite matrix Σ by assumption A5. Therefore, by the martingale difference sequence CLT,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t \epsilon_t \xrightarrow{d} \mathcal{N}[\mathbf{0},\, \Sigma].$$

Therefore, by Slutsky's theorem,

$$\begin{aligned} \sqrt{T} \left[\hat{\beta}_T - \beta \right] &\xrightarrow{d} (Q'_{zx} W Q_{zx})^{-1} Q'_{zx} W \cdot \mathcal{N}[\mathbf{0},\, \Sigma] \\ &= \mathcal{N} \left[\mathbf{0},\, \underbrace{(Q'_{zx} W Q_{zx})^{-1} Q'_{zx} W \Sigma W Q_{zx} (Q'_{zx} W Q_{zx})^{-1}}_{V} \right]. \end{aligned}$$

Q.E.D.

Note how we were able to establish the existence and consistency of the GMM estimator of β for large enough T without having to the turn to the basic consistency theorem. This was made possible because of the linear structure of the moment equation g. In particular, note how we can establish consistency without even having to directly check that the universal identification condition holds, and also under the assumption of a non-compact parameter space. Which assumption is the key for identification?

To investigate this, we once again investigate the population moment condition, which is taken as the probability limit of the sample analogue of our primitive moment conditions. Under the exogeneity assumption for the instruments, we have L (primitive) moment conditions

$$\mathbb{E}\left[Z_t\epsilon_t\right] = \mathbb{E}\left[Z_t(Y_t - X_t'\beta)\right] = O_{L\times 1}.$$

The sample analogue of these conditions is

$$\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) = O_{L\times 1},$$

and letting β_0 be the true value of β, the left hand side has probability limit

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) &= \frac{1}{T}\sum_{t=1}^{T} Z_t(X_t'(\beta_0 - \beta) + \epsilon_t) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)(\beta_0 - \beta) + \frac{1}{T}\sum_{t=1}^{T} Z_t\epsilon_t \xrightarrow{p} Q_{zx}(\beta_0 - \beta).\end{aligned}$$

Here, because Q_{zx} has full rank k, the population moment conditions are exactly satisfied if and only if $\beta = \beta_0$, meaning that the parameters in the model are identified. Thus, the relevance condition guarantees that the parameters are appropriately identified.

6.2.4 The Optimal Weighting Matrix

Now that we have established that, given our assumptions, any GMM estimator of β obtained under a sequence of weighting matrices $\{W_T\}_{T\in N_+}$ that converges in probability to some positive definite matrix W is consistent and asymptotically normal. We want to use the most asymptotically efficient of these estimators; fortunately, it turns out that the most efficient GMM estimator can be obtained under a simple asymptotic weighting matrix W, which also produces a simpler expression for the asymptotic variance. We first present two lemmas which are of independent interest.

Lemma 6.2 (Simultaneous Diagonalization of Positive Definite Matrices)
Let $A, B \in \mathbb{R}^{k\times k}$ be symmetric matrices such that A is positive definite and B is positive semidefinite. Then, there exists a non-singular $P \in \mathbb{R}^{k\times k}$ and a diagonal matrix $D \in \mathbb{R}^{k\times k}$ with non-negative diagonal entries such that

$$A = PP' \quad \text{and} \quad B = PDP'.$$

In addition, $P'P$ has the same set of eigenvalues as A, and the diagonal elements of D are solutions to the eigenvalue equation

$$\det(B - \lambda \cdot A) = 0.$$

Proof) To see this, let $A = \Lambda_1 D_1 \Lambda_1'$ be the orthogonal diagonalization of A (which exists by the principal axis theorem). Defining $D_1^{\frac{1}{2}}$ as the diagonal matrix whose diagonal elements are the square roots of the eigenvalues of A, $D_1^{\frac{1}{2}} D_1^{\frac{1}{2}\prime} = D_1$ and $D_1^{\frac{1}{2}}$ is nonsingular because A is positive definite. Defining $C = \Lambda_1 D_1^{\frac{1}{2}}$, it follows that $A = CC'$.

Since $C^{-1}B(C')^{-1}$ is a positive semidefinite matrix, it has eigendecomposition

$$C^{-1}B(C')^{-1} = \Lambda_2 D \Lambda_2',$$

where Λ_2 is an orthogonal matrix and D collects the eigenvalues of $C^{-1}B(C')^{-1}$, which are all non-negative valued by the positive semidefiniteness of $C^{-1}B(C')^{-1}$. Defining $P = C\Lambda_2$, we can immediately see that

$$B = (C\Lambda_2) D_2 (C\Lambda_2)' = PD_2P',$$

and that

$$PP' = C\Lambda_2\Lambda_2'C' = CC' = A.$$

To establish the claims about the eigenvalues of $P'P$, note that

$$P'P = \Lambda_2' C' C \Lambda_2 = \Lambda_2' D_1^{\frac{1}{2}} \underbrace{\Lambda_1' \Lambda_1}_{I_k} D_1^{\frac{1}{2}} \Lambda_2 = \Lambda_2' D_1 \Lambda_2.$$

As such, the characteristic polynomial of $P'P$ is given by

$$ch_{P'P}(\lambda) = \det(P'P - \lambda \cdot I_k) = \det(\Lambda_2' D_1 \Lambda_2 - \lambda \cdot I_k) = \det(D_1 - \lambda \cdot I_k),$$

where we used the fact that Λ_2 is orthogonal to justify the last equality. Thus, $P'P$ shares the same eigenvalues as D_1 and therefore A.

Finally, the final claim can be seen by first making the observation that the diagonal elements of D are exactly equal to its eigenvalues. The characteristic polynomial of D is given as

$$ch_D(\lambda) = \det(D - \lambda \cdot I_k) = \det\Big(P^{-1}\Big)^2 \cdot \det(PDP' - \lambda \cdot PP') = \det\Big(P^{-1}\Big)^2 \cdot \det(B - \lambda A),$$

so the eigenvalues of D are exactly the solutions to the equation $\det(B - \lambda \cdot A) = 0$.

Q.E.D.

Lemma 6.3 (Evaluating Positive Definiteness of Matrix Differences)
Let $A, B \in \mathbb{R}^{k \times k}$ be positive definite matrices. Then, the following are equivalent:

i) $A - B$ is positive (semi)definite.

ii) $B^{-1} - A^{-1}$ is positive (semi)definite.

iii) The solutions to the equation $\det(B - \lambda \cdot A) = 0$, which are all real positive valued, have modulus less than (or equal to) 1.

Proof) By the preceding lemma, there exists a non-singular $P \in \mathbb{R}^{k \times k}$ and a diagonal matrix $D \in \mathbb{R}^{k \times k}$ with non-negative diagonal entries such that

$$A = PP' \quad \text{and} \quad B = PDP'.$$

Since B is positive definite, the diagonal elements of D are all positive; otherwise, B has reduced rank, which contradicts is positive definiteness. We saw above that the diagonal elements of D are the solutions to the equation $\det(B - \lambda \cdot A) = 0$; thus, the solutions to the equation $\det(B - \lambda \cdot A) = 0$ are all positive. Let $\lambda_1 \geq \cdots \geq \lambda_k > 0$ be the diagonal elements of D.

The difference $A-B$ can now be written as

$$A-B=P(I_k-D)P'=P\begin{pmatrix}1-\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1-\lambda_k\end{pmatrix}P'.$$

Therefore, for any non-zero $\alpha\in\mathbb{R}^k$,

$$\alpha'(A-B)\alpha=\alpha'P\begin{pmatrix}1-\lambda_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1-\lambda_k\end{pmatrix}P'\alpha\geq\left(1-\max_{1\leq i\leq k}\lambda_i\right)\cdot\left|P'\alpha\right|^2,$$

where $P'\alpha$ is non-zero because P' is non-singular and α is non-zero. It follows that $A-B$ is positive (semi)definite if and only if

$$1>(\geq)\max_{1\leq i\leq k}\lambda_i.$$

Similarly,

$$B^{-1}-A^{-1}=(P')^{-1}\left(D^{-1}-I_k\right)P^{-1}=(P')^{-1}\begin{pmatrix}\frac{1}{\lambda_1}-1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_k}-1\end{pmatrix}P^{-1}$$

and, for any non-zero $\alpha\in\mathbb{R}^k$,

$$\alpha'(B^{-1}-A^{-1})\alpha\geq\left(\min_{1\leq i\leq k}\frac{1}{\lambda_i}-1\right)\cdot\left|P^{-1}\alpha\right|^2.$$

We can thus see that $B^{-1}-A^{-1}$ is positive (semi)definite if and only if

$$\min_{1\leq i\leq k}\frac{1}{\lambda_i}>(\geq)1,$$

which is equivalent to the inequality

$$1>(\geq)\max_{1\leq i\leq k}\lambda_i.$$

This establishes the claim.

Q.E.D.

Theorem 6.4 (Choice of Optimal Weighting Matrix)
Suppose assumptions A1 to A5 hold, and let $\{\hat{\beta}_T\}_{T\in N_+}$ be the sequence of GMM estimators obtained under the sequence $\{W_T\}_{T\in N_+}$ of weighting matrices. Assume in addition that

$$W_T \xrightarrow{p} \Sigma^{-1}.$$

Then, the asymptotic distribution of $\hat{\beta}_T$ is given as

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0},\ (Q'_{zx}\Sigma^{-1}Q_{zx})^{-1}\right]$$

and the asymptotic variance $(Q'_{zx}\Sigma^{-1}Q_{zx})^{-1}$ is the most efficient asymptotic variance that can be achieved by a GMM estimator.

Proof) That $\hat{\beta}_T$ is asymptotically normal with asymptotic variance equal to $(Q'_{zx}\Sigma^{-1}Q_{zx})^{-1}$ follows immediately from the preceding theorem by substituting Σ^{-1} in for W in the expression

$$V = (Q'_{zx}WQ_{zx})^{-1}Q'_{zx}W\Sigma WQ_{zx}(Q'_{zx}WQ_{zx})^{-1}.$$

It remains to establish the asymptotic efficiency of $\hat{\beta}_T$. To this end, we choose any $L\times L$ positive definite matrix W and show that

$$(Q'_{zx}WQ_{zx})^{-1}Q'_{zx}W\Sigma WQ_{zx}(Q'_{zx}WQ_{zx})^{-1} - (Q'_{zx}\Sigma^{-1}Q_{zx})^{-1}$$

is positive semidefinite.

In light of the preceding lemma, we need only show that the matrix

$$\Delta = Q'_{zx}\Sigma^{-1}Q_{zx} - (Q'_{zx}WQ_{zx})(Q'_{zx}W\Sigma WQ_{zx})^{-1}(Q'_{zx}WQ_{zx})$$

is positive semidefinite. Letting $\Sigma = CC'$ be the Cholesky decomposition of the positive definite matrix Σ, we can expand Δ as follows

$$\begin{aligned}\Delta &= Q'_{zx}\left[\Sigma^{-1} - WQ_{zx}(Q'_{zx}W\Sigma WQ_{zx})^{-1}Q'_{zx}W\right]Q_{zx}\\ &= Q'_{zx}(C')^{-1}\left[I_k - C'WQ_{zx}\left[(Q'_{zx}WC)(C'WQ_{zx})\right]^{-1}Q'_{zx}WC\right]C^{-1}Q_{zx}.\end{aligned}$$

Denoting $D = C'WQ_{zx}$, the above equation can be written as

$$\Delta = Q'_{zx}(C')^{-1}\underbrace{\left[I_k - D(D'D)^{-1}D'\right]}_{M_D}C^{-1}Q_{zx},$$

where M_D is a symmetric and idempotent matrix. Since the eigenvalues of M_D are equal to 1 or 0, M_D is positive semidefinite, which indicates that Δ is also positive

semidefinite.

Q.E.D.

In light of the above theorem,

$$\hat{\beta}_T^E = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \Sigma^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \Sigma^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right).$$

is called the efficient GMM estimator of β, and its asymptotic distribution is

$$\sqrt{T} \left[\hat{\beta}_T^E - \beta \right] \xrightarrow{p} \mathcal{N} \left[\mathbf{0},\ (Q_{zx}' \Sigma^{-1} Q_{zx})^{-1} \right].$$

6.2.5 Feasible Efficient GMM

One problem with the above estimator is that we do not know Σ. As such, for testing and inference to be made feasible, we must use the feasible efficient GMM estimator

$$\hat{\beta}_T^{FE} = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right),$$

where $\{\hat{\Sigma}_T\}_{T \in N_+}$ is a sequence of random $L \times L$ matrices that converge in probability to Σ. Since $\{\hat{\Sigma}_T\}_{T \in N_+}$ can be viewed as a sequence of weighting matrices under which the asymptotic variance is efficient, $\{\hat{\beta}_T^{FE}\}_{T \in N_+}$ can itself be viewed as a sequnece of efficient GMM estimators.

To construct a consistent estimator of Σ, we take inspiration from White's heteroskedasticity robust covariance estimator and consider an estimator of the following form:

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{\epsilon}_{i,t}^2 Z_t Z_t'$$

where we define

$$\hat{\epsilon}_{i,t}^2 = (Y_t - X_t' \tilde{\beta}_T)^2$$

for some consistent estimator $\tilde{\beta}_T$ of β. It turns out that this estimator is fortunately consistent for Σ, as we show below:

Theorem 6.5 (Consistency of Weighting Matrix Estimator)
Suppose assumptions A1 to A5 hold, and let $\tilde{\beta}_T$ be a consistent estimator for β. Then, the sequence $\{\hat{\Sigma}_T\}_{T \in N_+}$ of $L \times L$ random matrices defined as

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t' \tilde{\beta}_T)^2 Z_t Z_t'$$

for any $T \in N_+$ converges in probability to Σ.

Proof) The proof is almost identical to the proof of the consistency of White's covariance estimator. Like in the proof of that result, we first use the fact that $Y_t = X_t' \beta + \epsilon_t$ to decompose $\hat{\Sigma}_T$ as follows:

$$\begin{aligned}
\hat{\Sigma}_T &= \frac{1}{T} \sum_{t=1}^{T} \left(X_t'(\beta - \tilde{\beta}_T) + \epsilon_t \right)^2 Z_t Z_t' \\
&= \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t X_t' (\beta - \tilde{\beta}_T) \right] Z_t Z_t' \\
&\quad + 2 \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t \epsilon_t \right] Z_t Z_t' + \frac{1}{T} \sum_{t=1}^{T} \epsilon_t^2 Z_t Z_t'.
\end{aligned}$$

Once again, the proof will be complete if we can show that the first two terms are $o_p(1)$ as $T \to \infty$.

As before, we can expand the trace of the first term out as

$$\begin{aligned}
\left\| \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t X_t' (\beta - \tilde{\beta}_T) \right] Z_t Z_t' \right\|^2 &= \sum_{j=1}^{L} \sum_{m=1}^{L} \left| \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t X_t' (\beta - \tilde{\beta}_T) \right] Z_{jt} Z_{jm} \right|^2 \\
&\leq \left| \tilde{\beta}_T - \beta \right|^4 \cdot \sum_{j=1}^{L} \sum_{m=1}^{L} \left\| \frac{1}{T} \sum_{t=1}^{T} X_t X_t' Z_{jt} Z_{jm} \right\|^2 \\
&= \left| \tilde{\beta}_T - \beta \right|^4 \cdot \sum_{j=1}^{L} \sum_{m=1}^{L} \sum_{l=1}^{k} \sum_{p=1}^{k} \left(\frac{1}{T} \sum_{t=1}^{T} X_{lt} X_{pt} Z_{jt} Z_{jm} \right)^2 .
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E} \left| \frac{1}{T} \sum_{t=1}^{T} X_{lt} X_{pt} Z_{jt} Z_{jm} \right| &\leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} |X_{lt} X_{pt} Z_{jt} Z_{mt}| \\
&\leq \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E} |X_{lt} X_{pt}|^2 \right)^{\frac{1}{2}} \left(\mathbb{E} |Z_{jt} Z_{mt}|^2 \right)^{\frac{1}{2}} \\
&\leq \frac{1}{T} \sum_{t=1}^{T} \left(\mathbb{E} |X_{lt}|^4 \right)^{\frac{1}{4}} \left(\mathbb{E} |X_{pt}|^4 \right)^{\frac{1}{4}} \left(\mathbb{E} |Z_{jt}|^4 \right)^{\frac{1}{4}} \left(\mathbb{E} |Z_{mt}|^4 \right)^{\frac{1}{4}} \\
&\leq \left(\sup_{t \in N_+} \mathbb{E} |X_t|^4 \right)^{\frac{1}{2}} \left(\sup_{t \in N_+} \mathbb{E} |Z_t|^4 \right)^{\frac{1}{2}} < +\infty,
\end{aligned}$$

we can see that $\frac{1}{T} \sum_{t=1}^{T} X_{lt} X_{pt} Z_{jt} Z_{jm} = O_p(1)$ as $T \to \infty$. Since $\tilde{\beta}_T - \beta = o_p(1)$, this implies that

$$\left\| \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t X_t' (\beta - \tilde{\beta}_T) \right] Z_t Z_t' \right\|^2 = o_p(1) \quad \text{as } T \to \infty.$$

Similarly,

$$\begin{aligned}
\left\| \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t \epsilon_t \right] Z_t Z_t' \right\|^2 &= \sum_{j=1}^{L} \sum_{m=1}^{L} \left| \frac{1}{T} \sum_{t=1}^{T} \left[(\beta - \tilde{\beta}_T)' X_t \epsilon_t \right] Z_{jt} Z_{mt} \right|^2 \\
&\leq \left| \tilde{\beta}_T - \beta \right|^2 \cdot \sum_{j=1}^{L} \sum_{m=1}^{L} \left| \frac{1}{T} \sum_{t=1}^{T} X_t \epsilon_t Z_{jt} Z_{mt} \right|^2
\end{aligned}$$

$$= \left|\tilde{\beta}_T - \beta\right|^2 \cdot \sum_{j=1}^{L}\sum_{m=1}^{L}\sum_{p=1}^{k}\left(\frac{1}{T}\sum_{t=1}^{T} X_{pt}\epsilon_t Z_{jt} Z_{mt}\right)^2.$$

Since

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T} X_{pt}\epsilon_t Z_{jt} Z_{mt}\right| &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_{pt}\epsilon_t Z_{jt} Z_{mt}| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}|\epsilon_t X_{pt}|^2\right)^{\frac{1}{2}}\left(\mathbb{E}|Z_{jt}|^4\right)^{\frac{1}{4}}\left(\mathbb{E}|Z_{mt}|^4\right)^{\frac{1}{4}} \\
&\leq \left(\sup_{t\in N_+}\mathbb{E}|Z_t|^4\right)^{\frac{1}{2}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}|\epsilon_t X_t|^2\right)^{\frac{1}{2}}\right] \\
&\leq \left(\sup_{t\in N_+}\mathbb{E}|Z_t|^4\right)^{\frac{1}{2}}\left(\sup_{t\in N_+}\mathbb{E}|\epsilon_t X_t|^2\right)^{\frac{1}{2}} < +\infty,
\end{aligned}$$

we can see that $\frac{1}{T}\sum_{t=1}^{T} X_{pt}\epsilon_t Z_{jt} Z_{mt} = O_p(1)$ as $T \to \infty$, so that, as above,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}\left[(\beta - \tilde{\beta}_T)' X_t \epsilon_t\right] Z_t Z_t'\right\|^2 = o_p(1) \quad \text{as } T \to \infty.$$

Therefore,

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 Z_t Z_t' + o_p(1) \xrightarrow{p} \Sigma.$$

Q.E.D.

The theorem above shows us that, to complete our derivation of a feasible efficient GMM estimator, we need only procure a consistent estimator of β that does not depend on any unknown quantities (this rules out using $\hat{\beta}_T^E$ to construct $\hat{\Sigma}_T$). The next section concerns the most popular such estimator, which is of independent interest on its own as well.

6.2.6 The 2SLS Estimator

This is perhaps the most well-known estimator of β based on instrumental variables, and is defined as

$$\hat{\beta}_T^{2SLS} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right).$$

In other words, it is the GMM estimator of β obtained using a sequence of weighting matrices $\{W_T^{2SLS}\}_{T\in N_+}$ given as

$$W_T^{2SLS} = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}.$$

There are two main ways we can motivate this choice of weighting matrices: one purely statistical and the other more intuitive.

Motivation 1: Efficient GMM under Homoskedasticity

The 2SLS estimator is the efficient GMM estimator of β under homoskedasticity. To see this, suppose that the data is i.i.d. and

$$\mathbb{E}\left[\epsilon_t^2 \mid Z_t\right] = \sigma^2$$

for any $t \in N_+$, so that

$$\mathbb{E}\left[\epsilon_t^2 Z_t Z_t'\right] = \mathbb{E}\left[\mathbb{E}\left[\epsilon_t^2 \mid Z_t\right] Z_t Z_t'\right] = \sigma^2 \underbrace{\mathbb{E}\left[Z_t Z_t'\right]}_{Q_{zz}}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 Z_t Z_t' \xrightarrow{p} \mathbb{E}\left[\epsilon_t^2 Z_t Z_t'\right] = \sigma^2 \cdot Q_{zz}.$$

This means that $\Sigma = \sigma^2 \cdot Q_{zz}$, and as such, the GMM estimator obtained under the sequence of weighting matries $\{W_T^*\}_{T\in N_+}$ given as

$$W_T^* = \left(\sigma^2 \cdot \frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}$$

will be the asymptotically efficient GMM estimator $\hat{\beta}_T^E$, since $W_T^* \xrightarrow{p} \left(\sigma^2 Q_{zz}\right)^{-1} = \Sigma^{-1}$. It turns out that this efficient GMM estimator is exactly the 2SLS estimator $\hat{\beta}_T^{2SLS}$:

$$\hat{\beta}_T^E = \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T^* \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T^* \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right)$$

$$= \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right)$$

$$= \hat{\beta}_T^{2SLS}.$$

Motivation 2: Partialling Out the Endogeneity in X_t

This is a more intuitive derivation of the 2SLS estimator. We already saw that the usual way of regressing Y_t on X_t to estimate β does not work when X_t is correlated with ϵ_t. But what if we extract the part of X_t that is unrelated to ϵ_t and regress Y_t on that part? In this case, we would be able to obtain a consistent estimator of β, since we will be regressing Y_t on a quantity that is orthogonal to ϵ_t.

The question then becomes how to decompose X_t into a part related to ϵ_t and a part that is orthogonal to the errors. This is where the instruments come in; by the way they are chosen, our instruments Z_t are exogenous and relevant, that is, they are orthogonal to the errors ϵ_t but are correlated with the regressors X_t of interest. As such, we can obtain the desired decomposition of X_t by decomposing it into the sum of its projection on the L^2 subspace spanned by the variables in Z_t, and the residual. The projection of X_t on the subspace spanned by the Z_ts will be a linear function of the Z_ts and thus orthogonal to ϵ_t. The residual will be contained in the orthogonal complement of the subspace spanned by Z_t, and therefore contain the part of X_t that is related to ϵ_t.

The estimation scheme then becomes: first obtain the projection of X_t on the subspace spanned by Z_t, and then regress Y_t on this projection, which will be orthogonal to ϵ_t by design. For a simple illustration of how this estimation may proceed, suppose (X_t, Z_t) are jointly normal with mean 0, so that $\mathbb{E}[X_t \mid Z_t] = \Pi' Z_t$, where $\Pi \in \mathbb{R}^{L \times k}$. Then, defining $v_t = X_t - \Pi' Z_t$, we can see that, by design,

$$\mathbb{E}[v_t \mid Z_t] = \mathbb{E}[X_t \mid Z_t] - \Pi' Z_t = \mathbf{0},$$

so that $\mathbb{E}[v_t Z_t'] = O_{k \times L}$; v_t and Z_t are uncorrelated. This means that X_t can be written as

$$X_t = \Pi' Z_t + v_t,$$

where Z_t are the "regressors" orthogonal to the "errors" v_t. This regression equation is called the first stage regression equation.

Now we return to our original regression equation $Y_t = X_t' \beta + \epsilon_t$. Here, the problem is that

X_t is correlated with ϵ_t, so substitute in for X_t the first stage regression equation:

$$\begin{aligned} Y_t &= X_t'\beta + \epsilon_t \\ &= Z_t'\Pi \cdot \beta + \underbrace{v_t'\beta + \epsilon_t}_{u_t}. \end{aligned}$$

This is the second stage regression equation. Here, since v_t' is orthogonal to Z_t by design, and ϵ_t is orthogonal to Z_t by assumption, the above equation can be interpreted as a regression equation with regressors $\Pi' Z_t$ and errors u_t, which are uncorrelated with one another. Classical linear regression theory then indicates that we can consistently estimate β by regressing Y_t on the projection $\Pi' Z_t$.

The usual facts from conditional distributions of jointly normally distributed random variables suggests that

$$\Pi' = \mathbb{E}\left[X_t Z_t'\right] \left(\mathbb{E}\left[Z_t Z_t'\right]\right)^{-1} = Q_{zx}' Q_{zz}^{-1}.$$

The sample analogue of this would be

$$\hat{\Pi}_T' = \left(\frac{1}{T}\sum_{t=1}^{T} X_t Z_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1},$$

which is just the OLS estimator of Π' from the first stage regression. Afterward, our estimator of β would be obtained by regressing Y_t on $\hat{\Pi}_T' Z_t$, that is, as the OLS estimator from the second stage regression:

$$\begin{aligned} \hat{\beta}_T &= \left[\hat{\Pi}_T'\left(\sum_{t=1}^{T} Z_t Z_t'\right)\hat{\Pi}_T\right]^{-1} \hat{\Pi}_T'\left(\sum_{t=1}^{T} Z_t Y_t\right) \\ &= \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right). \end{aligned}$$

This is simply the 2SLS estimator of β, and we have just illustrated where the estimator earned its name; it is the estimator of β obtained by running a regression in two stages. The remarkable thing then is that, even though the above estimator was motivated under restrictive i.i.d. and normality assumptions, it is consistent and asymptotically normal even in more general settings.

The Bottom Line

Since the 2SLS estimator is a special case of the GMM estimator with weighting matrices given by $\{W_T^{2SLS}\}_{T\in N_+}$, it is consistent and asymptotically normal with asymptotic variance

$$(Q_{zx}' Q_{zz}^{-1} Q_{zx})^{-1}.$$

Under i.i.d. data and homoskedasticity, this is the asymptotically efficient variance.

Under our more general assumptions, the asymptotically efficient variance is

$$(Q_{zx}' \Sigma^{-1} Q_{zx})^{-1},$$

and we can obtain a feasible efficient GMM estimator of β as

$$\hat{\beta}_T^{FE} = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right),$$

where one possible choice for the sequence of weighting matrices $\{\hat{\Sigma}_T\}_{T\in N_+}$ that converges in probability to Σ is

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - X_t' \hat{\beta}_T^{2SLS})^2 Z_t Z_t'.$$

6.3 A Test for Instrument Exogeneity: The J Test

In choosing to use efficient GMM estimation over OLS estimation, we are making the huge assumption that the regressors X_t are endogeous. This is huge precsiely because the OLS estimator of β has such good asymptotic properties when X_t and ϵ_t are uncorrelated; it is consistent, asymptotically normal, and is even asymptotically efficient among the class of linear estimators of β. What this means is that we must have a very strong justification for using an estimator of β other than the OLS estimator.

In the current setting, the justification for using efficient GMM over OLS comes from three sources. First, OLS is inconsistent because X_t is endogenous. On the other hand, our instruments Z_t are not only exogenous, or uncorrelated with ϵ_t, but also relevant, or correlated with X_t. Therefore, whenever using efficient GMM, we must run three tests: we must test whether X_t truly is exogeous; we must test whether the instruments we have procured are truly exogenous; and finally, we must test whether our instruments are sufficiently highly correlated with X_t. In the sections below we develop tests that deal with each type of problem in turn.

We first study test statistics designed to test $H_0 : \mathbb{E}[\epsilon_t Z_t] = \mathbf{0}$ against $H_1 : \mathbb{E}[\epsilon_t Z_t] \neq \mathbf{0}$, that is, to test whether our instruments satisfy the exogeneity requirement. To motivate our search for an appropriate test statistic, suppose that Z_t truly is exogenous, so that $\mathbb{E}[\epsilon_t Z_t] = \mathbf{0}$. Then, we would expect the sample analogue

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \epsilon_t = \frac{1}{T}\sum_{t=1}^{T} Z_t (Y_t - X_t'\beta)$$

to also be close to 0. Although we cannot observe β directly, because $\hat{\beta}_T^{FE}$ is sufficiently close to β if assumptions A1 to A5 hold, we can expect

$$\frac{1}{T}\sum_{t=1}^{T} Z_t (Y_t - X_t'\hat{\beta}_T^{FE})$$

to be close to 0. Suppose $k < L$[1], so that the sample moment above does not exactly equal 0. In this case, if $\mathbb{E}[Z_t \epsilon_t] = \mathbf{0}$, then some quadratic form involving $\frac{1}{T}\sum_{t=1}^{T} Z_t (Y_t - X_t'\hat{\beta}_T^{FE})$ will be close

[1]Here is a catch at this point: if $k = L$, that is, if the model is just identified, then $\hat{\beta}_T^{FE}$ is precisely the estimator that solves the system of k equations

$$\frac{1}{T}\sum_{t=1}^{T} Z_t (Y_t - X_t'b) = \mathbf{0}$$

for b.

In this case,

$$\hat{\beta}_T^{FE} = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right)$$

and $\hat{\beta}_T^{FE}$ is called the IV estimator of β. Thus, in this case we cannot construct a test statistic based on $\frac{1}{T}\sum_{t=1}^{T} Z_t (Y_t - X_t'\hat{\beta}_T^{FE})$, since it will equal 0 every time instead of admitting a distribution when scaled appropriately.

to (but not equal to) 0. We can then test for instrument exogeneity by comparing this quadratic form to a critical value chosen from an appropriate one-tailed distribution (chi-squared, F or otherwise) and reject the null when the former is larger than the chosen critical value.

We now formally develop this intuition into a test for instrument exogeneity. Earlier, we noted that the GMM objective function

$$\hat{Q}_T(b) = \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'b)\right]' \hat{\Sigma}_T^{-1} \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'b)\right]$$

converges in probability to

$$Q(b) = (\beta - b)'Q_{zx}'\Sigma^{-1}Q_{zx}(\beta - b)$$

as $T \to \infty$ pointwise for any $b \in \Theta$. Therefore, it seems as if, under the proper scaling, the GMM objective function might converge in distribution to a quadratic form with a chi-squared distribution. This turns out to indeed be the case, as we show below:

Theorem 6.6 (Asymptotic Distribution of General J Stastistic)
Suppose assumptions A1 to A5 hold, and let $\hat{\beta}_T^{FE}$ be the feasible efficient GMM estimator of β obtained under the sequence of weighting matrices $\{\hat{\Sigma}_T^{-1}\}_{T \in N_+}$ computed using the 2SLS estimator. Suppose $L > k$. Then,

$$T \cdot \hat{Q}_T(\hat{\beta}_T^{FE}) = \left[\sum_{t=1}^{T} Z_t(Y_t - X_t'\hat{\beta}_T^{FE})\right]' (T \cdot \hat{\Sigma}_T)^{-1} \left[\sum_{t=1}^{T} Z_t(Y_t - X_t'\hat{\beta}_T^{FE})\right] \xrightarrow{d} \chi^2_{L-k}.$$

Proof) Consider the following expansion:

$$\begin{aligned}
T \cdot \hat{Q}_T(\hat{\beta}_T^{FE}) &= \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\hat{\beta}_T^{FE})\right]' \hat{\Sigma}_T^{-1} \left[\sum_{t=1}^{T} Z_t(Y_t - X_t'\hat{\beta}_T^{FE})\right] \\
&= \left[\frac{1}{T}\sum_{t=1}^{T} \Big(Z_t X_t'(\beta - \hat{\beta}_T^{FE}) + Z_t\epsilon_t\Big)\right]' \hat{\Sigma}_T^{-1} \left[\sum_{t=1}^{T} \Big(Z_t X_t'(\beta - \hat{\beta}_T^{FE}) + Z_t\epsilon_t\Big)\right] \\
&= \Big(\sqrt{T}\Big[\hat{\beta}_T^{FE} - \beta\Big]\Big)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right) \Big(\sqrt{T}\Big[\hat{\beta}_T^{FE} - \beta\Big]\Big) \\
&\quad - 2\Big(\sqrt{T}\Big[\hat{\beta}_T^{FE} - \beta\Big]\Big)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right) \\
&\quad + \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right).
\end{aligned}$$

The asymptotic distribution of this quantity can be obtained by observing that

$$\sqrt{T}\left[\hat{\beta}_T^{FE} - \beta\right] = \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right).$$

Thus,

$$\begin{aligned}
&\left(\sqrt{T}\left[\hat{\beta}_T^{FE} - \beta\right]\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right) \left(\sqrt{T}\left[\hat{\beta}_T^{FE} - \beta\right]\right) \\
&\quad = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right) \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \\
&\qquad\qquad \times \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right),
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\left(\sqrt{T}\left[\hat{\beta}_T^{FE} - \beta\right]\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right) \\
&\quad \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right) \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \\
&\qquad\qquad \times \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
T \cdot \hat{Q}_T(\hat{\beta}_T^{FE}) = &\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right) \\
&- \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right) \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \\
&\times \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \hat{\Sigma}_T^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right).
\end{aligned}$$

Let $\hat{\Sigma}_T^{\frac{1}{2}}$ be the Cholesky factor of $\hat{\Sigma}_T$, so that

$$\hat{\Sigma}_T^{-1} = \hat{\Sigma}_T^{-\frac{1}{2}\prime} \hat{\Sigma}_T^{-\frac{1}{2}}.$$

Defining

$$C_T = \hat{\Sigma}_T^{-\frac{1}{2}} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)$$

and

$$M_T = I_L - C_T(C_T'C_T)^{-1}C_T',$$

we can see that $T \cdot \hat{Q}_T(\hat{\beta}_T^{FE})$ assumes the familiar form

$$T \cdot \hat{Q}_T(\hat{\beta}_T^{FE}) = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right)' \hat{\Sigma}_T^{-\frac{1}{2}\prime} M_T \hat{\Sigma}_T^{-\frac{1}{2}} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right).$$

Letting $\Sigma^{\frac{1}{2}}$ be the Cholesky factor of Σ,

$$C_T \xrightarrow{p} C := \Sigma^{-\frac{1}{2}} Q_{zx}$$

$$M_T \xrightarrow{p} M := I_L - C(C'C)^{-1}C' = I_L - \Sigma^{-\frac{1}{2}} Q_{zx}(Q_{zx}'\Sigma^{-1}Q_{zx})^{-1}Q_{zx}'\Sigma^{-\frac{1}{2}}$$

$$\hat{\Sigma}_T^{-\frac{1}{2}} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t\right) \xrightarrow{d} \Sigma^{-\frac{1}{2}}\mathcal{N}[\mathbf{0},\ \Sigma] = \mathcal{N}[\mathbf{0}, I_L].$$

M is a symmetric and idempotent matrix with rank

$$\mathrm{tr}(M) = L - \mathrm{tr}\left(\Sigma^{-\frac{1}{2}} Q_{zx}(Q_{zx}'\Sigma^{-1}Q_{zx})^{-1}Q_{zx}'\Sigma^{-\frac{1}{2}}\right) = L - k.$$

Therefore, by the continuous mapping theorem and lemma 4.3,

$$T \cdot \hat{Q}_T(\hat{\beta}_T^{FE}) \xrightarrow{d} \mathbb{Z}'M\mathbb{Z} \sim \chi^2_{L-k},$$

where $\mathbb{Z} \sim \mathcal{N}[\mathbf{0}, I_L]$.

Q.E.D.

The statistic defined as

$$J_T = T \cdot \hat{Q}_T(\hat{\beta}_T^{FE})$$

is called Sargan's J statistic, and it can be used to conduct an asymptotic chi-squared test of the null $H_0 : \mathbb{E}[Z_t\epsilon_t] = \mathbf{0}$ against the alternative $H_1 : \mathbb{E}[Z_t\epsilon_t] \neq \mathbf{0}$; the intuition we presented in the beginning of this section suggests that the J test has power against this alternative.

6.4 The Hausman Specification Test

In this section we introduce a class of tests called Hausman specification tests. A Hausman test can be used to test for whether the original regressors are endogenous.

6.4.1 Model Assumptions

The general setting is given as follows. Let there be a statistical model $\mathcal{M}_1$ for a parameter of interest $\beta \in \mathbb{R}^k$. We want to test the null hypotheses

$$H_0 : \mathcal{M}_1 \text{ is the true model} \quad \text{v.s.} \quad H_1 : \mathcal{M}_1 \text{ is not the true model.}$$

Suppose $\hat{\beta}_T$ and $\tilde{\beta}_T$ are two estimators of β that satisfy the following assumptions:

Haus1 : **Consistency Assumptions**
$\hat{\beta}_T$ and $\tilde{\beta}_T$ are consistent for β under H_0, but only $\tilde{\beta}_T$ is consistent for β under H_1.

Haus2 : **Joint Asymptotic Normality under the Null**
$\hat{\beta}_T$ and $\tilde{\beta}_T$ are jointly asymptotically normal under H_0. Specifically, there exists a sequence $\{a_T\}_{T \in N_+}$ of real numbers monotonically diverging to $+\infty$ such that

$$a_T \begin{bmatrix} \hat{\beta}_T - \beta \\ \tilde{\beta}_T - \beta \end{bmatrix} \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \underbrace{\begin{pmatrix} V_1 & C \\ C' & V_2 \end{pmatrix}}_{V} \right],$$

where V is a $2k \times 2k$ positive definite matrix.

Haus3 : **Efficiency Assumptions**
We assume that $\hat{\beta}_T$ is asymptotically efficient under H_0 in the sense that, for any linear combination

$$\hat{q}_T = B_1 \hat{\beta}_T + B_2 \tilde{\beta}_T$$

where $B_1, B_2 \in \mathbb{R}^{k \times k}$ such that $B_1 + B_2 = I_k$, $\hat{\beta}_T$ is more asymptotically efficient than $\hat{q}_T$.

Haus4 : **Consistent Covariance Matrix Estimation**
We assume that there exist estimators $\hat{V}_{1,T}$ and $\hat{V}_{2,T}$ of V_1 and V_2 such that $\hat{V}_{2,T} - \hat{V}_{1,T}$ has the same rank as $V_2 - V_1$ for any $T \in N_+$, and which are consistent under both H_0 and H_1.

6.4.2 Hausman's Result for Asymptotically Efficient Estimators

We now establish prelimiary results concerning the asymptotic behavior of the two estimators $\hat{\beta}_T$ and $\tilde{\beta}_T$ under the null. The following is the key observation, namely that under the efficiency assumption above, the asymptotic variance of $\tilde{\beta}_T - \hat{\beta}_T$ is simply the difference $V_2 - V_1$ of their asymptotic variances. This lets us ignore the asymptotic covariance term and therefore shows us that, asymptotically, $\hat{\beta}_T$ is orthogonal to $\tilde{\beta}_T - \hat{\beta}_T$.

Theorem 6.7 (Hasuman's Result for Asymptotically Efficient Estimators) Suppose assumptions Haus1 to Haus3 are satisfied, so that $\hat{\beta}_T$ is an asymptotically efficient estimator of β in the sense of assumption Haus3. Then,

$$a_T \left[\tilde{\beta}_T - \hat{\beta}_T\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0},\ V_2 - V_1\right]$$

under the null hypothesis.

Proof) For any $B_1, B_2 \in \mathbb{R}^{k\times k}$ such that $B_1 + B_2 = I_k$, define the estimator

$$\hat{q}_T(B_1, B_2) = B_1 \cdot \hat{\beta}_T + B_2 \cdot \tilde{\beta}_T.$$

By assumption Haus1, under H_0

$$\hat{q}_T(B_1, B_2) \xrightarrow{p} B_1 \cdot \beta + B_2 \cdot \beta = \beta,$$

so that $\hat{q}_T(B_1, B_2)$ is a consistent estimator of β under the null hypothesis. Furthermore, by the continuous mapping theorem,

$$\begin{aligned} a_T\left[\hat{q}_T(B_1, B_2) - \beta\right] &= \begin{pmatrix} B_1 & B_2 \end{pmatrix} \cdot \left(a_T \begin{bmatrix} \hat{\beta}_T - \beta \\ \tilde{\beta}_T - \beta \end{bmatrix} \right) \\ &\xrightarrow{d} \begin{pmatrix} B_1 & B_2 \end{pmatrix} \cdot \mathcal{N}\left[\mathbf{0}, V\right] \\ &= \mathcal{N}\left[\mathbf{0},\ B_1 V_1 B_1' + B_2 V_2 B_2' + B_2 C' B_1 + B_1 C B_2'\right]. \end{aligned}$$

By assumption Haus3,

$$B_1 V_1 B_1' + B_2 V_2 B_2' + B_2 C' B_1 + B_1 C B_2' - V_1$$

is positive semidefinite. Since $B_1 + B_2 = I_k$, we have

$$V_1 = (B_1 + B_2) V_1 (B_1' + B_2') = B_1 V_1 B_1' + B_2 V_1 B_2' + B_2 V_1 B_1 + B_1 V_1 B_2'$$

and therefore

$$B_2(V_2 - V_1)B_2' + B_2(C - V_1)'B_1 + B_1(C - V_1)B_2'$$

is positive semidefinite. In particular, let $B_1 = (1-\alpha)I_k$ and $B_2 = \alpha \cdot I_k$ for some $\alpha \in \mathbb{R}$. Then, $B_1 + B_2 = I_k$, so that

$$\begin{aligned}
&\alpha^2(V_2 - V_1) + \alpha(1-\alpha)(C - V_1)' + \alpha(1-\alpha)(C - V_1) \\
&\qquad = \alpha^2 \left(V_2 - V_1 - (C - V_1)' - (C - V_1)\right) + \alpha \left(C' + C - 2V_1\right) \\
&\qquad\qquad = \alpha^2 \left(V_2 + V_1 - C' - C\right) + \alpha \left(C' + C - 2V_1\right)
\end{aligned}$$

is positive semidefinite.

Note that

$$\tilde{\beta}_T - \hat{\beta}_T = \begin{pmatrix} I_k & -I_k \end{pmatrix} \begin{pmatrix} \tilde{\beta}_T \\ \hat{\beta}_T \end{pmatrix},$$

so that, by the continuous mapping theorem,

$$a_T \left[\tilde{\beta}_T - \hat{\beta}_T\right] \xrightarrow{d} \mathcal{N}\left[\mathbf{0},\, V_1 + V_2 - C' - C\right].$$

We want to show that

$$D_1 = V_1 + V_2 - C' - C, \quad D_2 = V_2 - V_1$$

are the same matrix, where D_1 is positive semidefinite because it is the asymptotic variance of $\tilde{\beta}_T - \hat{\beta}_T$.

We just showed above that, for any $\alpha \in \mathbb{R}$,

$$\alpha^2 D_1 + \alpha(D_2 - D_1)$$

is positive semidefinite. For any $\alpha \in (0,1)$, it follows that

$$\alpha D_1 + (D_2 - D_1) = (\alpha - 1)D_1 + D_2$$

is positive semidefinite, and likewise,

$$-\alpha D_1 - (D_2 - D_1) = (1-\alpha)D_1 - D_2$$

is positive semidefinite. Choose any non-zero vector $v \in \mathbb{R}^k$. For any $\alpha \in (0,1)$,

$$v'((\alpha-1)D_1 + D_2)v = (\alpha-1)\cdot v'D_1v + v'D_2v \geq 0$$

and

$$v'((1-\alpha)D_1 - D_2)v = (1-\alpha)\cdot v'D_1v - v'D_2v \geq 0,$$

so we have the chain of inequalities

$$(1-\alpha)\cdot v'D_1v \leq v'D_2v \leq (1-\alpha)v'D_1v,$$

and therefore $(1-\alpha)v'D_1v = v'D_2v$, for any $\alpha \in (0,1)$. Taking $\alpha \to 0$ yields the equality

$$v'D_1v = v'D_2v.$$

This holds for any non-zero $v \in \mathbb{R}^k$, so we have $D_1 = D_2$, as desired.

Q.E.D.

The result above actually implies that

$$2V_1 = C' + C,$$

so that, if $k = 1$, $C - V_1 = 0$. In particular, this implies that

$$a_T \begin{bmatrix} \tilde{\beta}_T - \hat{\beta}_T \\ \hat{\beta}_T - \beta \end{bmatrix} = \begin{pmatrix} -I_k & I_k \\ I_k & O_{k\times k} \end{pmatrix} \left(a_T \begin{bmatrix} \hat{\beta}_T - \beta \\ \tilde{\beta}_T - \beta \end{bmatrix} \right) \xrightarrow{d} \mathcal{N}\left[O_{2\times 1}, \begin{pmatrix} V_2 + V_1 - 2C & C - V_1 \\ C - V_1 & V_1 \end{pmatrix} \right]$$

$$= \mathcal{N}\left[O_{2\times 1}, \begin{pmatrix} V_2 - V_1 & 0 \\ 0 & V_1 \end{pmatrix} \right],$$

so that $\tilde{\beta}_T - \hat{\beta}_T$ and $\hat{\beta}_T$ are asymptotically independent if $k = 1$.

6.4.3 Pseudoinversion of Positive Semidefinite Matrices

Now we can construct the desired test statistic. To this end, we first define the pseudoinverse of a positive definite matrix $A \in \mathbb{R}^{k\times k}$ with non-zero rank r. By the principal axis theorem, such a positive definite matrix A has eigendecomposition $A = PDP'$, where D is a diagonal matrix collecting the eigenvalues of A, and P is an orthogonal matrix whose columns are mutually orthogonal eigenvectors correspnding to the eigenvalues in D. Since A has rank r, D assumes

the form

$$D = \begin{pmatrix} \tilde{D} & O_{r\times(k-r)} \\ O_{(k-r)\times r} & O_{(k-r)\times(k-r)} \end{pmatrix},$$

where $\tilde{D}$ is an $r \times r$ matrix collecting the positive eigenvalues of A. Likewise, let $\tilde{P} \in \mathbb{R}^{k\times r}$ collect the first r columns of P, which are orthonormal eigenvectors corresponding to the non-zero eigenvalues in $\tilde{D}$, and as such satisfies $\tilde{P}'\tilde{P} = I_r$. Then, we can see that

$$A = \tilde{P}\tilde{D}\tilde{P}';$$

we refer to this as the reduced eigendecomposition of A.

We define the pseudoinverse of A as

$$A^\dagger = \tilde{P}\tilde{D}^{-1}\tilde{P}',$$

where $\tilde{D}^{-1}$ exists because the diagonal entries of $\tilde{D}$ are positive. $A^\dagger$ is a $k \times k$ matrix with rank r; when $k = r$, so that A is positive definite, we can see that $A^\dagger = A^{-1}$, so that the pseudoinversion operation can be seen as a sort of generalized inverse matrix for positive semidefinite matrices.

Despite the non-uniqueness of the eigendecomposition of A, its pseudoinverse is unique. To see this, let $\tilde{Q} \in \mathbb{R}^{k\times r}$ be another matrix whose columns are r orthonormal eigenvectors of A corresponding to the eigenvalues in $\tilde{D}$. Then,

$$A = \tilde{Q}\tilde{D}\tilde{Q}',$$

and we want to show that $A^\dagger = \tilde{Q}\tilde{D}^{-1}\tilde{Q} := B'$. Since

$$A\tilde{Q} = \tilde{Q}\tilde{D}, \quad A\tilde{P} = \tilde{P}\tilde{D}$$

by the definition of eigenvectors, we have

$$A\underbrace{(\tilde{Q}\tilde{D}^{-1}\tilde{Q}')}_{B} = A\underbrace{(\tilde{P}\tilde{D}^{-1}\tilde{P}')}_{A^\dagger} = I_k$$

and therefore

$$B = B\underbrace{(AB)}_{I_k} = B(AB)\underbrace{(AA^\dagger)}_{I_k} = \underbrace{(BA)}_{I_k}\underbrace{(BA)}_{I_k}A^\dagger = A^\dagger.$$

This demonstrates that the pseudoinverse is unique; more formally, the pseudoinverse is invariant to the choice of eigenbasis.

The next result motivates our definition of the pseudoinverse of reduced rank positive semidefinite matrices.

Lemma 6.8 (Quadratic Forms with Pseudoinverses)

Let $A \in \mathbb{R}^{k \times k}$ be a positive semidefinite matrix with rank $0 < r \leq k$, and suppose $\mathbb{Z} \sim \mathcal{N}[\mathbf{0}, A]$. Then,

$$\mathbb{Z}' A^{\dagger} \mathbb{Z} \sim \chi^2_r.$$

Proof) Let $A = \tilde{P}\tilde{D}\tilde{P}'$ be the reduced eigendecomposition of A, as defined above. Recall that $A^{\dagger}$ is defined as $A^{\dagger} = \tilde{P}\tilde{D}^{-1}\tilde{P}'$.

Suppose $\mathbb{Z} \sim \mathcal{N}[O_{k \times 1}, A]$. Then,

$$\tilde{P}'\mathbb{Z} \sim \mathcal{N}\left[O_{r \times 1},\ \tilde{P}' A \tilde{P}\right].$$

Here,

$$\tilde{P}' A \tilde{P} = (\tilde{P}'\tilde{P})\tilde{D}(\tilde{P}'\tilde{P}) = \tilde{D},$$

so that

$$(\tilde{P}'\mathbb{Z})'\tilde{D}^{-1}(\tilde{P}'\mathbb{Z}) \sim \chi^2_r.$$

Rewriting the left hand side, we can see that this result is equivalent to saying that

$$\mathbb{Z}' A^{\dagger} \mathbb{Z} \sim \chi^2_r.$$

Q.E.D.

Therefore, defining the pseudoinverse of a positive semidefinite matrix in the above manner allows us to continue to form chi squared random variables using normal random vectors with reduced rank covariance matrices.

The pseudoinverse operation is also continuous on sets of positive semidefinite matrices with constant rank, as we show below:

Lemma 6.9 (Continuity of Pseudoinversion for Matrices of Constant Rank)
Let $S_r^{k \times k}$ be the set of all positive semidefinite matrices with rank $0 < r \leq k$, metrized under the trace norm $\|\cdot\|$. Define the operation $G_r : S_r^{k \times k} \to S_r^{k \times k}$ as

$$G_r(A) = A^{\dagger} \quad \text{for any } A \in S_r^{k \times k}.$$

Then, G_r is a continuous function on $S_r^{k \times k}$ with respect to the trace norm.

Proof) Choose some sequence $\{A_n\}_{n \in N_+} \subset S_r^{k \times k}$ with limit $A \in S_r^{k \times k}$. For any $n \in N_+$, let $A_n =$

$P_n D_n P_n'$ be the eigendecomposition of A_n, $\tilde{D}_n$ the $k \times k$ upper block of D_n collecting the first r eigenvalues of A_n, and $\tilde{P}_n$ the $k \times r$ matrix that collects the first r columns of P_n. By definition,

$$G_r(A_n) = A_n^\dagger = \tilde{P}_n \tilde{D}_n^{-1} \tilde{P}_n'.$$

Letting $\tilde{D}$ collect the r positive eigenvalues of A, by the continuity of ordered eigenvalues of symmetric matrices, $\tilde{D}_n \to \tilde{D}$ and $\tilde{D}_n^{-1} \to \tilde{D}^{-1}$ as $n \to \infty$ in the trace norm.

Furthermore, since $\left\|\tilde{P}_n\right\| = \mathrm{tr}(\tilde{P}_n' \tilde{P}_n)^{\frac{1}{2}} = \sqrt{r}$, the elements of $\tilde{P}_n$ take values in a compact set. Thus, the Bolzano-Weierstrass theorem tells us that any subsequence of $\{\tilde{P}_n\}_{n \in N_+}$ has a further convergent subsequence with limit $\tilde{P} \in \mathbb{R}^{k \times r}$. For the sake of notational simplicity, assume without loss of generality that $\{\tilde{P}_n\}_{n \in N_+}$ itself converges to $\tilde{P}$. Then, since $\tilde{P}_n' \tilde{P}_n = I_r$ for any $n \in N_+$, taking limits on both sides implies that

$$I_k = \lim_{n \to \infty} \tilde{P}_n' \tilde{P}_n = \tilde{P}' \tilde{P}.$$

This shows us that the columns of $\tilde{P}$ collect r orthonormal vectors. In addition, since

$$A_n \tilde{P}_n = \tilde{P}_n \tilde{D}_n$$

for any $n \in N_+$ and $A_n \to A$ and $\tilde{D}_n \to \tilde{D}$, taking $n \to \infty$ on both sides yields

$$A\tilde{P} = \tilde{P}\tilde{D},$$

This shows us that the r columns of $\tilde{P}$ are orthonormal eigenvectors corresponding to the first r eigenvalues of A; therefore,

$$\tilde{P}\tilde{D}^{-1}\tilde{P}' = A^\dagger = G_r(A).$$

This indicates that

$$G_r(A_n) = \tilde{P}_n \tilde{D}_n^{-1} \tilde{P}_n' \to \tilde{P}\tilde{D}^{-1}\tilde{P}' = G_r(A)$$

as $n \to \infty$. We have just shown that any subsequence of $\{G_r(A_n)\}_{n \in N_+}$ has a further subsequence that converges to $G_r(A)$; by the selection principle for subsequences, this implies that $\{G_r(A_n)\}_{n \in N_+}$ itself converges to $G_r(A)$. Therefore, G_r is a continuous function on $S_r^{k \times k}$.

Q.E.D.

The above lemma has an important corollary. Let $\{W_T\}_{T \in N_+}$ be a sequence of $k \times k$ positive semidefinite valued random matrices with rank $0 < r \leq k$ converging in probability to a $k \times$

k positive semidefinite matrix W, also with rank r. Then, from what we showed above, the continuous mapping theorem tells us that

$$W_T^\dagger = G_r(W_T) \xrightarrow{p} G_r(W) = W^\dagger$$

as $T \to \infty$. This means that we can use the usual method to construct consistent estimators of the pseudoinverse of a positive semidefinite matrix W, with the added caveat that the sequence of random matrices converging to W in probability must all have the same rank as W.

6.4.4 Construction of Hausman Test Statistic

We can now define and derive the asymptotic distribution of the Hausman test statistic. The test statistic is defined as

$$H_T = a_T^2 \left(\tilde{\beta}_T - \hat{\beta}_T\right)' \left[\hat{V}_{2,T} - \hat{V}_{1,T}\right]^\dagger \left(\tilde{\beta}_T - \hat{\beta}_T\right),$$

which is the usual quadratic form found in many tests, for instance, in the Wald and J tests. Below, we show that this test statistic has the desired chi squared asymptotic distribution.

Theorem 6.10 (Asymptotic Distribution of Hausman Test Statistic)
Suppose assumptions Haus1 to Haus4 are satisfied, and let $V_2 - V_1$ have rank $0 < r \leq k$. Then, under the null hypothesis,

$$H_T = a_T^2 \left(\tilde{\beta}_T - \hat{\beta}_T\right)' \left[\hat{V}_{2,T} - \hat{V}_{1,T}\right]^\dagger \left(\tilde{\beta}_T - \hat{\beta}_T\right) \xrightarrow{d} \chi_r^2.$$

Proof) First, by assumption Haus4 and lemma 5.9 above, the continuity of the pseudoinverse for matrices of rank r and the consistency of $\hat{V}_{2,T} - \hat{V}_{1,T}$ for $V_2 - V_1$ imply that

$$\left[\hat{V}_{2,T} - \hat{V}_{1,T}\right]^\dagger \xrightarrow{p} (V_2 - V_1)^\dagger.$$

Furthermore, theorem 5.8 tells us that

$$a_T \left[\tilde{\beta}_T - \hat{\beta}_T\right] \xrightarrow{d} \mathbb{Z} \sim \mathcal{N}\left[\mathbf{0},\, V_2 - V_1\right].$$

Therefore, by Slutsky's theorem,

$$H_T = a_T^2 \left(\tilde{\beta}_T - \hat{\beta}_T\right)' \left[\hat{V}_{2,T} - \hat{V}_{1,T}\right]^\dagger \left(\tilde{\beta}_T - \hat{\beta}_T\right) \xrightarrow{d} \mathbb{Z}'(V_2 - V_1)^\dagger \mathbb{Z}.$$

Finally, lemma 5.8 tells us that, because $\mathbb{Z} \sim \mathcal{N}[\mathbf{0},\ V_2 - V_1]$,

$$\mathbb{Z}'(V_2 - V_1)^{\dagger}\mathbb{Z} \sim \chi^2_r.$$

This completes the proof.

Q.E.D.

We can further show that the Hausman test is consistent under certain alternatives. Specifically, consider the alternative under which $\hat{\beta}_T \xrightarrow{p} \gamma$, where $\gamma \neq \beta$ and $(\beta - \gamma)'(V_2 - V_1)^{\dagger}(\beta - \gamma) \neq 0$; since $(V_2 - V_1)^{\dagger}$ is a non-zero matrix, the set of all γ that makes this matrix product equal to 0 is a measure 0 set under the Lebesgue measure.

In this case, because

$$\tilde{\beta}_T - \hat{\beta}_T \xrightarrow{p} \beta - \gamma \neq \mathbf{0},$$

we have

$$\left| a_T \left[\tilde{\beta}_T - \hat{\beta}_T \right] \right| \xrightarrow{p} +\infty.$$

Since

$$\hat{V}_{2,T} - \hat{V}_{1,T} \xrightarrow{p} V_2 - V_1$$

even under H_1, it follows that

$$\left[\hat{V}_{2,T} - \hat{V}_{1,T} \right]^{\dagger} \xrightarrow{p} (V_2 - V_1)^{\dagger}$$

and therefore

$$\left(\tilde{\beta}_T - \hat{\beta}_T \right)' \left[\hat{V}_{2,T} - \hat{V}_{1,T} \right]^{\dagger} \left(\tilde{\beta}_T - \hat{\beta}_T \right) \xrightarrow{p} (\beta - \gamma)'(V_2 - V_1)^{\dagger}(\beta - \gamma) \neq 0.$$

It follows that

$$H_T = a_T^2 \cdot \left(\tilde{\beta}_T - \hat{\beta}_T \right)' \left[\hat{V}_{2,T} - \hat{V}_{1,T} \right]^{\dagger} \left(\tilde{\beta}_T - \hat{\beta}_T \right) \xrightarrow{p} +\infty$$

as $T \to \infty$, so that we reject with probability 1 under the given alternative. This shows us that the Hausman test is consistent under certain alternatives.

6.4.5 The Hausman Test for Testing Endogeneity

Now we want to test $H_0 : X_t$ is exogenous against $H_1 : X_t$ is not exogenous. This can be achieved using the Hausman test when we observe that the OLS estimator is consistent under H_0 but not under H_1, while the feasible efficient GMM estimator $\hat{\beta}_T^{FE}$ of β is consistent regardless of whether X_t is exogenous.

For simplicity, we focus on the homoskedastic case, in which we impose the additional assumptions to the base assumptions A1 to A5:

A6 : **Regularity Conditions for** X_t

Assume that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q_{xx},$$

where Q_{xx} is a positive definite $k \times k$ matrix. By implication, $\frac{1}{T}\sum_{t=1}^{T} X_t X_t'$ is nonsingular with probability approaching 1 as $T \to \infty$.

A7 : **Homoskedasticity**

We assume that $\mathbb{E}\left[\epsilon_t^2 \mid X_t, Z_t\right] = \sigma^2$ for any $t \in N_+$. By implication, since $\mathbb{E}\left[(\epsilon_t^2 - \sigma^2) X_t X_t'\right] = O_{k\times k}$ and $\{(\epsilon_t^2 - \sigma^2) X_t X_t'\}_{t\in N_+}$ has bounded second moments, Chebyshev's WLLN tells us that

$$\frac{1}{T}\sum_{t=1}^{T} (\epsilon_t^2 - \sigma^2) X_t X_t' \xrightarrow{p} O_{k\times k}.$$

Therefore,

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 X_t X_t' \xrightarrow{p} \sigma^2 Q_{xx}.$$

Likewise,

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 \begin{pmatrix} X_t X_t' & X_t Z_t' \\ Z_t X_t' & Z_t Z_t' \end{pmatrix} \xrightarrow{p} \sigma^2 \begin{pmatrix} Q_{xx} & Q_{zx}' \\ Q_{zx} & Q_{zz} \end{pmatrix}.$$

Under the homoskedasticity assumption, recall that the two stage least squares estimator $\hat{\beta}_T^{2SLS}$ defined as

$$\hat{\beta}_T^{2SLS} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right)$$

is asymptotically efficient. Meanwhile, the error variance σ^2 can be consistently estimated via

$$\hat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}(Y_t - X_t'\hat{\beta}_T^{2SLS})^2,$$

under both the null and the alternative.

Let $\hat{\beta}_T^{OLS}$ be the OLS estimator of β defined as

$$\hat{\beta}_T^{OLS} = \left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}X_tY_t\right) = \beta + \left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}X_t\epsilon_t\right).$$

We now verify that $\hat{\beta}_T^{OLS}$ and $\hat{\beta}_T^{2SLS}$ satisfy the assumptions of the general Hausman test:

1. **Consistency Assumptions**
   Under H_0, because $\mathbb{E}[X_t\epsilon_t] = \mathbf{0}$ for any $t \in N_+$ and $\{X_t\epsilon_t\}_{t\in N_+}$ is an independent process with bounded second moments, Chebyshev's WLLN tells us that

   $$\frac{1}{T}\sum_{t=1}^{T}X_t\epsilon_t \xrightarrow{p} \mathbf{0}$$

   and therefore

   $$\hat{\beta}_T^{OLS} - \beta \xrightarrow{p} Q_{xx}^{-1}\mathbf{0} = \mathbf{0}.$$

   This shows us that $\hat{\beta}_T^{OLS}$ is consistent under H_0.

   Meanwhile, we already showed above that $\hat{\beta}_T^{2SLS}$ is consistent for β regardless of whether X_t is exogenous.

2. **Joint Asymptotic Normality under the Null**
   Under H_0,

   $$\begin{bmatrix}\hat{\beta}_T^{OLS} - \beta \\ \hat{\beta}_T^{2SLS} - \beta\end{bmatrix} = \begin{pmatrix}\left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)^{-1} & O_{k\times k} \\ O_{k\times k} & \left[\left(\frac{1}{T}\sum_{t=1}^{T}Z_tX_t'\right)'\left(\frac{1}{T}\sum_{t=1}^{T}Z_tZ_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}Z_tX_t'\right)\right]^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}Z_tX_t'\right)'\left(\frac{1}{T}\sum_{t=1}^{T}Z_tZ_t'\right)^{-1}\end{pmatrix}$$

   $$\times\left[\frac{1}{T}\sum_{t=1}^{T}\begin{pmatrix}X_t \\ Z_t\end{pmatrix}\epsilon_t\right].$$

   Since

   $$\mathbb{E}\left[\begin{pmatrix}X_t \\ Z_t\end{pmatrix}\epsilon_t\right] = \mathbf{0},$$

   the stochastic process $\{\epsilon_t(X_t', Z_t')'\}_{t\in N_+}$ is independent, has bounded fourth moments, and

satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2\begin{pmatrix}X_tX_t' & X_tZ_t'\\ Z_tX_t' & Z_tZ_t'\end{pmatrix}\xrightarrow{p}\sigma^2\cdot\begin{pmatrix}Q_{xx} & Q_{zx}'\\ Q_{zx} & Q_{zz}\end{pmatrix}$$

by assumption. The standard martingale difference sequence CLT tells us that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\begin{pmatrix}X_t\\ Z_t\end{pmatrix}\epsilon_t\xrightarrow{d}\mathcal{N}\left[O_{(k+L)\times 1},\ \sigma^2\begin{pmatrix}Q_{xx} & Q_{zx}'\\ Q_{zx} & Q_{zz}\end{pmatrix}\right].$$

By Slutsky's theorem,

$$\begin{aligned}\sqrt{T}\begin{bmatrix}\hat{\beta}_T^{OLS}-\beta\\ \hat{\beta}_T^{2SLS}-\beta\end{bmatrix}&\xrightarrow{d}\begin{pmatrix}Q_{xx}^{-1} & O_{k\times k}\\ O_{k\times k} & (Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}Q_{zx}'Q_{zz}^{-1}\end{pmatrix}\times\mathcal{N}\left[O_{2k\times 1},\ \sigma^2\begin{pmatrix}Q_{xx} & Q_{zx}'\\ Q_{zx} & Q_{zz}\end{pmatrix}\right]\\ &=\mathcal{N}\left[O_{2k\times 1},\ \sigma^2\begin{pmatrix}Q_{xx}^{-1} & Q_{xx}^{-1}\\ Q_{xx}^{-1} & (Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}\end{pmatrix}\right].\end{aligned}$$

3. **Efficiency Assumption**

For any $B_1, B_2 \in \mathbb{R}^{k\times k}$ such that $B_1 + B_2 = I_k$, define

$$\hat{q}_T = B_1\cdot\hat{\beta}_T^{OLS} + B_2\cdot\hat{\beta}_T^{2SLS}.$$

Then, under H_0,

$$\begin{aligned}\sqrt{T}\left[\hat{q}_T-\beta\right]&=\begin{pmatrix}B_1 & B_2\end{pmatrix}\left(\sqrt{T}\begin{bmatrix}\hat{\beta}_T^{OLS}-\beta\\ \hat{\beta}_T^{2SLS}-\beta\end{bmatrix}\right)\\ &\xrightarrow{d}\mathcal{N}\left[O_{k\times 1},\ \sigma^2\cdot V_q\right],\end{aligned}$$

where

$$V_q = B_1Q_{xx}^{-1}B_1' + B_2(Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1}B_2' + B_1Q_{xx}^{-1}B_2' + B_2Q_{xx}^{-1}B_1'.$$

Note that

$$\begin{aligned}V_q - Q_{xx}^{-1} &= V_q - (B_1+B_2)Q_{xx}^{-1}(B_1+B_2)\\ &= B_2\left[(Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1} - Q_{xx}^{-1}\right]B_2'.\end{aligned}$$

Here, we can show that $(Q_{zx}'Q_{zz}^{-1}Q_{zx})^{-1} - Q_{xx}^{-1}$ is positive semidefinite. In light of lemma

5.3, we need only show that

$$Q_{xx} - Q'_{zx}Q_{zz}^{-1}Q_{zx}$$

is positive semidefinite. Note that this is the probability limit of

$$\frac{1}{T}\left[\sum_{t=1}^{T} X_t X_t' - \left(\sum_{t=1}^{T} X_t Z_t'\right)\left(\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\sum_{t=1}^{T} Z_t X_t'\right)\right]$$

$$= \frac{1}{T}\cdot \mathbb{X}_T'\left(I_T - \mathbb{Z}_T\left(\mathbb{Z}_T'\mathbb{Z}_T\right)^{-1}\mathbb{Z}_T'\right)\mathbb{X}_T,$$

where

$$\mathbb{X}_T = \begin{pmatrix} X_1' \\ \vdots \\ X_T' \end{pmatrix}, \quad \mathbb{Z}_T = \begin{pmatrix} Z_1' \\ \vdots \\ Z_T' \end{pmatrix};$$

this is a positive semidefinite-valued random matrix. It follows that

$$(Q'_{zx}Q_{zz}^{-1}Q_{zx})^{-1} - Q_{xx}^{-1}$$

is also positive semidefinite, and thus $V_q - Q_{xx}^{-1}$ is positive semidefinite, indicating that $\hat{\beta}_T^{OLS}$ is asymptotically more efficient than $\hat{q}_T$.

4. **Consistent Covariance Matrix Estimation**

The estimators

$$\hat{V}_{x,T} = \hat{\sigma}_T^2 \cdot \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$$

and

$$\hat{V}_{z,T} = \hat{\sigma}_T^2 \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1}$$

are consistent estimators of the asymptotic variances

$$\sigma^2 Q_{xx}^{-1}, \quad \sigma^2 \cdot (Q'_{zx}Q_{zz}^{-1}Q_{zx})^{-1}$$

of $\hat{\beta}_T^{OLS}$ and $\hat{\beta}_T^{2SLS}$, under both H_0 and H_1. Furthermore,

$$\hat{V}_{z,T} - \hat{V}_{x,T} = \hat{\sigma}_T^2 \left(\left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} - \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \right)$$

is a positive semidefinite valued random matrix with rank equal to thatt of $(Q_{zx}' Q_{zz}^{-1} Q_{zx})^{-1} - Q_{xx}^{-1}$.

It follows that the Hausman test statistic for testing H_0 against H_1 is

$$H_T = T \cdot \left(\hat{\beta}_T^{2SLS} - \hat{\beta}_T^{OLS} \right)' \left[\hat{V}_{z,T} - \hat{V}_{x,T} \right]^{\dagger} \left(\hat{\beta}_T^{2SLS} - \hat{\beta}_T^{OLS} \right) \xrightarrow{d} \chi_r^2.$$

6.5 Heterogeneity and Local Average Treatment Effects

Here we study the somewhat related topic of estimating treatment effects in the presence of heterogeneous effects.

6.5.1 A Model of Random Coefficients

Formally, we make the following assumptions. Note that we use here the subscripts i because the average treatment effect framework is almost exclusively used in the study of cross-sectional data.

A1 : **Heterogeneous Effects**

Let $\{(Y_i, X_i, \beta_i, Z_i)\}_{i \in N_+}$ be an i.i.d. sequence of random vectors with finite second moments, where X_i and β_i are k-dimensional random vectors, while Z_i is an L-dimensional random vector. We define the errors ϵ_i as

$$Y_i = X_i' \beta_i + \epsilon_i,$$

so that our model becomes one of heterogeneous effects; the effect β_i of X_i on the dependent variable Y_i is now dependent on the individual i. We assume that $\mathbb{E}[\epsilon_i] = 0$, and that β_i has mean $\bar{\beta} = \mathbb{E}[\beta_i]$.

For later reference, we denote by V_{ab} the covariance between quantities a_i and b_i, and by μ_a the mean of a_i; for example, V_{zx} is the covariance of Z_i and X_i. We let μ_Z, μ_X be non-zero in general; this accommodates the cases where Z_i and X_i include the constant 1.

A2 : **No Perfect Collinearity**

We assume that

$$\mathbb{E}[X_i X_i'] = Q_{xx}, \quad \mathbb{E}[Z_i Z_i'] = Q_{zz},$$

where $Q_{xx} \in \mathbb{R}^{k \times k}$ and $Q_{zz} \in \mathbb{R}^{L \times L}$ are positive definite. By implication, $\sum_{i=1}^n X_i X_i'$ and $\sum_{i=1}^n Z_i Z_i'$ are positive definite with probability approaching 1 as $n \to \infty$.

A3-1 : **Exogeneity of Regressors**

This is the first of two versions of the exogeneity assumptions. Under this assumption, the regressors X_i are exogenous, so that $\mathbb{E}[\epsilon_i X_i] = \mathbf{0}$ for any $i \in N_+$. We also impose the assumption that X_i is independent of the individual effects β_i.

A3-2 : **Exogeneity of Instruments**

Under this assumption, only the instruments Z_i are exogenous, so that $\mathbb{E}[\epsilon_i Z_i] = \mathbf{0}$ for any $i \in N_+$. We actually impose the stronger assumption that Z_i is a random assignment, so

that Z_i is independent of both ϵ_i and β_i.

A4 : **Control Function Specification**

We assume that the regressors X_i and the instruments Z_i are related by the control function equation

$$X_i = \Pi_i Z_i + v_i,$$

where Π_i is a $k \times L$ valued random matrix, v_i is an error term with mean 0, and Z_i is independent of both Π_i and v_i. We also assume that Π_i has rank k almost everywhere on the sample space, and denote $\bar{\Pi} = \mathbb{E}[\Pi_i]$. Finally, we assume that v_i is independent of β_i.

The i.i.d. assumption allows us to use Khinchin's WLLN, under which we can impose the WLLN on i.i.d. sequences of random vectors that have finite first moments. Therefore, we have all the convergence in probability results that we might need right at the outset.

It is also worth summarizing the dependence relations among the variables involved under assumptions A1, A2, A3-2 and A4:

$$\begin{aligned} Z_i &\quad \text{is uncorrelated with } \epsilon_i, v_i \\ Z_i &\quad \text{is independent of } \beta_i, \Pi_i \\ v_i &\quad \text{is uncorrelated with } \beta_i \end{aligned}$$

The instruments Z_i are uncorrelated with both error terms, and independent of any random coefficients. These assumptions will be satisfied if Z_i is random, for instance, if it randomly assigns some treatment to individuals as in the draft lottery of Angrist (1990). The final assumption that v_i is uncorrelated with β_i simply means that the entirety of the (possible) interplay between X_i and β_i is captured in the term Π_i. Thus, Π_i and β_i will be correlated if X_i and β_i are correlated. More formally, as we will show later, this assumption implies that

$$\text{Cov}[X_i, \beta_i] = \text{Cov}[\mu_Z' \Pi_i, \beta_i].$$

This elucidates why we require $\mu_Z \neq \mathbf{0}$.

6.5.2 Average Treatment Effects under Exogenous Regressors

We first study the case where assumptions A1, A2, and A3-1 hold, that is, the case where the original control variables X_i are properly exogenous. This means that the ith individual's regressors X_i are uncorrelated with his error ϵ_i, and also independent of his treatment effect β_i. This is a strong assumption that will be relaxed in the next section.

Consider the OLS estimator from regressing Y_i on X_i. What is this estimator estimating in this case? To investigate this, note first that the estimator is given as

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right).$$

Using the fact that $Y_i = X_i'\beta_i + \epsilon_i$, we can expand this estimator as

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} (X_i X_i')\beta_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right).$$

By Khinchin's WLLN,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} Q_{xx}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{p} \mathbf{0}.$$

To find the probability limit of the OLS estimator, it remains to evaluate the expected value $\mathbb{E}[(X_i X_i')\beta_i]$. When the effects are homogeneous, we can pull β out from the expectation, and use the above results to conclude that $\hat{\beta}_n$ consistently estimates this effect. In our present case, the independence assumption between X_i and β_i helps us mimick this procedure. Specifically, since X_i and β_i are independent,

$$\mathbb{E}\left[(X_i X_i')\beta_i\right] = \mathbb{E}\left[X_i X_i'\right] \cdot \mathbb{E}[\beta_i] = Q_{xx} \cdot \bar{\beta},$$

and from Khinchin's WLLN we have

$$\frac{1}{n}\sum_{i=1}^{n} (X_i X_i')\beta_i \xrightarrow{p} Q_{xx} \cdot \bar{\beta}.$$

Putting all this together shows us that

$$\hat{\beta}_n \xrightarrow{p} Q_{xx}^{-1} Q_{xx} \cdot \bar{\beta} + Q_{xx}^{-1}\mathbf{0} = \bar{\beta}.$$

In other words, the OLS estimator consistently estimates the average individual effect $\bar{\beta} = \mathbb{E}[\beta_i]$. This average value is called the average treatment effect (ATE).

6.5.3 Local Average Treatment Effects under Endogeneity

We now turn to the more interesting case where assumptions A1, A2, A3-2 and A4 hold. That is, in this case the regressors X_i are not necessarily uncorrelated with the error ϵ_i, nor are they independent of the individual effects β_i. Thus, two of the three probability results above fail to hold, and the OLS estimator from regressing Y_i on X_i is no longer consistent for the ATE.

In the presence of endogenous regressors, we turn to the usual GMM methods using instrumental variables. Putting aside the problem of asymptotic efficiency, let us consider the 2SLS estimator for the regression of Y_i on X_i using Z_i as the instruments. Recall that this estimator is defined as

$$\hat{\beta}_n^{2SLS} = \left[\left(\sum_{i=1}^n Z_i X_i'\right)' \left(\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\sum_{i=1}^n Z_i X_i'\right)\right]^{-1} \left(\sum_{i=1}^n Z_i X_i'\right)' \left(\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\sum_{i=1}^n Z_i Y_i\right).$$

Using the fact that $Y_i = X_i'\beta_i + \epsilon_i$, we can expand this estimator as

$$\hat{\beta}_n^{2SLS} = \left[\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)' \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)' \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n (Z_i X_i')\beta_i\right)$$

$$+ \left[\left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)' \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i X_i'\right)' \left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n Z_i \epsilon_i\right).$$

Let us get the simple convergence results out of the way first:

$$\frac{1}{n}\sum_{i=1}^n Z_i Z_i' \xrightarrow{p} Q_{zz}, \quad \frac{1}{n}\sum_{i=1}^n Z_i \epsilon_i \xrightarrow{p} \mathbf{0}.$$

Using the control function equation,

$$X_i = \Pi_i Z_i + v_i,$$

we can see that the following relationships between the variables in the model:

1. **Cross Product of X_i and Z_i**
   Premultiplying both sides by Z_i yields

   $$Z_i X_i' = (Z_i Z_i')\Pi_i' + Z_i v_i'$$

   for any $i \in N_+$,so taking expectations on both sides yields

   $$\mathbb{E}\left[Z_i X_i'\right] = \mathbb{E}\left[(Z_i Z_i')\Pi_i'\right] + \mathbb{E}\left[Z_i v_i'\right] = \mathbb{E}\left[Z_i Z_i'\right] \cdot \mathbb{E}\left[\Pi_i\right]' = Q_{zz} \cdot \bar{\Pi}'$$

where we used the fact that Z_i is independent of both Π_i and v_i. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} Z_i X_i' \xrightarrow{p} Q_{zz} \cdot \bar{\Pi}'.$$

Meanwhile, taking expectations on both sides of the control equation shows us that

$$\mu_X = \mathbb{E}[\Pi_i \cdot Z_i] = \bar{\Pi}\mu_Z,$$

where we used the fact that Π_i and Z_i are independent. Thus,

$$\begin{aligned} V_{zx} = \operatorname{Cov}[Z_i, X_i] &= \mathbb{E}\left[Z_i X_i'\right] - \mu_Z \cdot \mu_X' \\ &= \underbrace{\left(Q_{zz} - \mu_Z \mu_Z'\right)}_{V_{zz}} \cdot \bar{\Pi}'. \end{aligned}$$

2. **Cross Product of Z_i, X_i and β_i**
   First define

$$\bar{C} = \mathbb{E}\left[(\Pi_i - \bar{\Pi})'(\beta_i - \bar{\beta})\right],$$

so that

$$\mathbb{E}\left[\Pi_i' \beta_i\right] = \bar{C} + \bar{\Pi}'\bar{\beta}.$$

Now premultiply by Z_i and postmultiply by β_i to obtain

$$(Z_i X_i')\beta_i = (Z_i Z_i')\Pi_i'\beta_i + Z_i v_i'\beta_i.$$

taking expectations on both sides yields

$$\begin{aligned} \mathbb{E}\left[(Z_i X_i')\beta_i\right] &= \mathbb{E}\left[Z_i Z_i'\right] \cdot \mathbb{E}\left[\Pi_i'\beta_i\right] + \mathbb{E}\left[Z_i\right] \cdot \mathbb{E}\left[v_i'\beta_i\right] \\ &= Q_{zz} \cdot \left(\bar{C} + \bar{\Pi}'\bar{\beta}\right), \end{aligned}$$

where we used the fact that $\mathbb{E}[v_i'\beta_i] = 0$ since v_i and β_i are uncorrelated and the former

has mean 0. By implication,

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i X_i')\beta_i \xrightarrow{p} Q_{zz}\cdot\left[\bar{C}+\bar{\Pi}'\bar{\beta}\right].$$

3. **Cross Product of Y_i and Z_i**

   From the

   Post-multiplying both sides of the control equation by β_i yields

$$Y_i-\epsilon_i = X_i'\beta_i = Z_i'\Pi_i'\beta_i + v_i'\beta_i,$$

   and taking expectations on both sides shows us that

$$\mu_Y = \mathbb{E}\left[Z_i'\Pi_i'\beta_i\right]+\mathbb{E}\left[v_i'\beta_i\right] = \mu_Z'\left(\bar{C}+\bar{\Pi}'\bar{\beta}\right).$$

   Therefore,

$$\mu_Z\cdot\mu_Y = \mu_Z\mu_Z'\cdot\left(\bar{C}+\bar{\Pi}'\bar{\beta}\right).$$

   Similarly,

$$Z_iY_i = Z_iX_i'\beta_i + Z_i\epsilon_i,$$

   and taking expectations on both sides shows us that

$$\mathbb{E}\left[Z_iY_i\right] = \mathbb{E}\left[(Z_iX_i')\beta_i\right] = Q_{zz}\cdot\left(\bar{C}+\bar{\Pi}'\bar{\beta}\right).$$

   It follows that

$$V_{zy} = \mathrm{Cov}\left[Z_i,Y_i\right]-\mu_Z\cdot\mu_Y = \underbrace{(Q_{zz}-\mu_Z\mu_Z')}_{V_{zz}}\cdot\left(\bar{C}+\bar{\Pi}'\bar{\beta}\right).$$

4. **Cross Product of X_i and β_i**

   Finally, as promised earlier, we show that the correlation between X_i and β_i is fully captured in the correlation between Π_i and β_i. Recall that

$$\mu_X = \bar{\Pi}\cdot\mu_Z,$$

and that

$$\mathbb{E}\left[X_i'\beta_i\right] = \mu_Y = \mu_Z'\left(\bar{C} + \bar{P}\bar{i}'\bar{\beta}\right).$$

It follows that

$$\begin{aligned}\operatorname{Cov}\left[X_i, \beta_i\right] &= \mathbb{E}\left[X_i'\beta_i\right] - \mu_X'\bar{\beta} \\ &= \mu_Z'\left(\bar{C} + \bar{P}\bar{i}'\bar{\beta}\right) - \mu_Z' \cdot \bar{\Pi}'\bar{\beta} \\ &= \mu_Z' \cdot \bar{C} = \operatorname{Cov}\left[\mu_Z'\Pi_i, \beta_i\right].\end{aligned}$$

Putting these results together shows us that

$$\begin{aligned}\hat{\beta}_n^{2SLS} &\xrightarrow{p} \left(\bar{\Pi}Q_{zz}\bar{\Pi}'\right)^{-1}\bar{\Pi}Q_{zz}\left(\bar{C} + \bar{\Pi}'\bar{\beta}\right) + \left(\bar{\Pi}Q_{zz}\bar{\Pi}'\right)^{-1}\bar{\Pi}\cdot\mathbf{0} \\ &= \left(\bar{\Pi}Q_{zz}\bar{\Pi}'\right)^{-1}\bar{\Pi}Q_{zz}\left(\bar{C} + \bar{\Pi}'\bar{\beta}\right).\end{aligned}$$

If the effects Π_i and β_i are independent, so that $\bar{C} = \mathbf{0}$, then the 2SLS estimator converges in probability to the ATE $\bar{\beta}$.

The problem is that Π_i and β_i are seldom independent; this is the case when, for instance, self-selection occurs when assigning treatments. How can we interpret the probability limit of $\hat{\beta}_n^{2SLS}$ in this case? To this end, consider the case $k = L$. Then, the probability limit of $\hat{\beta}_n^{IV}$ (since the number of instruments and the number of regressors are the same, the 2SLS estimator reduces to the IV estimator in this case) can be written as

$$\begin{aligned}\check{\beta} &= \left(\bar{\Pi}'\right)^{-1}\left(\bar{C} + \bar{\Pi}'\bar{\beta}\right) \\ &= \left(V_{zz}\cdot\bar{\Pi}'\right)^{-1}\left[V_{zz}\left(\bar{C} + \bar{\Pi}'\bar{\beta}\right)\right] = V_{zx}^{-1}V_{zy}.\end{aligned}$$

In other words, the IV estimator is consistent for the covariance of Y_i and Z_i divided by the covariance of Z_i and X_i. This can be interpreted as the average treatment effect for individuals who affect the dependent variable through the instrument Z_i; these individuals are often called compliers, and this average treatment effect is called the local average treatment effect (LATE) of the compliers.

Chapter 7

Multivariate Linear Regression Models

So far, we have studied linear regression models with a single dependent variable and (possibly) multiple regressors. In many economic problems, especially structural models, we are interested in the interplay between multiple endogenous variables, and in these cases we need to extend the univariate regression framework to a multivariate one. In this chapter we briefly introduce and study the properties of the simultaneous equations model, which is the most general framework under which we multivariate regressions can be studied. Afterward, we introduce some methods to estimate the model parameterrs, including maximum likelihood methods and multiple equation GMM.

7.1 Vectorization and Kronecker Product

The vectorization operator and the kronecker product appear very frequently in the theory of multivariate linear regressions. Here we define these operations and present some of their most useful properties.

Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$. The Kronecker product $A \bigotimes B$ of A and B is the $mp \times pq$ matrix defined as

$$A \bigotimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix}.$$

Clearly, the Kronecker product is not commutative. The following are additional properties of the Kronecker product:

Lemma 7.1 (Properties of the Kronecker Product)
Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times q}$, $C \in \mathbb{C}^{k\times l}$, and $D \in \mathbb{C}^{s\times r}$. Then, the following hold true:

i) $(A \bigotimes B) \bigotimes C = A \bigotimes (B \bigotimes C)$

ii) If $m = p$, $n = q$, $k = s$ and $l = r$,

$$(A+B) \bigotimes (C+D) = A \bigotimes C + B \bigotimes C + A \bigotimes D + B \bigotimes D$$

iii) If $n = p$ and $l = s$, then

$$(A \bigotimes C)(B \bigotimes D) = AB \bigotimes CD$$

iv) If A and B are nonsingular square matrices, then

$$(A \bigotimes B)^{-1} = (A^{-1} \bigotimes B^{-1})$$

v) If A and B are square matrices, then

$$\mathrm{tr}\left(A \bigotimes B\right) = \mathrm{tr}(A)\,\mathrm{tr}(B)$$

vi) Let A be an $m \times m$ square matrix with eigenvalues $\lambda_1, \cdots, \lambda_m \in \mathbb{C}$, and B a $p \times p$ square matrix with eigenvalues $\mu_1, \cdots, \mu_p \in \mathbb{C}$. Then, the eigenvalues of $A \bigotimes B$ are collected in the set

$$\{\lambda_i \mu_j \mid 1 \le i \le m, 1 \le j \le p\}.$$

vii) Let A be an $m \times m$ square matrix and B a $p \times p$ square matrix. Then,

$$\left|A \bigotimes B\right| = |A|^p |B|^m.$$

Proof) i) This follows immediately by noting that

$$\begin{aligned}(A \bigotimes B) \bigotimes C &= \begin{pmatrix} A_{11}B & \cdots & A_{1n}B \\ \vdots & \ddots & \vdots \\ A_{m1}B & \cdots & A_{mn}B \end{pmatrix} \bigotimes C \\ &= \begin{pmatrix} A_{11}(B \bigotimes C) & \cdots & A_{1n}(B \bigotimes C) \\ \vdots & \ddots & \vdots \\ A_{m1}(B \bigotimes C) & \cdots & A_{mn}(B \bigotimes C) \end{pmatrix} = A \bigotimes (B \bigotimes C).\end{aligned}$$

ii) This follows immediately by noting that

$$(A+B)\bigotimes(C+D) = \begin{pmatrix} (A_{11}+B_{11})(C+D) & \cdots & (A_{1n}+B_{1n})(C+D) \\ \vdots & \ddots & \vdots \\ (A_{m1}+B_{m1})(C+D) & \cdots & (A_{mn}+B_{mn})(C+D) \end{pmatrix}$$

$$= \begin{pmatrix} A_{11}C+B_{11}C+A_{11}D+B_{11}D & \cdots & A_{1n}C+B_{1n}C+A_{1n}D+B_{1n}D \\ \vdots & \ddots & \vdots \\ A_{m1}C+B_{m1}C+A_{m1}D+B_{m1}D & \cdots & A_{mn}C+B_{mn}C+A_{mn}D+B_{mn}D \end{pmatrix}$$

$$= A\bigotimes C + B\bigotimes C + A\bigotimes D + B\bigotimes D.$$

iii) This also follows immediately by noting that

$$(A\bigotimes C)(B\bigotimes D) = \begin{pmatrix} A_{11}C & \cdots & A_{1n}C \\ \vdots & \ddots & \vdots \\ A_{m1}C & \cdots & A_{mn}C \end{pmatrix} \begin{pmatrix} B_{11}D & \cdots & B_{1q}D \\ \vdots & \ddots & \vdots \\ B_{n1}D & \cdots & B_{nq}D \end{pmatrix}$$

$$= \begin{pmatrix} (\sum_{i=1}^{n} A_{1i}B_{i1})CD & \cdots & (\sum_{i=1}^{n} A_{1i}B_{iq})CD \\ \vdots & \ddots & \vdots \\ (\sum_{i=1}^{n} A_{mi}B_{i1})CD & \cdots & (\sum_{i=1}^{n} A_{mi}B_{iq})CD \end{pmatrix}$$

$$= AB\bigotimes CD.$$

iv) Suppose A and B are nonsingular square matrices. Then, by the preceding result,

$$(A\bigotimes B)(A^{-1}\bigotimes B^{-1}) = (I_m\bigotimes I_p) = I_{mp},$$

so that the inverse of $A\bigotimes B$ exists and is given as

$$(A\bigotimes B)^{-1} = (A^{-1}\bigotimes B^{-1}).$$

v) Suppose A and B are again nonsingular square matrices. Then,

$$\operatorname{tr}(A\bigotimes B) = \sum_{i=1}^{m} A_{ii}\operatorname{tr}(B) = \operatorname{tr}(A)\operatorname{tr}(B).$$

vi) Let the Schur decompositions of A and B be given as

$$A = PDP^{-1} \quad \text{and} \quad B = Q\Lambda Q^{-1},$$

where P, Q are unitary matrices, so that their inverses are their conjugate transposes, and D, Λ are upper triangular matrices. Since the characteristic polynomial of A and B are given by

$$ch_A(z) = |A - z \cdot I_m| = |D - z \cdot I_m| = \prod_{i=1}^{m}(D_{ii} - z)$$
$$ch_B(z) = |B - z \cdot I_p| = |\Lambda - z \cdot I_p| = \prod_{i=1}^{p}(\Lambda_{ii} - z).$$

Therefore, the diagonal entries of D and Λ are the eigenvalues of A and B. We can now see that

$$\begin{aligned} A \bigotimes B &= (PDP^{-1}) \bigotimes (Q\Lambda Q^{-1}) \\ &= (P \bigotimes Q)((DP^{-1}) \bigotimes (\Lambda Q^{-1})) \\ &= (P \bigotimes Q)(D \bigotimes \Lambda)(P^{-1} \bigotimes Q^{-1}) \\ &= (P \bigotimes Q)(D \bigotimes \Lambda)(P \bigotimes Q)^{-1}. \end{aligned}$$

Therefore, the characteristic polynomial of $A \bigotimes B$ is given as

$$\begin{aligned} ch_{A \bigotimes B}(z) &= \left|(P \bigotimes Q)(D \bigotimes \Lambda)(P \bigotimes Q)^{-1} - z \cdot I_{mp}\right| \\ &= \left|(D \bigotimes \Lambda) - z \cdot I_{mp}\right| \\ &= \prod_{i=1}^{m}\prod_{j=1}^{p}(D_{ii}\Lambda_{jj} - z) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{p}(\lambda_i \mu_j - z) \end{aligned}$$

The eigenvalues of $A \bigotimes B$ are thus collected in the set

$$\{\lambda_i \mu_j \mid 1 \leq i \leq m, 1 \leq j \leq p\}.$$

vii) In light of the above result, we can see that

$$\begin{aligned} \left|A \bigotimes B\right| = ch_{A \bigotimes B}(0) &= \prod_{i=1}^{m} \lambda_i \left(\prod_{j=1}^{p} \mu_j\right) \\ &= \left(\prod_{i=1}^{m} \lambda_i\right)^p \left(\prod_{j=1}^{p} \mu_j\right)^m = |A|^p |B|^m. \end{aligned}$$

Q.E.D.

The vectorization operator stacks the columns of a matrix to transform it into a vector. Formally, for any $A \in \mathbb{C}^{m\times n}$, the vectorization of A is the mn-dimensional vector defined as

$$\mathrm{vec}\,(A) = \begin{pmatrix} A_1 \\ \vdots \\ A_n \end{pmatrix},$$

where $A_1, \cdots, A_n \in \mathbb{C}^m$ are the columns of A. The vec operator has very useful properties, especially in conjunction with the Kronecker product.

Lemma 7.2 (Properties of the Vectorization Operator)
Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{p\times q}$, $C \in \mathbb{C}^{k\times l}$, and $D \in \mathbb{C}^{s\times r}$. Then, the following hold true:

i) If $n = q = 1$, we have

$$\mathrm{vec}\,(AB') = B \bigotimes A.$$

ii) If $m = p$ and $n = q$, we have

$$\mathrm{vec}\,(A)'\,\mathrm{vec}\,(B) = \mathrm{tr}\,(A'B).$$

iii) If $n = p$ and $q = k$, we have

$$\mathrm{vec}\,(ABC) = (C' \bigotimes A)\mathrm{vec}\,(B).$$

iv) Suppose the matrix product $ABCD$ is well-defined. Then,

$$\mathrm{tr}\,(ABCD) = \mathrm{vec}\,(D')'\,(C' \bigotimes A)\mathrm{vec}\,(B).$$

Proof) i) If A, B are column vectors, then

$$\mathrm{vec}\,(AB') = \mathrm{vec}\begin{pmatrix} A_{11}B_{11} & \cdots & A_{11}B_{p1} \\ \vdots & \ddots & \vdots \\ A_{m1}B_{11} & \cdots & A_{m1}B_{p1} \end{pmatrix}$$
$$= \begin{pmatrix} A_{11}B_{11} \\ \vdots \\ A_{m1}B_{11} \\ \vdots \\ A_{11}B_{p1} \\ \vdots \\ A_{m1}B_{p1} \end{pmatrix} = \begin{pmatrix} B_{11}A \\ \vdots \\ B_{p1}A \end{pmatrix} = B \bigotimes A.$$

ii) If A and B are matrices sharing the same dimensions, then

$$\text{vec}\,(A)'\,\text{vec}\,(B) = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}B_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} A'_{ji}B_{ij} = \text{tr}\,(A'B)\,.$$

iii) Suppose $n = p$ and $q = k$, so that the matrix product ABC is well-defined. Letting $B_1, \cdots, B_q \in \mathbb{C}^p$ be the columns of B, and $\{e_1, \cdots, e_p\} \subset \mathbb{R}^p$ the standard basis of $\mathbb{R}^p$,

$$B = \sum_{i=1}^{q} B_i e_i'.$$

It follows that

$$\begin{aligned}
\text{vec}\,(ABC) &= \sum_{i=1}^{q} \text{vec}\,(AB_i \cdot e_i'C) \\
&= \sum_{i=1}^{q} (C'e_i) \bigotimes (AB_i) \\
&= (C' \bigotimes A) \left(\sum_{i=1}^{q} (e_i \bigotimes B_i) \right) \\
&= (C' \bigotimes A) \left(\sum_{i=1}^{q} \text{vec}\,(B_i e_i') \right) \\
&= (C' \bigotimes A)\text{vec}\,(B)\,,
\end{aligned}$$

where we used the preceding result on vectorization and one of the properties of the Kronecker product.

iv) Suppose that the matrix product $ABCD$ is well-defined. In this case, the previous results tell us that

$$\text{vec}\,(D')'\,(C' \bigotimes A)\text{vec}\,(B) = \text{vec}\,(D')'\,\text{vec}\,(ABC) = \text{tr}(DABC) = \text{tr}(ABCD).$$

Q.E.D.

7.2 Convergence Results with MDS Errors

Multivariate regressions appear often in time series analysis, which deals almost exclusively with dependent processes. Here, we present some results pertaining to regressions involving dependent processes, specifically regressions in which the dependent variables may be regressed on their lagged values. In these regressions, it is customary to model the error terms as martingale difference sequences, which are a class of dependent processes that includes mean zero independent processes[1]. We have actually used central limit theorems pertaining to martingale difference sequences to derive earlier results concerning regression models with independent samples. Below we state some additional convergence results involving martingale difference sequences:

Theorem 7.3 (Convergence Results for MDS Errors)
For $1 \leq i \leq n$, let $\{X_{it}\}_{t\in\mathbb{Z}}$ be a k_i-dimensional real stochastic process. Define $K = \sum_{i=1}^{n} k_i$ and the $K \times n$ matrix valued stochastic process $\{X_t\}_{t\in\mathbb{Z}}$ as

$$X_t = \begin{pmatrix} X_{1t} & \cdots & O_{k_1\times 1} \\ \vdots & \ddots & \vdots \\ O_{k_n\times 1} & \cdots & X_{nt} \end{pmatrix}$$

for any $t \in \mathbb{Z}$. Let $\{\epsilon_t\}_{t\in\mathbb{Z}}$ be an n-dimensional real stochastic process, and $\mathcal{F}$ the filtration generated by $\{(X_{t+1}, \epsilon_t)\}_{t\in\mathbb{Z}}$. We make the following assumptions:

i) **MDS Errors**
$\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to $\mathcal{F}$, so that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is $\mathcal{F}$-adapted, integrable, and satisfies

$$\mathbb{E}\left[\epsilon_t \mid \mathcal{F}_{t-1}\right] = O_{n\times 1}$$

for any $t \in \mathbb{Z}$.

ii) **Homoskedasticity**
$\{\epsilon_t\}_{t\in\mathbb{Z}}$ is homoskedastic in the sense that there exists some positive definite $n \times n$ matrix Σ such that

$$\mathbb{E}\left[\epsilon_t\epsilon_t' \mid \mathcal{F}_{t-1}\right] = \Sigma$$

for any $t \in \mathbb{Z}$. Let the (i,j)th element of Σ be denoted by σ_{ij}.

iii) **Asymptotic Behavior of Regressors**

[1]For a review of martingale difference sequence theory, consult chapter 5 of the probability theory text that deals with WLLNs and CLTs for dependent processes.

For any $1 \leq i, j \leq n$, there exists a $k_i \times k_j$ matrix Q_{ij} such that

$$\frac{1}{T}\sum_{t=1}^{T} X_{it}X_{jt}' \xrightarrow{p} Q_{ij}.$$

Defining

$$\Omega_Q = \begin{pmatrix} \sigma_{11}Q_{11} & \cdots & \sigma_{1n}Q_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1}Q_{n1} & \cdots & \sigma_{nn}Q_{nn} \end{pmatrix} \in \mathbb{R}^{K\times K},$$

assume that Ω_Q is a positive definite matrix.

iv) **Boundedness of Moments**

For any $1 \leq i \leq n$, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ and $\{X_{it}\}_{t\in\mathbb{Z}}$ have finite fourth moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}\|X_t\|^4 < +\infty.$$

In addition, there exists some $\delta > 0$ such that, for any $1 \leq i, j \leq n$,

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|\epsilon_{it}\epsilon_{jt}X_{it}X_{jt}'\right\|^{1+\delta} < +\infty.$$

Then, we have

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t\epsilon_t' \xrightarrow{p} \Sigma$$

$$\frac{1}{T}\sum_{t=1}^{T} (X_t\epsilon_t)(X_t\epsilon_t)' \xrightarrow{p} \Omega_Q$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t\epsilon_t \xrightarrow{d} \mathcal{N}\left[O_{K\times 1},\, \Omega_Q\right].$$

Proof) We can show that each of the three results hold through very similar processes. First, we show that the stochastic processes involved are, under suitable normalizations, martingale difference sequences with respect to the appropriate filtration. Then, we show that they satisfy the conditions either of Andrews' margingale WLLN[2] or the MDS CLT[3]. The desired result then follows immediately from an application of the relevant theorem.

[2]See theorem 5.7 of the probability theory text.
[3]See theorem 5.8 of the probability theory text.

Outer Product of Error Terms

Choose any $1 \leq i, j \leq n$. $\{\epsilon_{it}\epsilon_{jt} - \sigma_{ij}\}_{t\in\mathbb{Z}}$ is an MDS with respect to $\mathcal{F}$. First, since each ϵ_t is $\mathcal{F}_t$-measurable by design, $\{\epsilon_{it}\epsilon_{jt} - \sigma_{ij}\}_{t\in\mathbb{Z}}$ is $\mathcal{F}$-adapted. In addition,

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_{it}\epsilon_{jt}|^2 \leq \sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty,$$

so $\{\epsilon_{it}\epsilon_{jt} - \sigma_{ij}\}_{t\in\mathbb{Z}}$ is L^2-bounded. The L^2-boundedness of $\{\epsilon_{it}\epsilon_{jt} - \sigma_{ij}\}_{t\in\mathbb{Z}}$ further implies that it is uniformly integrable by the De la Valleé Poussin characterization of uniform integrability. Finally, for any $t \in \mathbb{Z}$,

$$\mathbb{E}\left[\epsilon_{it}\epsilon_{jt} - \sigma_{ij} \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[\epsilon_{it}\epsilon_{jt} \mid \mathcal{F}_{t-1}\right] - \sigma_{ij} = 0$$

by the homoskedasticity assumption. This proves that $\{\epsilon_{it}\epsilon_{jt} - \sigma_{ij}\}_{t\in\mathbb{Z}}$ is a uniformly integrable MDS with respect to $\mathcal{F}$, and thus, by Andrews' martingale WLLN,

$$\frac{1}{T}\sum_{t=1}^{T} (\epsilon_{it}\epsilon_{jt} - \sigma_{ij}) \xrightarrow{p} 0,$$

or equivalently,

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}\epsilon_{jt} \xrightarrow{p} \sigma_{ij}.$$

This holds for any $1 \leq i, j \leq n$, so we have

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t\epsilon_t \xrightarrow{p} \Sigma.$$

Outer Product of Cross Product of Error and Regressors

Note that

$$X_t\epsilon_t = \begin{pmatrix} X_{1t}\epsilon_{1t} \\ \vdots \\ X_{nt}\epsilon_{nt} \end{pmatrix}$$

for any $t \in \mathbb{Z}$, so that the outer product $(X_t\epsilon_t)(X_t\epsilon_t)'$ is given as

$$(X_t\epsilon_t)(X_t\epsilon_t)' = \begin{pmatrix} \epsilon_{1t}^2 X_{1t}X_{1t}' & \vdots & \epsilon_{1t}\epsilon_{nt}X_{1t}X_{nt}' \\ \vdots & \ddots & \vdots \\ \epsilon_{nt}\epsilon_{1t}X_{nt}X_{1t}' & \vdots & \epsilon_{nt}^2 X_{nt}X_{nt}' \end{pmatrix}.$$

As such, we need only focus our attention on each block $\epsilon_{it}\epsilon_{jt}X_{it}X_{jt}'$ for any $1 \leq i, j \leq n$.

Choose any $1 \leq i, j \leq n$. We will show that $\{(\epsilon_{it}\epsilon_{jt} - \sigma_{ij})X_{it}X_{jt}'\}_{t\in\mathbb{Z}}$ is an MDS with respect to the filtration $\mathcal{F}$. This process is clearly $\mathcal{F}$-adapted, since each X_t and ϵ_t are $\mathcal{F}_t$-measurable by assumption. It is also $L^{1+\delta}$-bounded, since

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|\epsilon_{it}\epsilon_{jt}X_{it}X_{jt}'\right\|^{1+\delta} < +\infty$$

by assumption, and $\{X_t\}_{t\in\mathbb{Z}}$ has finite fourth moments. By the De la Valleé Poussin characterization of uniform integrability, $\{(\epsilon_{it}\epsilon_{jt} - \sigma_{ij})X_{it}X_{jt}'\}_{t\in\mathbb{Z}}$ is uniformly integrable. Finally, for any $t \in \mathbb{Z}$,

$$\begin{aligned}\mathbb{E}\left[(\epsilon_{it}\epsilon_{jt} - \sigma_{ij})X_{it}X_{jt}' \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\epsilon_{it}\epsilon_{jt} - \sigma_{ij} \mid \mathcal{F}_{t-1}\right] \cdot X_{it}X_{jt}' \\ &= \left[\mathbb{E}\left[\epsilon_{it}\epsilon_{jt} \mid \mathcal{F}_{t-1}\right] - \sigma_{ij}\right] \cdot X_{it}X_{jt}' = O_{k_i\times k_j}.\end{aligned}$$

We have just shown that $\{(\epsilon_{it}\epsilon_{jt} - \sigma_{ij})X_{it}X_{jt}'\}_{t\in\mathbb{Z}}$ is a uniformly integrable MDS with respect to $\mathcal{F}$; by Andrews' martingale WLLN,

$$\frac{1}{T}\sum_{t=1}^{T}\left[(\epsilon_{it}\epsilon_{jt} - \sigma_{ij})X_{it}X_{jt}'\right] \xrightarrow{p} O_{k_i\times k_j}.$$

Since

$$\sigma_{ij} \cdot \frac{1}{T}\sum_{t=1}^{T} X_{it}X_{jt}' \xrightarrow{p} \sigma_{ij}Q_{ij}$$

by assumption, we can see that

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}\epsilon_{jt}X_{it}X_{jt}' \xrightarrow{p} \sigma_{ij}Q_{ij}.$$

This holds for any $1 \leq i, j \leq n$, so

$$\frac{1}{T}\sum_{t=1}^{T}(X_t\epsilon_t)(X_t\epsilon_t)' \xrightarrow{p} \Omega_Q.$$

Cross Product of Error and Regressors

Finally, we move onto the process $\{X_t\epsilon_t\}_{t\in\mathbb{Z}}$. This is also a martingale difference sequence with respect to $\mathcal{F}$: since X_t and ϵ_t are $\mathcal{F}_t$-measurable, it is clearly $\mathcal{F}$-adapted.

Furthermore, it is $L^{2+2\delta}$-bounded:

$$\begin{aligned}
\mathbb{E}|X_t\epsilon_t|^{2+2\delta} &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\epsilon_{it}^2 X_{it}'X_{it}'\right)^{1+\delta}\right] \\
&\leq \left(\sum_{i=1}^{n}\left(\mathbb{E}\left(\epsilon_{it}^2 X_{it}'X_{it}\right)^{1+\delta}\right)^{\frac{1}{1+\delta}}\right)^{1+\delta} \qquad \text{(Minkowski's inequality)} \\
&= \left(\sum_{i=1}^{n}\left(\mathbb{E}\left\|\epsilon_{it}^2 X_{it}X_{it}'\right\|^{1+\delta}\right)^{\frac{1}{1+\delta}}\right)^{1+\delta}
\end{aligned}$$

for any $t \in \mathbb{Z}$, where we used the fact that

$$\begin{aligned}
\left\|X_{it}X_{it}'\right\| &= \left[\operatorname{tr}\left(X_{it}X_{it}'X_{it}X_{it}'\right)\right]^{\frac{1}{2}} \\
&= \left(|X_{it}|^4\right)^{\frac{1}{2}} = |X_{it}|^2 = X_{it}'X_{it}.
\end{aligned}$$

for each $1 \leq i \leq n$. Therefore,

$$\sup_{t\in\mathbb{Z}}\mathbb{E}|X_t\epsilon_t|^{2+2\delta} \leq \left(\sum_{i=1}^{n}\left(\sup_{t\in\mathbb{Z}}\mathbb{E}\left\|\epsilon_{it}^2 X_{it}X_{it}'\right\|^{1+\delta}\right)^{\frac{1}{1+\delta}}\right)^{1+\delta} < +\infty,$$

where the last inequality follows by assumption, which shows that the desired boundedness assumption holds.

For any $t \in \mathbb{Z}$,

$$\begin{aligned}
\mathbb{E}\left[X_t\epsilon_t \mid \mathcal{F}_{t-1}\right] &= X_t \cdot \mathbb{E}\left[\epsilon_t \mid \mathcal{F}_{t-1}\right] \\
&= X_t \cdot \mathbb{E}\left[\epsilon_t \mid \mathcal{F}_{t-1}\right] \\
&= X_t \cdot O_{n\times 1} = O_{K\times 1}.
\end{aligned}$$

This shows us that $\{X_t\epsilon_t\}_{t\in\mathbb{Z}}$ is an $L^{2+2\delta}$-bounded martingale difference sequence with respect to $\mathcal{F}$. Finally, we established above that

$$\frac{1}{T}\sum_{t=1}^{T}(X_t\epsilon_t)(X_t\epsilon_t)' \xrightarrow{p} \Omega_Q,$$

where Ω_Q is positive definite by assumption. Thus, by the MDS CLT,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}X_t\epsilon_t \xrightarrow{p} \mathcal{N}\left[O_{K\times 1},\, \Omega_Q\right].$$

Q.E.D.

Corollary to Theorem 7.3 Let $\{X_t\}_{t\in\mathbb{Z}}$ be a k-dimensional real stochastic process. Let $\{\epsilon_t\}_{t\in\mathbb{Z}}$ be an n-dimensional real stochastic process, and $\mathcal{F}$ the filtration generated by $\{(X_{t+1},\epsilon_t)\}_{t\in\mathbb{Z}}$. We make the following assumptions:

i) **MDS Errors**
$\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to $\mathcal{F}$, so that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is $\mathcal{F}$-adapted, integrable, and satisfies

$$\mathbb{E}\left[\epsilon_t \mid \mathcal{F}_{t-1}\right] = O_{n\times 1}$$

for any $t \in \mathbb{Z}$.

ii) **Homoskedasticity**
$\{\epsilon_t\}_{t\in\mathbb{Z}}$ is homoskedastic in the sense that there exists some positive definite $n \times n$ matrix Σ such that

$$\mathbb{E}\left[\epsilon_t\epsilon_t' \mid \mathcal{F}_{t-1}\right] = \Sigma$$

for any $t \in \mathbb{Z}$. Let the (i,j)th element of Σ be denoted by σ_{ij}.

iii) **Asymptotic Behavior of Regressors**
There exists a positive definite $k \times k$ matrix Q such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q.$$

iv) **Boundedness of Moments**
$\{\epsilon_t\}_{t\in\mathbb{Z}}$ and $\{X_t\}_{t\in\mathbb{Z}}$ have finite fourth moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}|X_t|^4 < +\infty.$$

In addition, there exists some $\delta > 0$ such that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|\epsilon_t\epsilon_t' \bigotimes X_t X_t'\right\|^{1+\delta} < +\infty.$$

Then, we have

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\epsilon_t' \xrightarrow{p} \Sigma$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{vec}\left(X_t\epsilon_t'\right)\mathrm{vec}\left(X_t\epsilon_t'\right)' \xrightarrow{p} \Sigma\bigotimes Q$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathrm{vec}\left(X_t\epsilon_t'\right) \xrightarrow{d} \mathcal{N}\left[O_{nk\times 1},\ \Sigma\bigotimes Q\right].$$

Proof) This is a special case of the setting in theorem 7.3 where $X_{it} = X_{jt} = X_t$ for any $1 \le i, j \le n$. Define

$$\mathbb{X}_t = I_n \bigotimes X_t,$$

and $K = nk$. Then,

$$\mathbb{X}_t\epsilon_t = \begin{pmatrix} X_t\epsilon_{1t} \\ \vdots \\ X_t\epsilon_{nt} \end{pmatrix} = \mathrm{vec}\left(X_t\epsilon_t'\right)$$

and

$$(\mathbb{X}_t\epsilon_t)(\mathbb{X}_t\epsilon_t)' = \mathrm{vec}\left(X_t\epsilon_t'\right)\mathrm{vec}\left(X_t\epsilon_t'\right)'.$$

Furthermore,

$$\Omega_Q = \begin{pmatrix} \sigma_{11}Q & \cdots & \sigma_{1n}Q \\ \vdots & \ddots & \vdots \\ \sigma_{n1}Q & \vdots & \sigma_{nn}Q \end{pmatrix} = \Sigma\bigotimes Q,$$

which is positive definite because Σ and Q are positive definite. Finally, since

$$\|\mathbb{X}_t\|^2 = \mathrm{tr}\left(\mathbb{X}_t\mathbb{X}_t'\right) = n\cdot|X_t|^2$$

for any $t \in \mathbb{Z}$, we have

$$\sup_{t\in\mathbb{Z}}\mathbb{E}\|\mathbb{X}_t\|^4 = n^2\cdot\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|X_t|^4\right) < +\infty,$$

and since, for any $1 \le i, j \le n$ and $t \in \mathbb{Z}$,

$$\mathbb{E}\left\|\epsilon_{it}\epsilon_{jt}X_tX_t'\right\|^{1+\delta} \le \mathbb{E}\left\|\epsilon_t\epsilon_t'\bigotimes X_tX_t'\right\|^{1+\delta},$$

we have

$$\sup_{t\in\mathbb{Z}}\mathbb{E}\left\|\epsilon_{it}\epsilon_{jt}X_{it}X_{jt}'\right\|^{1+\delta}=\sup_{t\in\mathbb{Z}}\mathbb{E}\left\|\epsilon_t\epsilon_t\bigotimes X_tX_t'\right\|^{1+\delta}<+\infty.$$

Therefore, by theorem 7.3,

$$\frac{1}{T}\sum_{t=1}^{T}\text{vec}\left(X_t\epsilon_t'\right)\text{vec}\left(X_t\epsilon_t'\right)'=\frac{1}{T}\sum_{t=1}^{T}(\mathbb{X}_t\epsilon_t)(\mathbb{X}_t\epsilon_t)'$$

$$\xrightarrow{p}\Omega_Q=\Sigma\bigotimes Q$$

and

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\left(X_t\epsilon_t'\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbb{X}_t\epsilon_t$$

$$\xrightarrow{d}\mathcal{N}\left[O_{K\times1},\Omega_Q\right]=\mathcal{N}\left[O_{nk\times1},\ \Sigma\bigotimes Q\right].$$

Q.E.D.

7.3 Weak Instruments

Before we study multivariate regression models, we first take a detour to illustrate the power of the asymptotic machinery developed above via the weak instruments problem. We say a univariate or multivariate linear regression model suffers from the weak instrument problem if the instruments are only weakly correlated with the original regressors; in other words, the weak instrument problem arises when the relevance condition only barely holds.

The weak instrument problem can be formulated in terms of a control function approach. Consider the single equation GMM model that we studied in the preceding chapter, with dependent variable Y_t, regressors X_t and instruments Z_t. There, we assumed that Y_t is linearly related to X_t as

$$Y_t = X_t'\beta + \epsilon_t,$$

where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an independent process of exogenous shocks that may be correlated with X_t. The instruments Z_t were assumed to satisfy the exogeneity condition, meaning that $\mathbb{E}[Z_t\epsilon_t] = O_{L\times 1}$, and the relevance assumption, that is, $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ converges to a $L \times k$ matrix with rank k. Under these assumptions, we showed that the efficient GMM estimator of β using the instruments Z_t are both consistent and asymptotically normal.

Below, we will specify the relationship between Z_t and X_t via a control function equation, that is, we assume that they are llnearly related as

$$X_t = \Pi Z_t + u_t,$$

where u_t is mean zero and uncorrelated with Z_t. This essentially means that we are imposing the additional assumption that the conditional expectation of X_t given Z_t is linear in Z_t. Under this specification,

$$\frac{1}{T}\sum_{t=1}^{T} Z_t X_t' = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)\Pi' + \frac{1}{T}\sum_{t=1}^{T} Z_t u_t,$$

and if we assume that $\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'$ converges in probability to some positive definite $L \times L$ matrix Q_{zz}, it follows under the usual regularity assumptions that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t X_t' \xrightarrow{p} Q_{zz}\Pi'.$$

Therefore, the relevance condition is now equivalent to the $k \times L$ coefficient matrix Π having rank k.

In this context, the weak instrument problem arises when Π continues to have rank k, so that the relevance condition is satisfied, but is almost short ranked, meaning that the miniimum modulus of the eigenvalues of Π is very close to 0. At first glance, this does not seem to be a serious problem because, when all is said and done, Π still has rank k, so that the asymptotic results derived in the previous chapter continue to hold.

The problem is that, if the asymptotic covariance $Q_{zz}\Pi'$ of the regressors and instruments is nearly short ranked, then its sample analogue $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ is likely to actually be short ranked in finite samples since it is only full rank in the limit. This means that the efficient GMM estimator, not to mention any sort of GMM estimator of β using Z_t as an instrument, is undefined because it involves taking the inverse of a reduced rank square matrix.

To illustrate the problem of weak instruments in a more formal manner, econometricians usually turn to local-to-zero approximations. That is, they assume that the coefficient matrix Π converges to the zero matrix at some prespecified rate as T goes to infinity. This framework mimics the fact that the coefficient matrix is nearly short ranked by sending the minimum modulus of the eigenvalues of Π to 0 as $T \to \infty$, and allows us to study what goes wrong when we use the usual efficient GMM estimator of β even though $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ is short ranked in finite samples. It turns out that, under weak instruments, the efficient GMM estimator is not consistent for the true parameter β, but instead is spread out in the limit with β at its center. This means we must search for different means of inference when our model suffers from the weak instrument problem, two of which we introduce below.

7.3.1 The Problem with Weak Instruments

To illustrate what can happen when our instruments are weak, we use a local-to-zero approximation for the correlation between the instrument and the original regressors, that is, we assume that it converges in probability to 0 at some pre-specified rate.

Specifically, consider the following univariate IV estimation framework, where the number of instruments and regressors are the same:

A1 : **Control Function Specification**

Let the data be given as $\{Y_t, X_t, Z_t\}_{t\in\mathbb{Z}}$, where Y_t is a real random variable, and X_t, Z_t are k-dimensional random vectors. Suppose the measurement equation and control function equation are given as

$$Y_t = X_t'\beta + \epsilon_t$$

$$X_t = \Pi Z_t + u_t,$$

where $\beta \in \mathbb{R}^{k\times 1}$ is some non-random vector and $\Pi \in \mathbb{R}^{k\times k}$ is a non-random matrix of full rank.

A2 : **Exogeneity of Instruments**

Assume that $\{v_t = (\epsilon_t, u_t')'\}_{t\in\mathbb{Z}}$ is backward-looking exogenous with respect to the instruments Z_t, that is, $\{v_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to the filtration $\mathcal{F}$ generated by

$\{(Z_{t+1}, X_t, Y_t)\}_{t\in\mathbb{Z}}$. In particular,

$$\mathbb{E}\left[v_t \mid Z_t, Z_{t-1}, \cdots\right] = O_{(k+1)\times 1}$$

for any $t \in \mathbb{Z}$, which is stronger than requiring v_t to be simply contemporaneously uncorrelated with Z_t. This stronger assumption is a consequence of the possibility of dependence among the sample observations.

Note further that, becuase v_t is a function of Y_t, X_t and Z_t, the filtration $\mathcal{G}$ generated by $\{(Z_{t+1}, v_t)\}_{t\in\mathbb{Z}}$ is coarser than $\mathcal{F}$, that is, $\mathcal{G}_t \subset \mathcal{F}_t$ for any $t \in \mathbb{Z}$. Therefore, $\{v_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}$, so that the assumptions of theorem 7.3 are satisfied.

A3 : **Homoskedasticity**

There exists a positive definite $(k+1)\times(k+1)$ matrix Σ such that

$$\mathbb{E}\left[v_t v_t' \mid \mathcal{F}_{t-1}\right] = \underbrace{\begin{pmatrix} \sigma_e^2 & \sigma_{eu}' \\ \sigma_{eu} & \Sigma_u \end{pmatrix}}_{\Sigma}$$

for any $t \in \mathbb{Z}$.

A4 : **Regularity Assumptions**

We assume that there exists a positive definite matrix $Q_{zz} \in \mathbb{R}^{k\times k}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t' \xrightarrow{p} Q_{zz}.$$

Furthermore, we assume the following moment boundedness conditions hold:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|v_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}|Z_t|^4 < +\infty,$$

and there exists a $\delta > 0$ such that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\| v_t v_t' \bigotimes Z_t Z_t' \right\|^{1+\delta} < +\infty.$$

Finally, assume that $\sum_{t=1}^{T} Z_t X_t'$ is non-singiular everywhere on the sample space for any $T \in N_+$.

A5 : **Weak Instruments**

Assume that Π is $O(T^{-1/2})$ as $T \to \infty$, that is, there exists some full rank matrix $P \in \mathbb{R}^{k \times k}$ such that

$$\Pi = \frac{1}{\sqrt{T}} P.$$

This is our local-to-zero assumption.

Under assumptions A1 to A4, we can immediately see, from the corollary to theorem 7.3, that

$$\frac{1}{T}\sum_{t=1}^{T} v_t v_t \xrightarrow{p} \Sigma,$$

$$\frac{1}{T}\sum_{t=1}^{T} \text{vec}\left(Z_t v_t'\right) \text{vec}\left(Z_t v_t'\right)' \xrightarrow{p} \Sigma \bigotimes Q_{zz}$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \text{vec}\left(Z_t v_t'\right) \xrightarrow{d} \mathcal{N}\left[O_{k(k+1)\times 1},\ \Sigma \bigotimes Q_{zz}\right].$$

Assumption A5 is our local-to-zero framework for weak instruments; the relevance condition is nearly violated because, even though Π has full rank, under any sample size the modulus of its eigenvalues goes to 0 as $T \to \infty$, meaning that it is very close to a reduced-rank matrix under large T.

Consider the usual IV estimator

$$\hat{\beta}_T^{IV} = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right) = \beta + \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t \epsilon_t\right)$$

of β, which is well-defined for any $T \in N_+$ by our nonsingularity assumption. Under our usual relevance assumption, $\hat{\beta}_T^{IV}$ would converge to the true parameter β at a speed of $\frac{1}{\sqrt{T}}$, with $\sqrt{T}\left(\hat{\beta}_T^{IV} - \beta\right)$ having an asymptotically normal distribution.

Now we consider what happens in the weak instruments case, where Π is nearly short ranked. From the control function equation

$$X_t = \Pi Z_t + u_t,$$

it follows that

$$\sum_{t=1}^{T} X_t Z_t' = P\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t Z_t'\right) + \sum_{t=1}^{T} u_t Z_t'.$$

Therefore,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t Z_t' = P\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} u_t Z_t',$$

or in vectorized form,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \text{vec}\left(Z_t X_t'\right) = \left[I_k \bigotimes \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)\right] \cdot \text{vec}\left(P'\right) + \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \text{vec}\left(Z_t u_t'\right).$$

Since

$$\text{vec}\left(Z_t v_t'\right) = \begin{pmatrix} Z_t \epsilon_t \\ \text{vec}\left(Z_T u_t'\right), \end{pmatrix}$$

we have

$$\begin{pmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t \\ \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \text{vec}\left(Z_t u_t'\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left[O_{k(k+1)\times 1},\ \begin{pmatrix} \sigma^2 \cdot Q_{zz} & \sigma_{eu}' \bigotimes Q_{zz} \\ \sigma_{eu} \bigotimes Q_{zz} & \Sigma_u \bigotimes Q_{zz} \end{pmatrix}\right].$$

By Slutsky's theorem, it follows that

$$\begin{pmatrix} \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t \\ \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \text{vec}\left(Z_t X_t'\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left[\begin{pmatrix} 0 \\ \left(I_k \bigotimes Q_{zz}\right) \cdot \text{vec}\left(P'\right) \end{pmatrix},\quad \begin{pmatrix} \sigma^2 \cdot Q_{zz} & \sigma_{eu}' \bigotimes Q_{zz} \\ \sigma_{eu} \bigotimes Q_{zz} & \Sigma_u \bigotimes Q_{zz} \end{pmatrix}\right].$$

Let $\mathcal{Z}$ be a $k \times k$ random matrix with distribution

$$\text{vec}\left(\mathcal{Z}\right) \sim \mathcal{N}\left[\left(I_k \bigotimes Q_{zz}\right) \cdot \text{vec}\left(P'\right),\ \Sigma_u \bigotimes Q_{zz}\right].$$

Then, because the reverse vectorization procedure is continuous,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t X_t' \xrightarrow{d} \mathcal{Z}.$$

By the joint convergence in distribution result above, we can see that

$$\hat{\beta}_T^{IV} - \beta = \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t X_t'\right)^{-1} \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right) \xrightarrow{d} \mathcal{Z}^{-1} \times \mathcal{N}\left[0, \sigma^2 Q_{zz}\right].$$

Being the ratio of two normal distributions, with one not even centerd at zero, the asymptotic distribution of $\hat{\beta}_T^{IV} - \beta$ is not a known distribution. A more serious problem, however, is that $\hat{\beta}_T^{IV}$ is no longer consistent for β. Note that

$$\frac{1}{\sqrt{T}}\left(\hat{\beta}_T^{IV} - \beta\right) \xrightarrow{p} O_{k\times 1},$$

since

$$\frac{1}{T}\sum_{t=1}^{T} Z_t\epsilon_t \xrightarrow{p} O_{k\times 1}.$$

In other words, $\hat{\beta}_T^{IV}$ is not by istelf consistent for β, but must be scaled by a factor $\frac{1}{\sqrt{T}}$. Thus, the presence of weak instruments has the effect of "spreading out" the usual IV estimator of β; the newfound imprecision of the IV estimator means that it cannot be relied on for inference in the usual manner.

7.3.2 Inference under Weak Instruments

The question then becomes: how do we perform tests in the presence of weak instruments? One solution is the performing the Anderson-Rubin (AR) test. The AR test works because it does not require us to invert any estimators of Π, thus avoiding the problem that $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ is short ranked in finite samples when the instruments are weak. Another closely related method makes use of Fieller's method; it does not involve any form of matrix inversion at all.

To study these two methods of inference under weak instruments, we make assumptions similar to the ones made above but without the local-to-zero assumption. Instead, as in the previous chapter, we continue to assume that the probability limit of $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ has full rank, except this time with the implicit understanding that it is nearly short ranked.

Formally, the model assumptions are as follows:

A1 : **Control Function Specification**

Let the data be given as $\{Y_t, X_t, Z_t\}_{t\in\mathbb{Z}}$, where Y_t is a real random variable, X_t is a k-dimensional random vector, and Z_t is an L-dimensional random vector. Suppose the measurement equation and control function equation are given as

$$Y_t = X_t'\beta + \epsilon_t$$

$$X_t = \Pi Z_t + u_t,$$

where $\beta \in \mathbb{R}^{k\times 1}$ is some non-random vector and $\Pi \in \mathbb{R}^{k\times L}$ is a non-random matrix of rank k. This immediately requires that $k \leq L$.

A2 : **Exogeneity of Instruments**

Assume that $\{v_t = (\epsilon_t, u_t')'\}_{t\in\mathbb{Z}}$ is backward-looking exogenous with respect to the instruments Z_t, that is, $\{v_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to the filtration $\mathcal{F}$ generated by

$\{(Z_{t+1}, X_t, Y_t)\}_{t\in\mathbb{Z}}$. In particular,

$$\mathbb{E}\left[v_t \mid Z_t, Z_{t-1}, \cdots\right] = O_{(k+1)\times 1}$$

for any $t \in \mathbb{Z}$, which is a stronger requirement than assuming Z_t and v_t are simply uncorrelated.

Note further that, becuase v_t is a function of Y_t, X_t and Z_t, the filtration $\mathcal{G}$ generated by $\{(Z_{t+1}, v_t)\}_{t\in\mathbb{Z}}$ is coarser than $\mathcal{F}$, that is, $\mathcal{G}_t \subset \mathcal{F}_t$ for any $t \in \mathbb{Z}$. Therefore, $\{v_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}$, so that the assumptions of theorem 7.3 are satisfied.

A3 : **Homoskedasticity**

There exists a positive definite $(k+1)\times(k+1)$ matrix Σ such that

$$\mathbb{E}\left[v_t v_t' \mid \mathcal{F}_{t-1}\right] = \underbrace{\begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon u}' \\ \sigma_{\epsilon u} & \Sigma_u \end{pmatrix}}_{\Sigma}$$

for any $t \in \mathbb{Z}$.

A4 : **Regularity Assumptions**

We assume that there exists a positive definite matrix $Q_{zz} \in \mathbb{R}^{k\times k}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t' \xrightarrow{p} Q_{zz}.$$

Furthermore, we assume the following moment boundedness conditions hold:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|v_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}|Z_t|^4 < +\infty,$$

and there exists a $\delta > 0$ such that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\| v_t v_t' \bigotimes Z_t Z_t' \right\|^{1+\delta} < +\infty.$$

By the corollary to theorem 7.3, it follows immediately that

$$\frac{1}{T}\sum_{t=1}^{T} v_t v_t' \xrightarrow{p} \Sigma$$

$$\frac{1}{T}\sum_{t=1}^{T}\text{vec}\left(Z_t v_t'\right)\text{vec}\left(Z_t v_t'\right) \xrightarrow{p} \Sigma \bigotimes Q_{zz}$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\text{vec}\left(Z_t v_t'\right) \xrightarrow{d} \mathcal{N}\left[O_{L(k+1)\times 1},\ \Sigma \bigotimes Q_{zz}\right].$$

By implication,

$$\frac{1}{T}\sum_{t=1}^{T} Z_t v_t' \xrightarrow{p} O_{L\times(k+1)}.$$

From the control function equation

$$X_t = \Pi Z_t + u_t,$$

we can see that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t X_t' = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)\Pi' + \frac{1}{T}\sum_{t=1}^{t} Z_t u_t' \xrightarrow{p} Q_{zz}\Pi'.$$

Since Q_{zz} has full rank and Π has rank $k \leq L$, it follows that $Q_{zx} := Q_{zz}\Pi'$ also has rank k. Therefore, the instruments Z_t satsify the relevance condition in addition to the exogeneity condition.

However, we are implicitly assuming that Π is nearly short ranked, so that, under reasonable sample sizes T, the random matrix $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$ might have rank less than k for a subset of the sample space with positive measure. We want to make inferences about β even when this is the case; below we outline two methods of doing so.

The Anderson-Rubin Method

We first discuss the Anderson-Rubin (AR) test for hypotheses concerning β. Suppose we want to test the null hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$. The usual efficient GMM estimator for β might be undefined under reasonable sample sizes T or possess undesirable asymptotic properties, as we repeatedly noted above.

Anderson and Rubin suggest the following reformulation of the model. Consider the equations

$$Y_t = X_t'\beta + \epsilon_t$$

$$X_t = \Pi Z_t + u_t.$$

Subtracting $X_t'\beta_0$ from both sides of the first equation (the measurement equation for the de-

pendent variable) yields

$$Y_t - X_t'\beta_0 = X_t'(\beta - \beta_0) + \epsilon_t.$$

Now, substituting in the second equation (the first stage equation) for X_t in this equation yields

$$Y_t - X_t'\beta_0 = Z_t'\left[\Pi'(\beta - \beta_0)\right] + u_t'(\beta - \beta_0) + \epsilon_t.$$

Defining $\gamma = \Pi'(\beta - \beta_0) \in \mathbb{R}^{L\times 1}$, suppose that $\gamma = O_{L\times 1}$. Then, because Π' is an $L \times k$ matrix with rank k, by the dimension theorem it has nullity 0, so that $\beta = \beta_0$. On the other hand, if $\beta = \beta_0$, then it is clear that $\gamma = O_{L\times 1}$. In other words, instead of testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, we could test $H_0 : \gamma = O_{L\times 1}$ against $H_1 : \gamma \neq O_{L\times 1}$.

Because the OLS estimator of γ can be obtained by regressing $Y_t^* = Y_t - X_t'\beta_0$ on Z_t, where Z_t is exogenous for the error term $u_t'(\beta - \beta_0) + \epsilon_t$, it follows that this estimator is consistent and asymptotically normal. Thus, we will be able to conduct inference on γ without ever having to be concerned about the rank of $\frac{1}{T}\sum_{t=1}^T Z_t X_t'$.

We can formalize the intuition of the AR test as follows:

Theorem 7.4 (Asymptotic Properties of AR Test Statistic)
Suppose assumptions A1 to A4 hold. Consider testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$. The AR test statistic is defined as

$$AR_T = \frac{\hat{\gamma}_T'\left(\sum_{t=1}^T Z_t Z_t'\right)\hat{\gamma}_T}{\hat{\sigma}^2_{\epsilon,T}},$$

where

$$\hat{\gamma}_T = \left(\frac{1}{T}\sum_{t=1}^T Z_t Z_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T Z_t(Y_t - X_t'\beta_0)\right)$$

and

$$\hat{\sigma}^2_{\epsilon,T} = \frac{1}{T}\sum_{t=1}^T (Y_t - X_t'\beta_0 - Z_t'\hat{\gamma}_T)^2.$$

Then, under the null, we have

$$AR_T \xrightarrow{d} \chi^2_L,$$

and under any alternative $H_1 : \beta = \beta_1 \neq \beta_0$,

$$AR_T \xrightarrow{p} +\infty.$$

Proof) We proceed in two steps.

Asymptotic Distribution of AR Test Statistic under H_0

Under H_0, $\beta = \beta_0$ and $\gamma = 0$, so that

$$Y_t - X_t'\beta_0 = \epsilon_t$$

and

$$\hat{\gamma}_T = \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_t \epsilon_t\right).$$

Note that, by the corollary to theorem 7.3,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t \xrightarrow{d} \mathcal{N}\left[O_{L\times 1},\ \sigma_\epsilon^2 \cdot Q_{zz}\right].$$

It follows from Slutsky's theorem that

$$\begin{aligned}\sqrt{T}\cdot\hat{\gamma}_T &= \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right) \\ &\xrightarrow{d} \mathcal{N}\left[O_{L\times 1},\ \sigma_\epsilon^2 \cdot Q_{zz}^{-1}\right].\end{aligned}$$

Again from theorem 7.3, we know that

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 \xrightarrow{p} \sigma_\epsilon^2.$$

Consider the estimator

$$\hat{\sigma}_{\epsilon,T}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - X_t'\beta_0 - Z_t'\hat{\gamma}_T\right)^2$$

of σ_ϵ^2. Using the fact that

$$Y_t - X_t'\beta = \epsilon_t$$

yields the following expansion of $\hat{\sigma}_{\epsilon,T}^2$:

$$\begin{aligned}\hat{\sigma}_{\epsilon,T}^2 &= \frac{1}{T}\sum_{t=1}^{T}(\epsilon_t - Z_t'\hat{\gamma}_T)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\epsilon_t^2 - 2\hat{\gamma}_T'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t\epsilon_t\right) + \hat{\gamma}_T'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)\hat{\gamma}_T.\end{aligned}$$

Since $\hat{\gamma}_T \xrightarrow{p} 0$ follows from the asymptotic distribution of $\hat{\gamma}_T$, we can easily see from what we have established so far that

$$\hat{\sigma}^2_{\epsilon,T} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 + o_p(1) \xrightarrow{p} \sigma^2_\epsilon.$$

Let $\mathcal{Z}$ be an L-dimensional random vector with distribution

$$\mathcal{Z} \sim \mathcal{N}\left[O_{L\times 1}, \sigma^2_\epsilon \cdot Q_{zz}\right].$$

Then, it follows from Slutsky's theorem that

$$\begin{aligned} AR_T &= \frac{\hat{\gamma}_T' \left(\sum_{t=1}^{T} Z_t Z_t'\right) \hat{\gamma}_T}{\hat{\sigma}^2_{\epsilon,T}} \\ &= \left(\sqrt{T}\cdot\hat{\gamma}_T\right)' \left[\hat{\sigma}^2_{\epsilon,T} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1}\right]^{-1} \left(\sqrt{T}\cdot\hat{\gamma}_T\right) \\ &\xrightarrow{d} \mathcal{Z}' \left(\sigma^2_\epsilon \cdot Q_{zz}^{-1}\right)^{-1} \mathcal{Z} \sim \chi^2_L. \end{aligned}$$

Consistency of the AR Test

Consider the alternative $H_1 : \beta = \beta_1 \neq \beta_0$. Then, because

$$Y_t - X_t'\beta_0 = X_t'(\beta_1 - \beta_0) + \epsilon_t = Z_t' \left(\Pi'(\beta_1 - \beta_0)\right) + u_t' \left(\beta_1 - \beta_0\right) + \epsilon_t,$$

we have

$$\begin{aligned} \hat{\gamma}_T &= \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t \left[Z_t' \left(\Pi'(\beta_1 - \beta_0)\right) + u_t' \left(\beta_1 - \beta_0\right) + \epsilon_t\right]\right) \\ &= \Pi' \left(\beta_1 - \beta_0\right) + \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t \left[u_t' \left(\beta_1 - \beta_0\right) + \epsilon_t\right]\right). \end{aligned}$$

By theorem 7.3,

$$\frac{1}{T}\sum_{t=1}^{T} Z_t v_t' \xrightarrow{p} O_{L\times(k+1)},$$

so it follows that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t \left[u_t'(\beta_1 - \beta_0) + \epsilon_t\right] \xrightarrow{p} O_{L\times 1}.$$

Therefore,

$$\hat{\gamma}_T \xrightarrow{p} \Pi'(\beta_1 - \beta_0),$$

where $\Pi'(\beta_1 - \beta_0) \neq O_{L\times 1}$ because otherwise $\beta_1 = \beta_0$, a contradiction. Define $\gamma_1 := \Pi'(\beta_1 - \beta_0)$.

Now note that

$$\begin{aligned}
\hat{\sigma}^2_{\epsilon,T} &= \frac{1}{T}\sum_{t=1}^{T}(Y_t - X_t'\beta_0 - Z_t'\hat{\gamma}_T)^2 \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[Z_t'(\gamma_1 - \hat{\gamma}_T) + u_t'(\beta_1 - \beta_0) + \epsilon_t\right]^2 \\
&= (\gamma_1 - \hat{\gamma}_T)'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t Z_t'\right)(\gamma_1 - \hat{\gamma}_T) \\
&\quad + 2\cdot(\gamma_1 - \hat{\gamma}_T)'\left(\frac{1}{T}\sum_{t=1}^{T} Z_t\left[u_t'(\beta_1 - \beta_0) + \epsilon_t\right]\right) \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\left(u_t'(\beta_1 - \beta_0) + \epsilon_t\right)^2.
\end{aligned}$$

By theorem 7.3,

$$\frac{1}{T}\sum_{t=1}^{T}\left(u_t'(\beta_1 - \beta_0) + \epsilon_t\right)^2 = \begin{pmatrix}1 & (\beta_1 - \beta_0)'\end{pmatrix}\left(\frac{1}{T}\sum_{t=1}^{T} v_t v_t'\right)\begin{pmatrix}1 \\ \beta_1 - \beta_0\end{pmatrix} \xrightarrow{p} \sigma_w^2,$$

where we define

$$\begin{aligned}
\sigma_w^2 &:= \begin{pmatrix}1 & (\beta_1 - \beta_0)'\end{pmatrix}\begin{pmatrix}\sigma_\epsilon^2 & \sigma_{\epsilon u}' \\ \sigma_{\epsilon u} & \Sigma_u\end{pmatrix}\begin{pmatrix}1 \\ \beta_1 - \beta_0\end{pmatrix} \\
&= \sigma_\epsilon^2 + 2(\beta_1 - \beta_0)'\sigma_{\epsilon u}\sigma_\epsilon^2 + (\beta_1 - \beta_0)'\Sigma_u(\beta_1 - \beta_0) > 0,
\end{aligned}$$

where the final inequality follows from the fact that Σ is positive definite. The other terms comprising $\hat{\sigma}^2_{\epsilon,T}$ are all $o_p(1)$ terms, so it follows that

$$\hat{\sigma}^2_{\epsilon,T} = \frac{1}{T}\sum_{t=1}^{T}\left(u_t'(\beta_1 - \beta_0) + \epsilon_t\right)^2 + o_p(1) \xrightarrow{p} \sigma_w^2.$$

Putting all these results together, we can see that

$$\frac{1}{T}AR_T = \frac{\hat{\gamma}_T' \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t' \right) \hat{\gamma}_T}{\hat{\sigma}_{\epsilon,T}^2} \xrightarrow{p} \frac{\gamma_1' Q_{zz} \gamma_1}{\sigma_w^2}.$$

Here, γ_1 is non-zero and Q_{zz} is positive definite, so the probability limit is positive. It follows that

$$AR_T = T \left(\frac{1}{T} AR_T \right) \xrightarrow{p} +\infty,$$

as we desired.

Q.E.D.

The preceding theorem shows us that the AR test statistic, constructed as the usual Wald test statistic for testing $H_0 : \gamma = O_{L \times 1}$, has the usual chi squared distribution under the null $H_0 : \beta = \beta_0$, and is consistent for any alternative $H_1 : \beta = \beta_1 \neq \beta_0$. Thus, it furnishes a viable way to test for hypotheses concerning β in the presence of weak instruments; note that we have not once used the expression $\frac{1}{T} \sum_{t=1}^{T} Z_t X_t'$ in the definition of the AR test statistic.

Inference via Fieller's Method

Another way to test the null hypothesis $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$ is to consider the reduced-form counterpart to the structural model

$$Y_t = X_t'\beta + \epsilon_t$$

$$X_t = \Pi Z_t + u_t.$$

This model is called the structural model because it describes the actual process by which Y_t and X_t are simultaneously determined, where the regressor X_t might be correlated with the error ϵ_t. The reduced form counterpart is a version of the above system that describes the same DGP, but which has only exogenous regressors on the right hand side. In this case, the reduced form model is given by simply substituting the second equation into the first one:

$$Y_t = Z_t'\gamma + u_t'\beta + \epsilon_t$$

$$X_t = \Pi Z_t + u_t,$$

where we define $\gamma = \Pi'\beta$. The reduced form model suggests the following way of testing for $H_0 : \beta = \beta_0$. We can recover β as a function of γ and Π as

$$\beta = (\Pi\Pi')^{-1}\Pi \cdot \gamma.$$

Here, $\Pi\Pi'$ is a nonsingular $k \times k$ matrix because Π is an $L \times k$ matrix of rank k. The problem with weak instruments is that, although $(\Pi\Pi')^{-1}$ might exist, $\Pi\Pi'$ might be too short ranked that taking the sample analogue of this expression as the estimator of β is not feasible under reasonable sample sizes.

To circumvent this problem, note that $\beta = \beta_0$ if and only if $(\Pi\Pi')^{-1}\Pi \cdot \gamma = \beta_0$, or equivalently, if and only if

$$\alpha := \Pi \cdot \gamma - (\Pi\Pi')\beta_0 = O_{k\times 1}.$$

As with the AR test, we can therefore reduce the problem of testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ to the problem of testing $H_0 : \alpha = O_{k\times 1}$ against $H_1 : \alpha \neq O_{k\times 1}$. This does not involve any complications from weak instruments because it does not require the inversion of any matrix involving $\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'$.

This approach mirrors Fieller's method for testing for the ratio of two means. Letting μ_X and μ_Y be the population means of two populations X and Y, Fieller's method tests the hypothesis $H_0 : \frac{\mu_Y}{\mu_X} = \theta$ against $H_1 : \frac{\mu_Y}{\mu_X} \neq \theta$ by instead testing $H_0 : \mu_Y - \theta\mu_X = 0$ against $H_1 : \mu_Y - \theta\mu_X \neq 0$. This is robust against cases where the mean μ_X is so close to 0 that using the sample analogue of $\frac{\mu_Y}{\mu_X}$ to construct the test statistic does not work well under reasonable sample sizes. Observe how this problem is the same as that faced in a model with weak instruments, where the covariance between Z_t and X_t serves the same role as μ_X.

We need an estimator of α to derive the desired test statistic. This is accomplished by taking the OLS estimators of γ and Π and constructing an estimator of α as a function of these OLS estimators. While we have not, strictly speaking, defined the OLS estimator of the coefficient matrix of a system of equations, we define it in a straightforward manner here and delve into the details in subsequent sections. In particular, we will see that the OLS estimator defined here is precisely the feasible efficient system GMM estimator in a system of equations where each equaition shares the same exogenous regressors. Formally, our estimators of α can be given as

$$\hat{\alpha}_T = \hat{\Pi}_T \cdot \hat{\gamma}_T - \left(\hat{\Pi}_T \hat{\Pi}_T'\right) \beta_0,$$

where

$$\begin{pmatrix} \hat{\gamma}_T & \hat{\Pi}_T' \end{pmatrix} = \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Z_t' \right)^{-1} \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} Z_t Y_t & \frac{1}{T}\sum_{t=1}^{T} Z_t X_t' \end{pmatrix}$$

is the OLS estimator of the $L \times (k+1)$ coefficient matrix (γ, Π').

We will see later on that the estimator $\begin{pmatrix} \hat{\gamma}_T \\ \text{vec}\left(\hat{\Pi}_T'\right) \end{pmatrix}$ is consistent and asymptotically normal, using which we can construct a test statistic using $\hat{\alpha}_T$, where we find the critical values using the relevant asymptotic distribution.

7.4 Efficient System GMM

Here we study efficient GMM estimation of a model with multiple equations. This general framework encompasses a variety of different models, including the seemingly unrelated regression equations (SURE) model, and can be used to estimate structural models such as the simultaneous equations model, which we will study below.

7.4.1 Model Assumptions

The assumptions of the model are as follows:

A1 : **The Sample**

There is a sample $\{(Y_{it}, X_{it}, Z_{it})\}_{t\in\mathbb{Z}, 1\leq i\leq n}$, where each Y_{it} is a real-valued dependent variable, X_{it} is a k_i-dimensional vector of regressors, and Z_{it} is an L_i-dimensional vector of instruments. The sample observations are assumed to be dependent, with the degree of dependence to be specified below. The more general assumption concerning the dependence structure of the model allows lagged values of the dependent variable to be included in the instruments Z_{it}.

The error of the ith equation, ϵ_{it}, satisfies the equation

$$Y_{it} = X_{it}'\beta_i + \epsilon_{it}$$

for any $t \in \mathbb{Z}$, where β_i is a non-random k_i-dimensional real vector.

Defining

$$Y_t = \begin{pmatrix} Y_{1t} \\ \vdots \\ Y_{nt} \end{pmatrix}, \quad X_t = \begin{pmatrix} X_{1t} & \cdots & O_{k_1\times 1} \\ \vdots & \ddots & \vdots \\ O_{k_n\times 1} & \cdots & X_{nt} \end{pmatrix}, \quad Z_t = \begin{pmatrix} Z_{1t} & \cdots & O_{L_1\times 1} \\ \vdots & \ddots & \vdots \\ O_{L_n\times 1} & \cdots & Z_{nt} \end{pmatrix}, \quad \epsilon_t = \begin{pmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{nt} \end{pmatrix}$$

for any $t \in \mathbb{Z}$ and letting $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$, the n equations can be stacked as

$$Y_t = X_t'\beta + \epsilon_t.$$

Letting $k = \sum_{i=1}^n k_i$ and $L = \sum_{i=1}^n L_i$, note that X_t and Z_t are $k \times n$ and $L \times n$ random matrices, respectively.

A2 : **Exogeneity of Instruments**

Let $\mathcal{F}$ be the filtration generated by $\{(Z_{t+1}, X_t, Y_t)\}_{t\in\mathbb{Z}}$. We assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to the filtration $\mathcal{F}$. Heuristically, this implies that, at each period $t \in \mathbb{Z}$, the error ϵ_t is a proper innovation, that is, it is mean independent with everything that came before it. The instruments and errors are also contemporaneously uncorrelated, so that they are properly exogenous. We impose the stronger requirement that the errors are not only uncorrelated with the current instruments but also uncorrelated with every past variable because the sample is no longer assumed to be independent.

In particular, because ϵ_t is $\sigma(Y_t, X_t)$-measurable for any $t \in \mathbb{Z}$, $\mathcal{F}$ is finer than the filtration $\mathcal{G}$ generated by $\{(X_{t+1}, \epsilon_t)\}_{t\in\mathbb{Z}}$. It follows that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}$, and as such the convergence results derived in the previous section are valid. Note that, formally, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is pairwise uncorrelated but not independent, and crucially,

$$\mathbb{E}\left[\epsilon_t \mid Z_t, Z_{t-1}, \cdots\right] = O_{n\times 1}$$

for any $t \in \mathbb{Z}$.

A3 : **Relevance of Instruments**

Note that, under our new notation,

$$Z_t X_t' = \begin{pmatrix} Z_{1t}X_{1t}' & \cdots & O_{L_1\times k_n} \\ \vdots & \ddots & \vdots \\ O_{L_n\times k_1} & \cdots & Z_{nt}X_{nt}' \end{pmatrix}$$

is a $L \times k$ valued random matrix. We assume that there exists an $L \times k$ matrix Q_{zx} with rank k such that

$$\frac{1}{T}\sum_{t=1}^{T} Z_t X_t' \xrightarrow{p} Q_{zx}.$$

A4 : **Homoskedasticity**

There exists an $n \times n$ matrix Σ such that

$$\mathbb{E}\left[\epsilon_t \epsilon_t' \mid \mathcal{F}_{t-1}\right] = \Sigma$$

for any $t \in \mathbb{Z}$. Denote the (i, j)th element of Σ by σ_{ij}.

A5 : **No Perfect Collinearity**

For any $1 \leq i, j \leq n$, there exists a $L_i \times L_j$ matrix $Q_{zz,ij}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} Z_{it} Z_{jt}' \xrightarrow{p} Q_{zz,ij}.$$

We assume $Q_{zz,ii}$ is positive definite for each $1 \leq i \leq n$. Defining

$$\Sigma_{zz} = \begin{pmatrix} \sigma_{11} Q_{zz,11} & \cdots & \sigma_{1n} Q_{zz,1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} Q_{zz,n1} & \vdots & \sigma_{nn} Q_{zz,nn} \end{pmatrix},$$

we assume that Σ_{zz} is a positive definite $L \times L$ matrix.

A6 : **Boundedness of Moments**

$\{\epsilon_t\}_{t\in\mathbb{Z}}$ and $\{Z_t\}_{t\in\mathbb{Z}}$ have finite fourth moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}\|Z_t\|^4 < +\infty.$$

$\{X_t\}_{t\in\mathbb{Z}}$ has finite second moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\|X_t\|^2 < +\infty.$$

In addition, there exists some $\delta > 0$ such that, for any $1 \leq i, j \leq n$,

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|\epsilon_{it}\epsilon_{jt} Z_{it} Z_{jt}'\right\|^{1+\delta} < +\infty.$$

Under these assumptions, by theorem 7.3, we can immediately tell that

$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t \epsilon_t' \xrightarrow{p} \Sigma$$

$$\frac{1}{T}\sum_{t=1}^{T} (Z_t\epsilon_t)(Z_t\epsilon_t)' \xrightarrow{p} \Sigma_{zz}$$

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\epsilon_t \xrightarrow{d} \mathcal{N}\left[O_{L\times 1}, \Sigma_{zz}\right].$$

7.4.2 Moment Conditions and Identification

Under the exogeneity assumption for the instruments, we have L moment conditions

$$\mathbb{E}[Z_t\epsilon_t] = \mathbb{E}\left[Z_t(Y_t - X_t'\beta)\right] = O_{L\times 1}.$$

The sample analogue of these conditions is

$$\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) = O_{L\times 1},$$

and letting β_0 be the true value of β, the left hand side has probability limit

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\beta) &= \frac{1}{T}\sum_{t=1}^{T} Z_t(X_t'(\beta_0 - \beta) + \epsilon_t) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)(\beta_0 - \beta) + \frac{1}{T}\sum_{t=1}^{T} Z_t\epsilon_t \xrightarrow{p} Q_{zx}(\beta_0 - \beta).\end{aligned}$$

Here, because Q_{zx} has full rank k, the population moment conditions are exactly satisfied if and only if $\beta = \beta_0$, meaning that the parameters in the model are identified. Once again, the relevance condition for the instruments is sufficient for the identification of the parameter of interest. Here, we can break down the relevance condition into two parts:

1. **The Order Condition**
   This condition requires the total number of instruments L must be larger than or equal to the total number of regressors k. This is a necessary condition of identification, since the rank of an $m \times n$ matrix A is bounded above by $\min(m,n)$, and Q_{zx} is an $L \times k$ matrix.

2. **The Rank Condition**
   This condition requires the rank of Q_{xz} to be equal to k. We can see that this is a sufficient condition for identification, and entails the stronger order condition, due to the block diagonal form of Q_{zx}, that L_i be larger than or equal to k_i for each $1 \leq i \leq n$.

We will encounter these two conditions again in a more familiar form when we study simultaneous equation models.

7.4.3 The Efficient System GMM Estimator

We can now obtain an efficient GMM estimator for β in the same manner as we obtained one in the univariate case. We go through all the steps again for the sake of completeness.

Under the sequence of positive definite weighting matrices $\{W_T\}_{T\in N_+}$ that converges in probability to some positive definite $L \times L$ matrix W, the GMM objective function becomes

$$\hat{Q}_T(b) = \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'b)\right]' W_T \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'b)\right]$$

for any $b \in \mathbb{R}^k$. The GMM estimator of β, which we denote $\hat{\beta}_T$, satisfies the first order conditions

$$\boldsymbol{\nabla}\hat{Q}_T(\hat{\beta}_T) = -2 \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left[\frac{1}{T}\sum_{t=1}^{T} Z_t(Y_t - X_t'\hat{\beta}_T)\right] = O_{k\times 1},$$

solving which yields the GMM estimator

$$\begin{aligned}\hat{\beta}_T &= \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t Y_t\right) \\ &= \beta + \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t \epsilon_t\right).\end{aligned}$$

Under our assumptions, we can easily obtain the asymptotic distribution of $\hat{\beta}_T$:

Theorem 7.5 (Asymptotic Behavior of System GMM Estimator)
Suppose assumptions A1 to A6 hold, and let $\hat{\beta}_T$ be the system GMM estimator of β under the positive definite weighting matrices $\{W_T\}_{T\in N_+}$, which converges in probability to some positive definite matrix $W \in \mathbb{R}^{L\times L}$. Then,

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\, V\right],$$

where

$$V = (Q_{zx}' W Q_{zx})^{-1} Q_{zx}' W \Sigma_{zz} W Q_{zx} (Q_{zx}' W Q_{zx})^{-1}.$$

Proof) We noted that, under our assumptions, theorem 7.3 implies

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t \xrightarrow{d} \mathcal{N}\left[O_{L\times 1},\, \Sigma_{zz}\right].$$

Since

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] = \left[\left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} Z_t X_t'\right)' W_T \left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \epsilon_t\right),$$

by Slutsky's theorem

$$\begin{aligned}\sqrt{T}\left[\hat{\beta}_T - \beta\right] &\xrightarrow{d} (Q_{zx}' W Q_{zx})^{-1} Q_{zx}' W \cdot \mathcal{N}\left[O_{L\times 1},\, \Sigma_{zz}\right] \\ &= \mathcal{N}\left[O_{k\times 1},\, (Q_{zx}' W Q_{zx})^{-1} Q_{zx}' W \Sigma_{zz} W Q_{zx} (Q_{zx}' W Q_{zx})^{-1}\right],\end{aligned}$$

as desired.

Q.E.D.

Note that the above result implies the consistency of $\hat{\beta}_T$ for β as well. To obtain the efficient system GMM estimator of β, we once again rely on the following result:

Lemma 7.6 (Weights for Efficient System GMM Estimation)
Suppose assumptions A1 to A6 hold. Then, for any positive definite $L \times L$ matrix W,

$$(Q_{zx}' W Q_{zx})^{-1} Q_{zx}' W \Sigma_{zz} W Q_{zx} (Q_{zx}' W Q_{zx})^{-1} - (Q_{zx}' \Sigma_{zz}^{-1} Q_{zx})^{-1}$$

is positive semidefinite.

Proof) Once again, it is sufficient to show that

$$\Delta = Q_{zx}' \Sigma_{zz}^{-1} Q_{zx} - Q_{zx}' W Q_{zx} \left(Q_{zx}' W \Sigma_{zz} W Q_{zx}\right)^{-1} Q_{zx}' W Q_{zx}$$

is positive semidefinite. Let $\Sigma_{zz} = BB'$ be the Cholesky decomposition of Σ_{zz}; it follows that $\Sigma_{zz} = B^{-1\prime} B^{-1}$. Now we have

$$\begin{aligned}\Delta &= Q_{zx}' B^{-1\prime} \left[I_L - B' W Q_{zx} \left(Q_{zx}' W B B' W Q_{zx}\right)^{-1} Q_{zx}' W B\right] B^{-1} Q_{zx} \\ &= Q_{zx}' B^{-1\prime} \cdot M_{B' W Q_{zx}} \cdot B^{-1} Q_{zx},\end{aligned}$$

where $M_{B' W Q_{zx}}$ is a symmetric and idempotent matrix. Therefore, Δ is positive semidefinite, and we have our desired result.

Q.E.D.

It follows that the efficient system GMM estimator of β is given as

$$\hat{\beta}_T^E = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \Sigma_{zz}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \Sigma_{zz}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right),$$

which has asymptotic distribution

$$\sqrt{T} \left[\hat{\beta}_T^E - \beta \right] \xrightarrow{d} \mathcal{N} \left[O_{k \times 1}, \, (Q_{zx}' \Sigma_{zz}^{-1} Q_{zx})^{-1} \right].$$

7.4.4 Feasible Efficient System GMM Estimation

Once again we run into the problem that Σ_{zz} is an unobservable quantity, and thus has to be estimated consistently. To this end, as usual, we first obtain a consistent but not necessarily efficient estimator of β and use it to construct a consistent estimator of Σ_{zz}, which we then use to construct the feasible efficient system GMM estimator of β.

Equation-by-Equation 2SLS Estimation

An easy candidate for a consistent estimator of β that does not rely on Σ_{zz} is once again the 2SLS estimator. Specifically, consider the 2SLS estimator of β_i from the univariate regression

$$Y_{it} = X_{it}' \beta_i + \epsilon_{it}.$$

The 2SLS estimator of β_i using Z_{it} as instruments is given as

$$\begin{aligned} \hat{\beta}_{i,T}^{2SLS} = &\left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} X_{it}' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} Z_{it}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} X_{it}' \right) \right]^{-1} \\ &\times \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} X_{it}' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} Z_{it}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} Y_{it} \right), \end{aligned}$$

where the estimator is well-defined because the probability limit of $\frac{1}{T} \sum_{t=1}^{T} Z_{it} X_{it}'$ has rank $k_i \leq L_i$ by the relevance condition and the probability limit $Q_{zz,ii}$ of $\frac{1}{T} \sum_{t=1}^{T} Z_{it} Z_{it}'$ is positive definite by assumption. It then follows easily from the continuous mapping theorem and the fact that

$$\frac{1}{T} \sum_{t=1}^{T} Z_t \epsilon_t \xrightarrow{p} O_{L \times 1}$$

that $\hat{\beta}_{i,T}^{2SLS}$ is consistent for β_i. In fact, since the errors are homoskedastic, the 2SLS estimator of β_i is the efficient GMM estimator of β_i when we restrict our attention to the ith equation.

Putting all these estimators together, we obtain the equation-by-equation 2SLS estimator

$$\hat{\beta}_T^{2SLS} = \begin{pmatrix} \hat{\beta}_{1,T}^{2SLS} \\ \vdots \\ \hat{\beta}_{n,T}^{2SLS} \end{pmatrix},$$

which is consistent for β.

Consistent Esimation of Σ_{zz}

Using the 2SLS estimator, we can estimate the error variance Σ via the estimator

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}(Y_t - X_t'\hat{\beta}_T^{2SLS})(Y_t - X_t'\hat{\beta}_T^{2SLS})'.$$

The consistency of $\hat{\Sigma}_T$ for Σ is the content of the next theorem:

Lemma 7.7 (Consistent Error Variance Estimation)
Suppose assumptions A1 to A6 hold. Let $\hat{\beta}_T^{2SLS}$ be the equation-by-equation 2SLS estimator of β, and $\hat{\Sigma}_T$ the estimator of Σ obtained using $\hat{\beta}_T^{2SLS}$. Then,

$$\hat{\Sigma}_T \xrightarrow{p} \Sigma.$$

Proof) Note that we can expand $\hat{\Sigma}_T$ as

$$\begin{aligned}
\hat{\Sigma}_T &= \frac{1}{T}\sum_{t=1}^{T}\left[X_t'(\beta - \hat{\beta}_T^{2SLS}) + \epsilon_t\right]\left[X_t'(\beta - \hat{\beta}_T^{2SLS}) + \epsilon_t\right]' \\
&= (\beta - \hat{\beta}_T^{2SLS})'\left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t'\right)(\beta - \hat{\beta}_T^{2SLS}) \\
&\quad + 2(\beta - \hat{\beta}_T^{2SLS})'\left(\frac{1}{T}\sum_{t=1}^{T}X_t\epsilon_t\right) \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\epsilon_t\epsilon_t'.
\end{aligned}$$

Note that,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}X_t\epsilon_t\right| &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}|X_t\epsilon_t| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}\|X_t\|^2\right)^{\frac{1}{2}}\left(\mathbb{E}|\epsilon_t|^2\right)^{\frac{1}{2}}
\end{aligned}$$

$$\leq \left(\sup_{t \in \mathbb{Z}} \mathbb{E}\|X_t\|^2 \right)^{\frac{1}{2}} (\mathrm{tr}(\Sigma))^{\frac{1}{2}} < +\infty$$

by the finiteness of the second moments of X_t, so that $\frac{1}{T}\sum_{t=1}^{T} X_t \epsilon_t = O_p(1)$ as $T \to \infty$. It follows that

$$(\beta - \hat{\beta}_T^{2SLS})' \left(\frac{1}{T} \sum_{t=1}^{T} X_t \epsilon_t \right) = o_p(1) \quad \text{as } T \to \infty,$$

since $\beta - \hat{\beta}_T^{2SLS} = o_p(1)$.

Similarly, since

$$\begin{aligned} \mathbb{E}\left\| \frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right\| &\leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|X_t X_t'\| \\ &\leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|X_t\|^2 \\ &\leq \sup_{t \in \mathbb{Z}} \mathbb{E}\|X_t\|^2 < +\infty \end{aligned}$$

for any $T \in N_+$, we can see that $\frac{1}{T}\sum_{t=1}^{T} X_t X_t' = O_p(1)$ as $T \to \infty$, and so

$$(\beta - \hat{\beta}_T^{2SLS})' \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right) (\beta - \hat{\beta}_T^{2SLS}) = o_p(1)$$

as $T \to \infty$. Putting all these results together, we have

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} \epsilon_t \epsilon_t' + o_p(1) \xrightarrow{p} \Sigma.$$

Q.E.D.

To the Feasible Efficient Estimator

Therefore, a consistent estimator of Σ_{zz} is given as

$$\hat{\Sigma}_{zz,T} = \begin{pmatrix} \hat{\sigma}_{11,T} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_{1t} Z_{1t}' \right) & \cdots & \hat{\sigma}_{1n,T} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_{1t} Z_{nt}' \right) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{n1,T} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_{nt} Z_{1t}' \right) & \cdots & \hat{\sigma}_{nn,T} \cdot \left(\frac{1}{T}\sum_{t=1}^{T} Z_{nt} Z_{nt}' \right) \end{pmatrix},$$

where $\hat{\sigma}_{ij,T}$ is the (i,j)th element of the matrix $\hat{\Sigma}_T$. Our feasible efficient system GMM estimator of β is now given as

$$\hat{\beta}_T^{FE} = \left[\left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_{zz,T}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right) \right]^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t X_t' \right)' \hat{\Sigma}_{zz,T}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} Z_t Y_t \right),$$

and it has the same asymptotic distribution as the efficient system GMM estimator $\hat{\beta}_T^E$.

7.4.5 Efficiency Gains of System GMM

We have seen that the feasible efficient system GMM estimator $\hat{\beta}_T^{FE}$ is the most asymptotically efficient GMM estimator of β. The natural question that arises is: when do we see large efficiency gains from using $\hat{\beta}_T^{FE}$ over the equation-by-equation 2SLS estimator $\hat{\beta}_T^{2SLS}$? The answer can be found by focusing on the main point of deviation of $\hat{\beta}_T^{FE}$ from $\hat{\beta}_T^{2SLS}$, the weighting matrix

$$\hat{\Sigma}_{zz,T} = \begin{pmatrix} \hat{\sigma}_{11,T} \cdot \left(\frac{1}{T} \sum_{t=1}^{T} Z_{1t} Z_{1t}' \right) & \cdots & \hat{\sigma}_{1n,T} \cdot \left(\frac{1}{T} \sum_{t=1}^{T} Z_{1t} Z_{nt}' \right) \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{n1,T} \cdot \left(\frac{1}{T} \sum_{t=1}^{T} Z_{nt} Z_{1t}' \right) & \cdots & \hat{\sigma}_{nn,T} \cdot \left(\frac{1}{T} \sum_{t=1}^{T} Z_{nt} Z_{nt}' \right) \end{pmatrix}.$$

It is easy to see that $\hat{\beta}_T^{2SLS} = \hat{\beta}_T^{FE}$ if and only if

$$\hat{\sigma}_{ij,T} \cdot \left(\frac{1}{T} \sum_{t=1}^{T} Z_{it} Z_{jt}' \right) = O_{L_i \times L_j}$$

for any $1 \leq i \neq j \leq n$, that is, when the off-diagonal terms of $\hat{\Sigma}_{zz,T}$ are equal to 0. The off-diagonal terms of this estimator is likely to be close to 0 if the off-diagonal terms of the limiting matrix

$$\Sigma_{zz} = \begin{pmatrix} \sigma_{11} Q_{zz,11} & \cdots & \sigma_{1n} Q_{zz,1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} Q_{zz,n1} & \vdots & \sigma_{nn} Q_{zz,nn} \end{pmatrix}$$

are close or equal to 0. The off-diagonal terms here represent two sources of correlation: the first between the error terms $\epsilon_{1t}, \cdots, \epsilon_{nt}$, and the second between the instruments $Z_{1t}, \cdots, Z_{nt}$. If we use the same instruments for every equation, as is the case in many models, then the off-diagonal terms end up capturing only the correlation between the errors in the vector ϵ_t. As such, we can expect to see efficiency gains by opting for system GMM over equation-by-equation GMM if the same confounders affect each regression equation, causing the errors to be highly correlated with one another.

7.5 The Seemingly Unrelated Regression Equations (SURE) Model

The seemingly unrelated regression equations (SURE) model is a special case of the system GMM method that we encounter often in the literature. Many econometric models, such as reduced-form VARs, are, in turn, special cases of the SURE model. The main difference between the SURE model and the general system GMM model is that the regressors are assumed to be exogenous, so that the regressors serve as their own instruments. This leads to a very simple form for the feasible efficient GMM estimator, and, in some cases, even reduces to equation-by-equation OLS even though the errors remain correlated with one another.

7.5.1 Model Assumptions

The assumptions of the model are as follows:

A1 : **The Sample**

There is a sample $\{(Y_{it}, X_{it})\}_{t\in\mathbb{Z}, 1\leq i\leq n}$, where each Y_{it} is a real-valued dependent variable, and X_{it} is a k_i-dimensional vector of regressors. The sample observations are assumed to be dependent, with the degree of dependence to be specified below. The more general assumption concerning the dependence structure of the model allows lagged values of the dependent variable to be included in the regressors X_{it}.

The error of the ith equation, ϵ_{it}, satisfies the equation

$$Y_{it} = X_{it}'\beta_i + \epsilon_{it}$$

for any $t \in \mathbb{Z}$, where β_i is a non-random k_i-dimensional real vector.

Defining

$$Y_t = \begin{pmatrix} Y_{1t} \\ \vdots \\ Y_{nt} \end{pmatrix}, \quad \mathbb{X}_t = \begin{pmatrix} X_{1t} & \cdots & O_{k_1\times 1} \\ \vdots & \ddots & \vdots \\ O_{k_n\times 1} & \cdots & X_{nt} \end{pmatrix}, \quad \epsilon_t = \begin{pmatrix} \epsilon_{1t} \\ \vdots \\ \epsilon_{nt} \end{pmatrix}$$

for any $t \in \mathbb{Z}$ and letting $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$, the n equations can be stacked as

$$Y_t = \mathbb{X}_t'\beta + \epsilon_t.$$

Denote the dimension of β by $K = \sum_{i=1}^n k_i$. We collect the distinct elements of $X_{1t}, \cdots, X_{nt}$ into the L-dimensional vector X_t.

A2 : **Exogeneity of Regressors**

Let $\mathcal{F}$ be the filtration generated by $\{(X_{t+1}, Y_t)\}_{t\in\mathbb{Z}}$. We assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to the filtration $\mathcal{F}$. Because ϵ_t is $\sigma(Y_t, X_t)$-measurable for any $t \in \mathbb{Z}$, and $\mathbb{X}_t$ is σX_t-measurable, $\mathcal{F}$ is finer than the filtration $\mathcal{G}$ generated by $\{(\mathbb{X}_{t+1}, \epsilon_t)\}_{t\in\mathbb{Z}}$ and therefore $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}$.

Formally, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is pairwise uncorrelated but not independent, and crucially,

$$\mathbb{E}\left[\epsilon_t \mid X_t, X_{t-1}, \cdots\right] = O_{n\times 1}$$

for any $t \in \mathbb{Z}$.

A3 : **Homoskedasticity**

There exists an $n \times n$ matrix Σ such that

$$\mathbb{E}\left[\epsilon_t \epsilon_t' \mid \mathcal{F}_{t-1}\right] = \Sigma$$

for any $t \in \mathbb{Z}$. Denote the (i, j)th element of Σ by σ_{ij}.

A4 : **No Perfect Collinearity**

There exists a positive definite $L \times L$ matrix Q such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q.$$

In particular, this implies that, for any $1 \le i \le n$, there exists an $L \times k_i$ matrix Q_i with rank k_i that collects the columns of Q corresponding to the regressors in X_t that are contained in X_{it} such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_{it}' \xrightarrow{p} Q_i.$$

A5 : **Boundedness of Moments**

$\{\epsilon_t\}_{t\in\mathbb{Z}}$ and $\{X_t\}_{t\in\mathbb{Z}}$ have finite fourth moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty, \quad \sup_{t\in\mathbb{Z}} \mathbb{E}\|X_t\|^4 < +\infty.$$

In addition, there exists some $\delta > 0$ such that, for any $1 \leq i, j \leq n$,

$$\sup_{t \in \mathbb{Z}} \mathbb{E} \left| \epsilon_t \epsilon_t \bigotimes X_t X_t' \right|^{1+\delta} < +\infty.$$

In this model, X_t serves as the same instrument for all n equations; in other words, the SURE model is a special case of the system GMM model in which $Z_{it} = X_t$ for any $1 \leq i \leq n$. Accordingly, define

$$\mathbb{Z}_t = I_n \bigotimes X_t.$$

The relevance condition for the instruments is satisfied: $k_i \leq L$, since X_t contains all the distinct variables in X_{it}, and

$$\frac{1}{T} \sum_{t=1}^{T} X_t X_{it}' \xrightarrow{p} Q_i,$$

where Q_i is an $L \times k_i$ matrix with rank k_i. It follows that the model satisfies both the order and rank conditions and is therefore fully identified.

Furthermore, since $Q_{zz,ij} = Q$ for any $1 \leq i, j \leq n$, the matrix Σ_{zz} reduces to

$$\Sigma_{zz} = \Sigma \bigotimes Q.$$

7.5.2 The Efficient System GMM Estimator

For each $1 \leq i \leq n$, the 2SLS estimator of β_i from the equation $Y_{it} = X_{it}' \beta_i + \epsilon_{it}$ obtained using the instruments X_t is given as

$$\begin{aligned} \hat{\beta}_{i,T}^{2SLS} = & \left[\left(\frac{1}{T} \sum_{t=1}^{T} X_t X_{it}' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_{it}' \right) \right]^{-1} \\ & \times \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_{it}' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} X_t Y_{it} \right). \end{aligned}$$

The equation-by-equation 2SLS estimator of β is now defined as

$$\hat{\beta}_T^{2SLS} = \begin{pmatrix} \hat{\beta}_{1,T}^{2SLS} \\ \vdots \\ \hat{\beta}_{n,T}^{2SLS} \end{pmatrix},$$

and, using this estimator, we can consistently estimate Σ via

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - \mathbb{X}_t' \hat{\beta}_T^{2SLS})(Y_t - \mathbb{X}_t' \hat{\beta}_T^{2SLS})'$$

$$= \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T}(Y_{1t} - X_{1t}'\hat{\beta}_{1,T}^{2SLS})^2 & \cdots & \frac{1}{T}\sum_{t=1}^{T}(Y_{1t} - X_{1t}'\hat{\beta}_{1,T}^{2SLS})(Y_{nt} - X_{nt}'\hat{\beta}_{n,T}^{2SLS}) \\ \vdots & \ddots & \vdots \\ \frac{1}{T}\sum_{t=1}^{T}(Y_{nt} - X_{nt}'\hat{\beta}_{n,T}^{2SLS})(Y_{1t} - X_{1t}'\hat{\beta}_{1,T}^{2SLS}) & \cdots & \frac{1}{T}\sum_{t=1}^{T}(Y_{nt} - X_{nt}'\hat{\beta}_{n,T}^{2SLS})^2 \end{pmatrix}.$$

It follows that a consistent estimator for $\Sigma \bigotimes Q$ is

$$\hat{\Sigma}_T \bigotimes \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right).$$

From the results we derived in the section on system GMM estimation, the feasible efficient system GMM estimator of β in the SURE model is given by

$$\hat{\beta}_T^{FE} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t \mathbb{X}_t' \right)' \left(\hat{\Sigma}_T^{-1} \bigotimes \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \right) \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t \mathbb{X}_t' \right) \right]^{-1}$$
$$\times \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t \mathbb{X}_t' \right)' \left(\hat{\Sigma}_T^{-1} \bigotimes \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \right) \left(\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t Y_t \right).$$

Since

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t \mathbb{X}_t' = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' & \cdots & O_{L \times k_n} \\ \vdots & \ddots & \vdots \\ O_{L \times k_1} & \cdots & \frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \end{pmatrix}$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{Z}_t Y_t = \frac{1}{T}\sum_{t=1}^{T} \mathrm{vec}\left(X_t Y_t' \right),$$

the feasible efficient system GMM estimator can be written as

$$\hat{\beta}_T^{FE} = \begin{pmatrix} \hat{\Sigma}_{T,11}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right) & \cdots & \hat{\Sigma}_{T,1n}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right) \\ \vdots & \ddots & \vdots \\ \hat{\Sigma}_{T,n1}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right) & \cdots & \hat{\Sigma}_{T,nn}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right) \end{pmatrix}^{-1}$$
$$\times \begin{pmatrix} \hat{\Sigma}_{T,11}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} & \cdots & \hat{\Sigma}_{T,1n}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{1t}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \\ \vdots & \ddots & \vdots \\ \hat{\Sigma}_{T,n1}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} & \cdots & \hat{\Sigma}_{T,nn}^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_{nt}' \right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \right)^{-1} \end{pmatrix} \left(\frac{1}{T}\sum_{t=1}^{T} \mathrm{vec}\left(X_t Y_t' \right) \right)$$

where $\hat{\Sigma}_{T,ij}^{-1}$ refers to the (i,j)th element of $\hat{\Sigma}_T^{-1}$. This estimator has the asymptotic distribution

$$\sqrt{T}\left[\hat{\beta}_T^{FE} - \beta \right] \xrightarrow{d} \mathcal{N}\left[O_{k \times 1}, \begin{pmatrix} \Sigma_{11}^{-1} \cdot Q_1' Q^{-1} Q_1 & \cdots & \Sigma_{1n}^{-1} \cdot Q_1' Q^{-1} Q_n \\ \vdots & \ddots & \vdots \\ \Sigma_{n1}^{-1} \cdot Q_n' Q^{-1} Q_1 & \cdots & \Sigma_{nn}^{-1} \cdot Q_n' Q^{-1} Q_n \end{pmatrix}^{-1} \right].$$

7.5.3 The Special Case $X_{it} = X_t$

So far, it seems like we just have a more complicated expression for the feasible efficient GMM estimator we studied in the previous section. However, it turns out that, when $X_{it} = X_t$ for each $1 \leq i \leq n$, that is, when each equation has the same regressors, the feasible efficient system GMM estimator reduces to the equation-by-equation OLS estimator.

When $X_{it} = X_t$ for any $1 \leq i \leq n$, we can see that the system of equations can be stacked as

$$Y_t = (I_n \bigotimes X_t')\beta + \epsilon_t = X_t' \underbrace{\begin{pmatrix} \beta_1 & \cdots & \beta_n \end{pmatrix}}_{\Pi'} + \epsilon_t.$$

For any $1 \leq i \leq n$,

$$\hat{\beta}_{i,T}^{2SLS} = \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} X_t Y_{it} \right) = \hat{\beta}_{i,T}^{OLS}$$

for any $1 \leq i \leq n$, so that the equation-by-equation 2SLS estimator becomes the equation-by-equation OLS esimtator. It follows that the equation-by-equation estimator of Π is given as

$$\hat{\Pi}_T^{OLS} = \begin{pmatrix} \hat{\beta}_{1,T}^{OLS\prime} \\ \vdots \\ \hat{\beta}_{n,T}^{OLS\prime} \end{pmatrix} = \left(\frac{1}{T} \sum_{t=1}^{T} Y_t X_t' \right) \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1}.$$

Note that now, the consistent estimator of Σ reduces to

$$\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - \hat{\Pi}_T^{OLS} X_t)(Y_t - \hat{\Pi}_T^{OLS} X_t)'.$$

Finally, the feasible efficient system GMM estimator of β is

$$\begin{aligned} \hat{\beta}_T^{FE} &= \left[\hat{\Sigma}_T \bigotimes \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right) \right]^{-1} \left(\hat{\Sigma}_T \bigotimes I_L \right) \left(\frac{1}{T} \sum_{t=1}^{T} \text{vec}\left(X_t Y_t' \right) \right) \\ &= \left(I_n \bigotimes \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \right) \left(\frac{1}{T} \sum_{t=1}^{T} \text{vec}\left(X_t Y_t' \right) \right) \\ &= \text{vec}\left(\left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} X_t Y_t' \right) \right). \end{aligned}$$

It is now clear that

$$\hat{\beta}_T^{OLS} = \text{vec}\left(\hat{\Pi}_T' \right) = \hat{\beta}_T^{FE},$$

so that the equation-by-equation OLS estimator of β is exactly the feasible efficient system GMM estimator of β, with asymptotic distribution given as

$$\sqrt{T}\left[\hat{\beta}_T^{OLS} - \beta\right] \xrightarrow{d} \mathcal{N}\left[O_{nL\times 1},\, \Sigma \bigotimes Q^{-1}\right].$$

This is a curious result, because both sources of the off-diagonal correlation in Σ_{zz} are still very much alive in this model. Specifically, the errors in ϵ_t are still allowed to be correlated, as evidenced by the fact that we do not impose any restrictions on the off-diagonal terms of Σ, and the instruments are perfectly correlated across equations because we are using the same instruments X_t across all equations. It just so happens that the error variance Σ cancels out as the various terms comprising $\hat{\beta}_T^{FE}$ are reduced and compressed.

Intuitively, this result makes more sense when we consult the stacked version of the model. Note that, under the setting $X_{it} = X_t$, the relationship between Y_t and X_t can be written as

$$Y_t = \Pi X_t + \epsilon_t.$$

This looks very much like the standard regression model, except now there are multiple endogenous variables and exogenous variables. Furthermore, by assumption X_t is backward exogenous for the homoskedastic error ϵ_t, so that we can expect the OLS estimator of Π to be asymptotically normal and efficient, with asymptotic variance given as some sort of product of the error variance and the inverse of the cross product of the X_ts. It turns out that this is exactly the result that we have derived above.

7.6 The Simultaneous Equations Model

The simultaneous equations model is a framework meant to help analyze how multiple endogenous and exogenous variables interact with one another. One example is the simple demand and supply problem, in which the equilibrium quantity and price are both endogenous variables determined by a number of exogenous demand and supply shifters. Another example is the vector autoregression (VAR) model that is used to model the dynamic behavior of multiple macroeconomic variables in time series analysis; we will study VARs in much greater detail later on.

7.6.1 Model Assumptions

The model assumptions are as follows.

A1 : **The Sample**

There is a sample $\{(Y_t, X_t)\}_{t\in\mathbb{Z}}$, where each Y_t takes values in $\mathbb{R}^n$ and X_t takes values in $\mathbb{R}^k$. In other words, there are n endogenous variables and k exogenous variables. The sample observations are assumed to be dependent, with the degree of dependence to be specified below. The errors satisfy the equation

$$AY_t = BX_t + \epsilon_t,$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times k}$ are non-random matrices. In particular, A is non-singular.

A2 :**No Perfect Collinearity**

We assume that there exists a positive definite matrix $Q \in \mathbb{R}^{k\times k}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} X_t X_t' \xrightarrow{p} Q.$$

By implication, $\frac{1}{T}\sum_{t=1}^{T} X_t X_t'$ is positive definite with probability approaching 1 as $T \to \infty$.

A3 : **Exogeneity**

Let $\mathcal{F}$ be the filtration generated by $\{(X_{t+1}, Y_t)\}_{t\in\mathbb{Z}}$. We assume that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}$. As usual, because ϵ_t is $\sigma(Y_t, X_t)$-measurable for any $t \in \mathbb{Z}$, $\mathcal{F}$ is finer than the filtration $\mathcal{G}$ generated by $\{(X_{t+1}, \epsilon_t)\}_{t\in\mathbb{Z}}$ and therefore $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}$.

By implication, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is pairwise uncorrelated and, crucially,

$$\mathbb{E}\left[\epsilon_t \mid X_t, X_{t-1}, \cdots\right] = O_{n\times 1}$$

for any $t \in \mathbb{Z}$.

A4 : **Homoskedasticity**

We assume that the errors are homoskedastic with positive definite covariance matrix $W \in \mathbb{R}^{n \times n}$. Formally,

$$\mathbb{E}\left[\epsilon_t \epsilon_t' \mid \mathcal{F}_{t-1}\right] = V$$

for any $t \in \mathbb{Z}$.

A5 : **Boundedness of Moments**

$\{\epsilon_t\}_{t \in \mathbb{Z}}$ an $\{X_t\}_{t \in \mathbb{Z}}$ have finite fourth moments:

$$\sup_{t \in \mathbb{Z}} \mathbb{E}|\epsilon_t|^4 < +\infty, \quad \sup_{t \in \mathbb{Z}} \mathbb{E}|X_t|^4 < +\infty.$$

In addition, there exists some $\delta > 0$ such that

$$\sup_{t \in \mathbb{Z}} \mathbb{E}\left\|\epsilon_t \epsilon_t \bigotimes X_t X_t'\right\|^{1+\delta} < +\infty.$$

7.6.2 The Structural and Reduced Form Models

The equation

$$AY_t = BX_t + \epsilon_t$$

is referred to as the structural form of the model. It suggests that the endogenous variables Y_t may determine each other simultaneously, that is, other endogenous variables can appear as regressors in the equation corresponding to each endogenous variable. In this sense, each equation in the structural model contains both endogenous and exogenous regressors, and we need to adopt estimation methods that account for the endogeneity problem.

The reduced form model that corresponds to this model can be obtained by pre-multiplying both sides of the model by A^{-1}; defining $A^{-1}B = \Pi \in \mathbb{R}^{n\times k}$ and $u_t = A^{-1}\epsilon_t$, we can now see that Y_t is determined as

$$Y_t = \Pi X_t + u_t.$$

Note that this is precisely the SURE model with identical regressors that we studied in the previous section. Here, $\{u_t\}_{t\in\mathbb{Z}}$ is the stochastic process representing the reduced-form error terms.

System GMM Estimation of the Reduced Form Model

Analysis of the reduced form model is particularly simple in light of the SURE model framework introduced above; as such, we take some time here to derive the asymptotic properties of the feasible efficient system GMM estimators $\hat{\Pi}_T$ and $\hat{\Sigma}_T$ of Π and Σ. To this end, note first that the reduced form errors satisfy the following conditions:

i) **Exogeneity**
$\{u_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to $\mathcal{F}$: $u_t = A^{-1}\epsilon_t$ is $\mathcal{F}_t$-measurable for any $t \in \mathbb{Z}$, making $\{u_t\}_{t\in\mathbb{Z}}$ an $\mathcal{F}$-adapted process; u_t is integrable because ϵ_t is; and for any $t \in \mathbb{Z}$,

$$\mathbb{E}[u_t \mid \mathcal{F}_{t-1}] = A^{-1} \cdot \mathbb{E}[\epsilon_t \mid \mathcal{F}_{t-1}] = O_{n\times 1}$$

because $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an MDS with respect to $\mathcal{F}$.

Defining $\mathcal{G}'$ as the filtration generated by $\{(X_{t+1}, u_t)\}_{t\in\mathbb{Z}}$, since $\mathcal{F}$ is finer than $\mathcal{G}'$ ($u_t = Y_t - A^{-1}BX_t$ for any $t \in \mathbb{Z}$, so that $\mathcal{G}'_t \subset \mathcal{F}_t$ for any $t \in \mathbb{Z}$), $\{u_t\}_{t\in\mathbb{Z}}$ is also an MDS with respect to $\mathcal{G}'$.

ii) **Homoskedasticity**

For any $t \in \mathbb{Z}$,

$$\mathbb{E}\left[u_t u_t' \mid \mathcal{F}_{t-1}\right] = A^{-1} \cdot \mathbb{E}\left[\epsilon_t \epsilon_t' \mid \mathcal{F}_{t-1}\right] A^{-1\prime} = A^{-1} V A^{-1\prime},$$

so that $\{u_t\}_{t\in\mathbb{Z}}$ is homoskedastic in the same sense as $\{\epsilon_t\}_{t\in\mathbb{Z}}$. Defining

$$\Sigma = A^{-1} V A^{-1\prime},$$

Σ is the reduced form error variance in contrast to the structural error variance V.

iii) **Boundedness of Moments**

$\{u_t\}_{t\in\mathbb{Z}}$ has finite fourth moments:

$$\sup_{t\in\mathbb{Z}} \mathbb{E}|u_t|^4 \leq \left\|A^{-1}\right\|^4 \left(\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^4\right) < +\infty.$$

In addition, since

$$u_t u_t' \bigotimes X_t X_t' = A^{-1}\left(\epsilon_t \epsilon_t' \bigotimes X_t X_t'\right) A^{-1\prime},$$

for any $t \in \mathbb{Z}$

$$\mathbb{E}\left\|u_t u_t' \bigotimes X_t X_t'\right\|^{1+\delta} \leq \left\|A^{-1}\right\|^{2(1+\delta)} \cdot \mathbb{E}\left\|\epsilon_t \epsilon_t' \bigotimes X_t X_t'\right\|^{1+\delta},$$

it follows that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|u_t u_t' \bigotimes X_t X_t'\right\|^{1+\delta} \leq \left\|A^{-1}\right\|^{2(1+\delta)} \left(\sup_{t\in\mathbb{Z}} \mathbb{E}\left\|\epsilon_t \epsilon_t' \bigotimes X_t X_t'\right\|^{1+\delta}\right) < +\infty.$$

As such, the reduced-form model

$$Y_t = \Pi X_t + u_t$$

satisfies all the assumptions of the SURE model we studied in the previous section. It follows that the feasible efficient system GMM estimator $\hat{\beta}_T = \operatorname{vec}\left(\hat{\Pi}_T'\right)$ of $\beta = \operatorname{vec}(\Pi')$ is given as the equation-by-equation OLS estimator

$$\hat{\beta}_T = \operatorname{vec}\left(\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} X_t Y_t'\right)\right) = \operatorname{vec}\left(\hat{\Pi}_T'\right),$$

and that a consistent estimator for the reduced form error variance Σ is given by

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{\Pi}_T X_t)(Y_t - \hat{\Pi}_T X_t)'.$$

Recall also that $\hat{\beta}_T$ is asymptotically normal with the following asymptotic distribution:

$$\sqrt{T}\left[\hat{\beta}_T - \beta\right] \xrightarrow{d} \mathcal{N}\left[O_{k\times 1},\, \Sigma \bigotimes Q^{-1}\right].$$

The Distinction between Structural and Reduced Form

Heuristically, if the structural model is closer to an economist's model, since it specifies the various interrelationships between the endogenous and exogenous variables, then the reduced form model describes the data generating process of the endogenous and exogenous variables that the econometrician observes. This is because the likelihood of the data depends only on the reduced form parameters, which will be studied in more detail below. Thus, it is the econometrician's job to recover the structural parameters, which are the parameters of interest, from the reduced form parameters, which the econometrician can observe.

Below, we study three ways to estimate the model. The first is maximum likelihood under the assumption of Gaussian errors[4]. The second is GMM using the primitive moment conditions $\mathbb{E}[X_t\epsilon_t'] = O_{k\times n}$, like what we used for the GMM estimation of the basic univariate linear regression model. It turns out that these two approaches produce the exact same estimators, given that the model is correctly identified. Finally, we study three stage least squares, which is the system GMM method studied in previous sections adapted to the simultaneous equations model.

[4]The Gaussian parametric assumption is made for the sake of illustration; we can impose any (known) distributive assumption on the errors and proceed with the same maximum likelihood analysis.

7.6.3 Full Information Maximum Likelihood

One way to estimate the model is through full information maximum likelihood (FIML) under the assumption of Gaussian errors. Formally, we make the following assumption:

A7 : **Gaussian Errors**

For any $t \in \mathbb{Z}$, assume that

$$\epsilon_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}[O_{n\times 1}, V].$$

Since the right hand side does not depend on $\mathcal{F}_{t-1}$, it follows that

$$\epsilon_t \mid X_t, Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1 \sim \mathcal{N}[O_{n\times 1}, V]$$

for any $t \geq 1$.

In addition, for any $t \geq 1$, the conditional distribution of X_t given $Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1$ is absolutely continuous with respect to the Lebesgue measure (so that it admits a density), and depends on the parameters θ, which are independent of A, B, V.

Note that, for any $t \geq 1$,

$$u_t \mid X_t, Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1 \sim \mathcal{N}[O_{n\times 1}, \Sigma],$$

and since $Y_t = \Pi X_t + u_t$,

$$Y_t \mid X_t, Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1 \sim \mathcal{N}[\Pi X_t, \Sigma].$$

Before maximizing the likelihood, we must first address the identification problem. In general, the structural model is not identified because there are more structural parameters than reduced form parameters, but the likelihood depends only on the reduced form parameters.

Identification under Gaussian Errors

In a sample of size T, the joint likelihood of $\{(Y_t, X_t)\}_{1\leq t\leq T}$ can be decomposed as

$$f(Y_T, X_T, \cdots, Y_1, X_1 \mid A, B, V) = \left[\prod_{t=2}^{T} f(Y_t, X_t \mid Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1; A, B, V)\right] \cdot f(Y_1, X_1; A, B, V)$$

$$= \left[\prod_{t=2}^{T} f(Y_t \mid X_t, Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1; \Pi, \Sigma) \times f(X_t \mid Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1; \theta)\right]$$
$$\times f(Y_1 \mid X_1; \Pi, \Sigma) \times f(X_1 \mid \theta)$$

$$= \left(\prod_{t=1}^{T} \mathcal{N}\left[Y_t \mid \Pi X_t,\ \Sigma \right] \right) \cdot \left(\prod_{t=1}^{T} f(X_t \mid Y_{t-1}, X_{t-1}, \cdots, Y_1, X_1; \theta) \right).$$

Thus, the (full information) log-likelihood becomes

$$l_T(A, B, V \mid Y_T, X_T, \cdots, Y_1, X_1) = \sum_{t=1}^{T} \left[-\frac{1}{2} \log|\Sigma| - \frac{1}{2} (Y_t - \Pi X_t)' \Sigma^{-1} (Y_t - \Pi X_t) \right] + \text{terms without } A, B, V$$

$$= -\frac{T}{2} \log|\Sigma| - \frac{1}{2} \sum_{t=1}^{T} (Y_t - \Pi X_t) \Sigma^{-1} (Y_t - \Pi X_t)' + \text{terms without } A, B, V.$$

This shows us that the log-likelihood depends only on the structural parameters A, B, V via the reduced form parameters Π and Σ. In other words, if two sets of structural parameters $\{A', B', V'\}$ and $\{A, B, V\}$ result in the same reduced form parameters $\{\Pi, \Sigma\}$, then both sets of structural parameters lead to the exact same likelihood and therefore the data cannot tell us anything about which of the two are the true structural parameters.

When are the structural parameters identified? They are identified when the equations

$$A^{-1}B = \Pi, \quad A^{-1} V A^{-1\prime} = \Sigma$$

can be solved for a unique set of structural parameters A, B, V. Once again, we can decompose this requirement into two familiar conditions:

i) **The Order Condition**

Note that there are $nk + \frac{n(n+1)}{2}$ equations for each of the elements in Π, Σ that we must solve for $n^2 + nk + \frac{n(n+1)}{2}$ unknowns in A, B, V; here, there are only $\frac{n(n+1)}{2}$ unique elements in Σ and V because they are symmetric. As such, we must impose at least n^2 restrictions on the structural parametes A, B, V in order for there to exist a unique solution to the above system of equations. This is once again a necessary condition for identification.

Usually, we immediately impose $\frac{n(n+1)}{2}$ restrictions by assuming that V is diagonal (that is, the structural errors are cross-sectionally uncorrelated) and by assuming in addition that either the diagonal entries of V are equal to 1 (so that the shocks are normalized to be of similar magnitudes), or the diagonal entries of A are equal to 1.

ii) **The Rank Condition**

In this context, the rank condition refers to any kind of sufficient conditions that A, B, V can satisfy in order for there to exist a unique solution to the above system of $nk + \frac{n(n+1)}{2}$ equations. One possible condition, popularly used in the literature on VARs, requires $V = I_n$ and for A to be lower triangular. In this case, A^{-1} is recovered uniquely (up to sign changes) as the Cholesky factor of Σ, and B is reovered uniquely as $B = A\Pi$.

Deriving the FIML Estimators

Suppose that the model is just identified, so that, given any $\Pi \in \mathbb{R}^{n\times k}$ and positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$, the equations

$$A^{-1}B = \Pi, \quad A^{-1}VA^{-1\prime} = \Sigma$$

can be uniquely solved for a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, and a positive definite matrix $V \in \mathbb{R}^{n\times n}$.

From our derivation of the log-likelihood above, we can see that, given a sample of size T, the FIML estimators of A, B, V can be found by maximizing the objective function

$$\begin{aligned}\mathbb{Q}_T(Y_T, X_T, \cdots, Y_1, X_1 \mid A, B, V) &= -\frac{T}{2}\log|\Sigma| - \frac{1}{2T}\sum_{t=1}^{T}(Y_t - \Pi X_t)'\Sigma^{-1}(Y_t - \Pi X_t) \\ &= n\cdot\log|A| - \frac{n}{2}\log|V| - \frac{1}{2T}\sum_{t=1}^{T}(AY_t - BX_t)'V^{-1}(AY_t - BX_t)\end{aligned}$$

with respect to A, B, V. We can find the FIMLEs in two steps: first, we find the FIML estimators of the reduced form parameters, denoted $\hat{\Pi}_T^{FIML}$ and $\hat{\Sigma}_T^{FIML}$. Then, the FIMLEs of A, B, V can be found by solving the equations

$$A^{-1}B = \hat{\Pi}_T^{FIML}, \quad A^{-1}VA^{-1\prime} = \hat{\Sigma}_T^{FIML}$$

for A, B and V.

To derive the first order conditions for maximization, it is useful to first simplify the objective function. For any $T \in N_+$, define

$$\mathbb{Y}^{(T)} = \begin{pmatrix} Y_1' \\ \vdots \\ Y_T' \end{pmatrix}, \quad \mathbb{X}^{(T)} = \begin{pmatrix} X_1' \\ \vdots \\ X_T' \end{pmatrix}.$$

Then,

$$\begin{aligned}\sum_{t=1}^{T}(Y_t - \Pi X_t)(Y_t - \Pi X_t)' &= \begin{pmatrix} Y_1 - \Pi X_1 & \cdots & Y_T - \Pi X_T \end{pmatrix}\begin{pmatrix} (Y_1 - \Pi X_1)' \\ \vdots \\ (Y_T - \Pi X_T)' \end{pmatrix} \\ &= \left(\mathbb{Y}^{(T)} - \mathbb{X}^{(T)}\Pi'\right)'\left(\mathbb{Y}^{(T)} - \mathbb{X}^{(T)}\Pi'\right) \\ &= \mathbb{Y}^{(T)\prime}\mathbb{Y}^{(T)} - \Pi\cdot\mathbb{X}^{(T)\prime}\mathbb{Y}^{(T)} + \mathbb{Y}^{(T)\prime}\mathbb{X}^{(T)}\cdot\Pi' + \Pi\cdot\mathbb{X}^{(T)\prime}\mathbb{X}^{(T)}\cdot\Pi'.\end{aligned}$$

Then, the objective function becomes

$$\begin{aligned}\mathbb{Q}_T(\Pi,\Sigma) &= -\frac{T}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\cdot\sum_{t=1}^{T}(Y_t-\Pi X_t)(Y_t-\Pi X_t)'\right)\\ &= -\frac{T}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\cdot\left(\mathbb{Y}^{(T)}-\mathbb{X}^{(T)}\Pi'\right)'\left(\mathbb{Y}^{(T)}-\mathbb{X}^{(T)}\Pi'\right)\right]\\ &= -\frac{T}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\cdot\mathbb{Y}^{(T)\prime}\mathbb{Y}^{(T)}\right] + \operatorname{tr}\left[\Pi\cdot\mathbb{X}^{(T)\prime}\mathbb{Y}^{(T)}\Sigma^{-1}\right]\\ &\quad - \frac{1}{2}\operatorname{tr}\left[\Pi\cdot\mathbb{X}^{(T)\prime}\mathbb{X}^{(T)}\cdot\Pi'\Sigma^{-1}\right].\end{aligned}$$

Recall the following results from matrix calculus:

$$\frac{\partial|D|}{\partial D} = |D|\cdot D^{-1\prime},$$

$$\frac{\partial\operatorname{tr}(MP)}{\partial M} = P',$$

$$\frac{\partial\operatorname{tr}(MPM'C)}{\partial M} = CMP + C'MP$$

for any square matrix D and conformable matrices M, P, C. As such, the first order conditions for maximization become

$$\frac{\partial\mathbb{Q}_T(\hat{\Pi}_T^{FIML},\hat{\Sigma}_T^{FIML})}{\partial\Pi} = \left(\hat{\Sigma}_T^{FIML}\right)^{-1}\mathbb{Y}^{(T)\prime}\mathbb{X}^{(T)} - \left(\hat{\Sigma}_T^{FIML}\right)^{-1}\hat{\Pi}_T^{FIML}\mathbb{X}^{(T)\prime}\mathbb{X}^{(T)} = O_{n\times k}$$

and

$$\frac{\partial\mathbb{Q}_T(\hat{\Pi}_T^{FIML},\hat{\Sigma}_T^{FIML})}{\partial\Sigma^{-1}} = \frac{T}{2}\hat{\Sigma}_T^{FIML} - \frac{1}{2}\left(\mathbb{Y}^{(T)}-X^{(T)}\hat{\Pi}_T^{FIML\prime}\right)'\left(\mathbb{Y}^{(T)}-X^{(T)}\hat{\Pi}_T^{FIML\prime}\right) = O_{n\times n}.$$

It follows that

$$\hat{\Pi}_T^{FIML} = \mathbb{Y}^{(T)\prime}\mathbb{X}^{(T)}\left(\mathbb{X}^{(T)\prime}\mathbb{X}^{(T)}\right)^{-1} = \left(\frac{1}{T}\sum_{t=1}^{T}Y_tX_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T}X_tX_t\right)^{-1}$$

and

$$\begin{aligned}\hat{\Sigma}_T^{FIML} &= \frac{1}{T}\left(\mathbb{Y}^{(T)}-X^{(T)}\hat{\Pi}_T^{FIML\prime}\right)'\left(\mathbb{Y}^{(T)}-X^{(T)}\hat{\Pi}_T^{FIML\prime}\right)\\ &= \frac{1}{T}\sum_{t=1}^{T}(Y_t-\hat{\Pi}_T^{FIML}X_t)(Y_t-\hat{\Pi}_T^{FIML}X_t)'.\end{aligned}$$

Note that these are exactly the feasible efficient system GMM estimators of the reduced form model parameters derived earlier.

7.6.4 GMM Estimation

Without the assumption of Gaussian errors, or indeed any distributional assumption on the errors, we can turn to GMM estimation. Suppose we make the normalization that $V = I_n$, so that we are left with only $n^2 + nk$ structural parameters to estimate in A, B. We have $nk + \frac{n(n+1)}{2}$ moment conditions: the first nk conditions are drawn from the exogeneity of the regressors,

$$\mathbb{E}\left[X_t \epsilon_t'\right] = \mathbb{E}\left[X_t (AY_t - BX_t)'\right] = O_{k \times n}$$

and the remaining $\frac{n(n+1)}{2}$ from the homoskedasticity condition:

$$\mathbb{E}\left[\epsilon_t \epsilon_t' - V\right] = \mathbb{E}\left[(AY_t - BX_t)(AY_t - BX_t)' - V\right] = O_{n \times n},$$

where the symmetry of the integrand indicates that there are only $\frac{n(n+1)}{2}$ distinct equations here. Again, the identification problem rears its head; in general, the structural model is unidentified even in the context of GMM.

Identification in the GMM Approach

Formally, note that the sample analogue of the moment conditions are

$$\frac{1}{T} \sum_{t=1}^{T} X_t (AY_t - BX_t)' = O_{k \times n}$$

and

$$\frac{1}{T} \sum_{t=1}^{T} (AY_t - BX_t)(AY_t - BX_t)' - V = O_{n \times n}.$$

Letting A_0, B_0, V_0 be the true structural parameters, the probability limits of the left hand side terms are given as

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} X_t (A(\Pi_0 \cdot X_t + u_t) - BX_t)' &= \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right) (A'\Pi_0' - B') + \frac{1}{T} \sum_{t=1}^{T} X_t u_t' \\ &\xrightarrow{p} Q(A'\Pi_0' - B') \end{aligned}$$

and

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} &(A(\Pi_0 \cdot X_t + u_t) - BX_t)(A(\Pi_0 \cdot X_t + u_t) - BX_t)' - V \\ &\xrightarrow{p} (A\Pi_0 - B) Q (A\Pi_0 - B)' + A\Sigma_0 A - V, \end{aligned}$$

where $\Pi_0 = A_0^{-1}B_0$ and $\Sigma_0 = A_0^{-1}V_0A_0^{-1\prime}$. These population moment conditions are exactly equal to 0 if and only if

$$\Pi_0 = A^{-1}B, \quad \Sigma_0 = A^{-1}VA^{-1\prime},$$

so that the model is unidentified unless there exist unique solutions A, B, V to the above $nk + \frac{n(n+1)}{2}$ equations. Once again, this leads to the same rank and order conditions as above.

Deriving the GMM Estimators

As in the FIML case, suppose that the model is just identified, so that, given any $\Pi \in \mathbb{R}^{n\times k}$ and positive definite matrix $\Sigma \in \mathbb{R}^{n\times n}$, the equations

$$A^{-1}B = \Pi, \quad A^{-1}VA^{-1\prime} = \Sigma$$

can be uniquely solved for a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$, and a positive definite matrix $V \in \mathbb{R}^{n\times n}$.

In this just-identified case, by the order condition there are a total of $nk + \frac{n(n+1)}{2}$ parameters that we must solve given a set of $nk + \frac{n(n+1)}{2}$ primitive moment conditions, whose sample analogues are given by

$$\frac{1}{T}\sum_{t=1}^{T} X_t(AY_t - BX_t)' = O_{k\times n}$$

$$\frac{1}{T}\sum_{t=1}^{T}(AY_t - BX_t)(AY_t - BX_t)' = V.$$

Since there are the same number of equations and unknowns, we need not specify a sequence of weighting matrices; instead, our GMM estimators of A, B, V are the values that directly solve the above sample moment conditions.

Like in the FIML case, the GMM estimators of the structural parameters can be found in two steps: first, we find the GMM estimators of the reduced form parameters, denoted $\hat{\Pi}_T$ and $\hat{\Sigma}_T$. Then, the GMM estimators of A, B, V can be found by solving the equations

$$A^{-1}B = \hat{\Pi}_T, \quad A^{-1}VA^{-1\prime} = \hat{\Sigma}_T$$

for A, B and V.

In terms of the reduced form parameters, the sample moment conditions become

$$\frac{1}{T}\sum_{t=1}^{T} X_t(Y_t - \Pi X_t)' = O_{k\times n}$$

$$\frac{1}{T}\sum_{t=1}^{T}(Y_t - \Pi X_t)(Y_t - \Pi X_t)' = \Sigma.$$

Therefore,

$$\hat{\Pi}_T = \left(\frac{1}{T}\sum_{t=1}^{T} Y_t X_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}$$

and

$$\hat{\Sigma}_T = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{\Pi}_T X_t)(Y_t - \hat{\Pi}_T X_t)'.$$

Note that the GMM estimators of Π and Σ are identical to their FIML estimators, which are in turn just the feasible efficient GMM estimators of the reduced form model parameters. This is due to the quadratic form of the Gaussian likelihood in conjunction with the homoskedasticity assumption, and mirrors how, in the univariate case, the MLE and GMM estimators are all equal to the OLS estimator.

It follows from the fact that the GMM estimators $\hat{A}_T, \hat{B}_T$ and $\hat{V}_T$ are continuous functions of $\hat{\Pi}_T$ and $\hat{\Sigma}_T$ that $\hat{A}_T, \hat{B}_T$ and $\hat{V}_T$ are consistent for A, B and V as well. Their asymptotic distributions, however, depend on the asymptotic distribution not only of $\hat{\Pi}_T$ but also that of $\hat{\Sigma}_T$. Given the asymptotic distribution of $\hat{\Sigma}_T$, we can obtain the asymptotic distributions of $\hat{A}_T, \hat{B}_T$ and $\hat{V}_T$ via the Delta method.

7.6.5 Three Stage Least Squares

Three stage least squares (3SLS) is another method that can be used to estimate the parameters of the structural model, and it is an adaptation of the system GMM method that we studied at the start of this chapter. 3SLS involves imposing restrictions on the matrices A and B in a manner that allows us to use the entire set of exogenous variables X_t as an instrument for each of the n equations in the model.

Formally, normalize the diagonal elements of A to 1. For any $1 \leq i \leq n$, let α_{-i} be an m_i-dimensional vector collecting the non-zero elements of the ith row of $-A$ excluding $-A_{ii}$. Similarly, let β_i be the k_i-dimensional vector collecting the non-zero elements of the ith row of B. Letting $Y_{-i,t}$ collect the endogenous variables corresponding to the non-zero elements of the ith row of A, and X_{it} collect the exogenous variables corresponding to the non-zero elements of the ith row of B, the ith equation in the model can be written as

$$Y_{it} = Y'_{-i,t}\alpha_{-i} + X'_{it}\beta_i + \epsilon_{it}.$$

Defining

$$W_{it} = \begin{pmatrix} Y_{-i,t} \\ X_{it} \end{pmatrix}, \quad \gamma_i = \begin{pmatrix} \alpha_{-i} \\ \beta_i \end{pmatrix},$$

the ith equation can now be written as

$$Y_{it} = W'_{it}\gamma_i + \epsilon_{it}.$$

This holds for each $1 \leq i \leq n$, so defining

$$W_t = \begin{pmatrix} W_{1t} & \cdots & O_{(m_1+k_1)\times(m_n+k_n)} \\ \vdots & \ddots & \vdots \\ O_{(m_n+k_n)\times(m_1+k_1)} & \cdots & W_{nt} \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{pmatrix},$$

the stacked version of the model is

$$Y_t = W'_t\gamma + \epsilon_t.$$

Since each W_{it} may contain some endogenous variables, in itself the structural model suffers from the endogeneity problem. This can be alleviated by using the same set of instruments $Z_{it} = X_t$ for each of the n equations, since X_t satisfies the exogeneity assumption with respect to the structural errors ϵ_t. For the model to be identified, however, we need to consider whether the order and rank conditions are satisfied.

Identification in the System GMM Approach

Recall that, in the system GMM model, the two conditions for identification are given as follows:

i) **The Order Condition**
For each $1 \leq i \leq n$, the number of total instruments for the ith equation, k, must be larger than or equal to the number of regressors for the ith equation, $m_i + k_i$. In other words, we must have $m_i + k_i \leq k$.

ii) **The Rank Condition**
The probability limit of the matrix

$$\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t'$$

must be a matrix of rank $K = \sum_{i=1}^{n}(k_i + m_i)$.

We show that these conditions are satisfied, and that the model is therefore identified, under the following additional assumptions:

A7 : **Appropriate Zero Restrictions**
We assume that zero restrictions have been imposed on the structural parameters A and B so that $k_i + m_i \leq k$ for any $1 \leq i \leq n$. This involves the imposition of at least n^2 zero restrictions on the elements of A and B.

A8 : **Full Rank Reduced Form Parameter**
For any $1 \leq i \leq n$, let $\Pi_{-i,i}$ be the $m_i \times (k - k_i)$ submatrix of $\Pi = A^{-1}B$, where the m_i rows are chosen to correspond to the endogenous variables in $Y_{-i,t}$ and the $k - k_i$ columns are chosen to correspond to the exogenous variables that are excluded from X_{it}. We assume that $\Pi_{-i,i}$ has rank m_i.

The following result shows us that, under the above assumptions, the rank condition is satisfied and therefore the model is identified.

Lemma 7.8 (Rank Condition for 3SLS)
Suppose assumptions A1 to A8 hold. Then, the rank condition for identification is satisfied, that is, the probability limit of the $nk \times K$ matrix $\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t$ has rank K.

Proof) Choose any $1 \leq i \leq n$, and let Π_{-i} collect the rows of Π corresponding to $Y_{-i,t}$, and choose $\epsilon_{-i,t}$ in a similar manner. Then,

$$Y_{-i,t} = \Pi_{-i}X_t + \epsilon_{-i,t},$$

so that

$$\frac{1}{T}\sum_{t=1}^{T} X_t Y'_{-i,t} = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X'_t\right)\Pi'_{-i} + \frac{1}{T}\sum_{t=1}^{T} X_t \epsilon'_{-i,t} \xrightarrow{p} Q\Pi'_{-i}.$$

Furthermore, let $S_i \in \mathbb{R}^{k\times k_i}$ be a matrix such that QS_i collects the columns of Q corresponding to the regressors chosen in X_{it}. It follows that

$$\frac{1}{T}\sum_{t=1}^{T} X_t W'_{it} = \left(\frac{1}{T}\sum_{t=1}^{T} X_t Y'_{-i,t} \quad \frac{1}{T}\sum_{t=1}^{T} X_t X'_{it}\right) \xrightarrow{p} Q\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}.$$

The rank condition will be satisfied if $Q\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}$ has rank $k_i + m_i$. Since this is $k \times (k_i + m_i)$ matrix, where there are more columns than rows, by the dimension theorem this matrix with have rank $k_i + m_i$ if it has nullity equal to 0, that is, if

$$Q\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = O_{k\times 1}$$

for some $a \in \mathbb{R}^{m_i}$ and $b \in \mathbb{R}^{k_i}$ implies $a = O_{m_i\times 1}$ and $b = O_{k_i\times 1}$. To this end, suppose that

$$Q\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = O_{k\times 1}.$$

Because Q is non-singular by assumption,

$$\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix} = \Pi'_{-i}a + S_i \cdot b = O_{k\times 1}.$$

Note that, because $S_i \cdot b$ is the $k \times 1$ vector formed by assinging each element of b to the indices corresponding to the exogenous variables chosen in X_{it}, and the columns of $\Pi_{-i,i}$ are chosen to correspond to the exogenous variables that are not included in X_{it}, the above equation implies that

$$\Pi'_{-i,i}a = O_{(k-k_i)\times 1}.$$

Since $\Pi'_{-i,i}$ is an $(k - k_i) \times m_i$ matrix with rank m_i, by the dimension theorem $\Pi'_{-i,i}$ has nullity 0. This implies that $a = O_{m_i\times 1}$. By implication,

$$S_i \cdot b = O_{k\times 1},$$

which immediately tells us that $b = O_{k_i\times 1}$ as well. This proves that $Q\begin{pmatrix}\Pi'_{-i} & S_i\end{pmatrix}$ has rank $k_i + m_i$, and since this holds for any $1 \le i \le n$, the probability limit of the block

diagonal matrix

$$\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} X_t W_{1t}' & \cdots & O_{k\times(k_n+m_n)} \\ \vdots & \ddots & \vdots \\ O_{k\times(k_1+m_1)} & \cdots & \frac{1}{T}\sum_{t=1}^{T} X_t W_{mt}' \end{pmatrix}$$

has rank $K = \sum_{i=1}^{n}(k_i + m_i)$.

Q.E.D.

In light of the above result, we can see that there exists a block-diagonal $k \times K$ matrix C_{xw} of rank K such that

$$\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t' \xrightarrow{p} (I_n \bigotimes Q)\cdot C_{xw}.$$

Deriving the 3SLS Estimators

Now that the model is identified, we can see that the efficient system GMM estimator of γ is given as

$$\begin{aligned}\hat{\gamma}_T^E = &\left[\left(\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t'\right)' \left(V^{-1} \bigotimes Q^{-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t'\right)\right]^{-1} \\ &\times \left(\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)W_t'\right)' \left(V^{-1} \bigotimes Q^{-1}\right)\left(\frac{1}{T}\sum_{t=1}^{T}(I_n \bigotimes X_t)Y_t\right),\end{aligned}$$

which has asymptotic distribution

$$\sqrt{T}\left[\hat{\gamma}_T^E - \gamma\right] \xrightarrow{d} \mathcal{N}\left[O_{K\times 1},\ \left(C_{xw}'\left(V^{-1} \bigotimes Q\right)C_{xw}\right)^{-1}\right].$$

Again, feasbility is the issue, since we do not observe the variance V of the structural errors. Once again, we turn to the two stage least squares estimators to obtain an initial consistent estimator of γ that does not rely on having consistently estimated Σ beforehand. For any $1 \leq i \leq n$, denote by $\hat{\gamma}_{i,T}^{2SLS}$ the 2SLS estimator of γ_i defined as

$$\hat{\gamma}_{i,T}^{2SLS} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t W_{it}'\right)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} X_t W_{it}'\right)\right]^{-1}$$

$$\times \left(\frac{1}{T} \sum_{t=1}^{T} X_t W_{it}' \right)' \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} X_t Y_{it} \right).$$

The equation-by-equation 2SLS estimator of γ is then given as

$$\hat{\gamma}_T^{2SLS} = \begin{pmatrix} \hat{\gamma}_{1,T}^{2SLS} \\ \vdots \\ \hat{\gamma}_{n,T}^{2SLS} \end{pmatrix},$$

and it is consistent for γ. Our estimator of V can finally be constructed as

$$\hat{V}_T = \frac{1}{T} \sum_{t=1}^{T} (Y_t - W_t' \hat{\gamma}_T^{2SLS})(Y_t - W_t' \hat{\gamma}_T^{2SLS})',$$

which we showed earlier is consistent for V. As such, the feasible efficient system GMM estimator of γ, also called the 3SLS estimator of γ, is given as

$$\begin{aligned} \hat{\gamma}_T^{3SLS} = & \left[\left(\frac{1}{T} \sum_{t=1}^{T} (I_n \bigotimes X_t) W_t' \right)' \left(\hat{V}_T^{-1} \bigotimes \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \right) \left(\frac{1}{T} \sum_{t=1}^{T} (I_n \bigotimes X_t) W_t' \right) \right]^{-1} \\ & \times \left(\frac{1}{T} \sum_{t=1}^{T} (I_n \bigotimes X_t) W_t' \right)' \left(\hat{V}_T^{-1} \bigotimes \left(\frac{1}{T} \sum_{t=1}^{T} X_t X_t' \right)^{-1} \right) \left(\frac{1}{T} \sum_{t=1}^{T} (I_n \bigotimes X_t) Y_t \right), \end{aligned}$$

and it has the same asymptotic distribution as the efficient system GMM estimator $\hat{\gamma}_T^E$.

To see where the term "three stage least squares" comes from, note that $\hat{\gamma}_T^{3SLS}$ is derived in three steps:

Step 1 : **Equation-by-Equation First Stage Regression**
For the ith equation, we compute the first stage regression by regressing the original regressors W_{it} on the instruments X_t.

Step 2 : **Equation-by-Equation Second Stage Regression**
For the ith equation, we compute the 2SLS estimator by regressing Y_{it} on the fitted values from the first stage regression.

Step 3 : **System Computation for Efficiency Gains**
Using the equation-by-equation 2SLS estimator of γ obtained in the first two steps, we construct the consistent estimator $\hat{V}_T$ of V and, using this estimator, compute the feasible efficient system GMM estimator of γ.

In other words, the extra stage in 3SLS comes from the last step, in which the equation-by-equation 2SLS estimator is used to construct a consistent estimator for the error variance V,

which is in turn used to compute the feasible efficient system GMM estimator. The added third stage has the advantage of resulting in efficiency gains. since $\hat{\gamma}_T^{3SLS}$ is the asymptotically efficient GMM estimator of γ; as mentioned earlier, these efficiency gains are large only when the structural errors are correlated, or when the off-diagonal terms in V are non-zero.

Chapter 8

Non-Parametric Methods

This chapter is a soft introduction to non-parametric methods. In traditional statistical problems, there is a (usually finite-dimensional) unknown parameter of interest that we wish to estimate using the data at hand. In parametric models, this parameter can be seen as a summary of the distribution from which the data are generated; think of the problem of estimating the population mean and variance given a normal DGP for the data. In semi-parametric models such as linear regressions and GMM, the parameter determines a key feature, but not the entirety, of the data. For instance, in linear regression models, the parameter of interest determines the conditional mean of the dependent variable conditional on the regressors.

Suppose the information about the underlying distribution or a key feature that we are interested in cannot be summarized by a finite-dimensional parameter. In this case, we must directly estimate the distribution or this feature of interest, and it is here that non-parametric methods come in handy. Kernel density estimation furnishes a method to learn about the underlying distribution of the data without having to impose any parametric assumptions; here, the entire density can be viewed as an infinite-dimensional parameter of interest. Likewise, kernel regression methods allow us to recover the conditional mean of the dependent variable without any linearity assumptions.

In what follows, we introduce methods of non-parametrically estimating the entire distribution of a random variable with a focus on the Glivenko-Cantelli theorem, which provides a robust theoretical basis for non-parametrically estimating the distribution function of a real random vector. Afterward, we shift our focus to densities of continuous random variables, and investigate methods of non-parametrically estimating the densities and, in a regression context, the conditional mean function.

8.1 Non-parametric Estimation of the Distribution Function

8.1.1 Concentration Inequalities

Before delving into the theorem of interest, we first touch on a set of useful inequalities called concentration inequalities. These, like Markov's inequality, often bound the tail behavior of various

kinds of random variables and prove useful when deriving bounds for the rate of convergence.

Perhaps the simplest result in this class of concentration inequalities is the Chernoff bound. Let X be a real valued random variable, so that, for any $t \in \mathbb{R}$, the expectation $\mathbb{E}[\exp(tX)]$ exists in $[0, +\infty]$; recall that the function $M_X : \mathbb{R} \to [0, +\infty]$ defined as $M_X(t) = \mathbb{E}[\exp(tX)]$ for any $t \in \mathbb{R}$ is the moment generating function of X. It follows that, for any $a \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}(X \geq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq e^{-ta}\mathbb{E}\left[e^{tX}\right] = e^{-ta}M_X(t),$$

where the first equality follows from the fact that $t > 0$ and the inequality follows from Markov's inequality. This holds for any $t > 0$, so it follows that

$$\mathbb{P}(X \geq a) \leq \inf_{t \in (0,+\infty)} e^{-ta}M(t).$$

The quantity on the right, which can also be written as

$$\inf_{t \in (0,+\infty)} \mathbb{E}\left[e^{t(X-a)}\right],$$

is called the Chernoff bound.

We can similarly bound the left tail of X by noting that

$$\mathbb{P}(X \leq a) = \mathbb{P}\left(e^{tX} \geq e^{ta}\right) \leq e^{-ta}\mathbb{E}\left[e^{tX}\right] = e^{-ta}M_X(t)$$

for any $t < 0$, so that

$$\mathbb{P}(X \leq a) \leq \inf_{t \in (-\infty,0)} e^{-ta}M_X(t).$$

The Chernoff bound is simple but, depending on the underlying distribution of X, the MGF might be equal to $+\infty$, at which point the bound is trivial. Even when the MGF exists in $(0, +\infty)$, there are easily cases where the bound is trivial. For instance, if X has bounded first moments and $\mathbb{E}[X] \geq a$, then by Jensen's inequality

$$\mathbb{E}\left[e^{tX}\right] \geq e^{t\mathbb{E}[X]} \geq e^{ta}$$

for any $t > 0$, so that the Chernoff bound in this case is larger than 1. Therefore, in many cases we require tighter bounds on the tail behavior of the random variables of interest. In light of the Chernoff bound, we actually only require tighter bounds on the MGF of our random variables of interest, which is where Hoeffding's lemma enters the picture.

Lemma 8.1 (Hoeffding's Lemma)
Let X be a real random variable and $\mathcal{F}$ a sub σ-algebra of $\mathcal{H}$ such that $\mathbb{E}[X \mid \mathcal{F}] = 0$. Suppose A, B are $\mathcal{F}$-measurable real random variables such that $A \leq X \leq B$ and $A < B$ almost surely.

Then, $A \leq 0 \leq B$ almost surely and, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda X) \mid \mathcal{F}] \leq \exp\left(\frac{\lambda^2 (B-A)^2}{8}\right).$$

Proof) Define $H = \{A > 0\}$. Then, by the $\mathcal{F}$-measurability of A and the defining property of conditional expectations,

$$\mathbb{E}[X \cdot I_H] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] I_H] = 0.$$

Suppose $\mathbb{P}(H) > 0$; since $A \leq X$ almost surely, $\mathbb{P}(H \cap \{A \leq X\}) = \mathbb{P}(H) > 0$ and therefore $X \cdot I_{H \cap \{A \leq X\}} \geq 0$ everywhere on the sample space. Meanwhile,

$$\mathbb{E}\left[X \cdot I_{H \cap \{A \leq X_0\}}\right] = \mathbb{E}[X \cdot I_H] = 0.$$

By the vanishing property for non-negative functions, it follows that $X \cdot I_{H \cap \{A \leq X_0\}} = 0$ almost surely. This contradicts the fact that $X = X \cdot I_{H \cap \{A \cap X_0\}}(\omega) > 0$ for any outcome ω in the event $H \cap \{A \cap X_0\}$ of non-zero probability, so it must be the case that $\mathbb{P}(H) = 0$, that is, $A \leq 0$ almost surely.

A symmetric argument for B shows us that $B \geq 0$ almost surely, so that $A \leq 0 \leq B$ almost surely.

Now choose any $\lambda \in \mathbb{R}$. Since $A \leq X \leq B$ and

$$X = \frac{B-X}{B-A} A + \frac{X-A}{B-A} B,$$

the convexity of the exponential function tells us that

$$\exp(\lambda X) \leq \frac{B-X}{B-A} \exp(\lambda A) + \frac{X-A}{B-A} \exp(\lambda B).$$

It follows that

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X) \mid \mathcal{F}] &\leq \frac{B - \mathbb{E}[X \mid \mathcal{F}]}{B-A} \exp(\lambda A) + \frac{\mathbb{E}[X \mid \mathcal{F}] - A}{B-A} \exp(\lambda B) \\
&= \frac{B}{B-A} \exp(\lambda A) - \frac{A}{B-A} \exp(\lambda B) \\
&= e^{\lambda A} \times \frac{B - A e^{\lambda (B-A)}}{B-A} \\
&= \exp\left(\frac{\lambda (B-A) A}{B-A} + \log\left(\frac{B - A e^{\lambda (B-A)}}{B-A}\right)\right).
\end{aligned}$$

Choose any constants $a < b$ such that $a \leq 0 \leq b$, and define the function $L : \mathbb{R} \to \mathbb{R}$ as

$$L(h) = \frac{ha}{b-a} + \log\left(\frac{b-ae^h}{b-a}\right)$$

for any $h \in \mathbb{R}$, where the function is well-defined because $b - ae^h > 0$ for any $h \in \mathbb{R}$ and $b - a > 0$ due to the fact that $a \leq 0 \leq b$ and $b - a > 0$. Note that

$$L'(h) = \frac{a}{b-a} - \frac{ae^h}{b-ae^h}$$

$$L''(h) = -\frac{abe^h}{(b-ae^h)^2}$$

for any $h \in \mathbb{R}$. A second-order Taylor expansion around 0 shows us that, for any $h > 0$, there exists some $\theta \in [0, 1]$ such that

$$L(h) = L(0) + L'(0)h + \frac{1}{2}h^2 L''(h\theta) = \frac{1}{2}h^2\left(-\frac{ab \cdot e^{h\theta}}{(b-ae^{h\theta})^2}\right)$$

The arithmetic mean-geometric mean inequality shows us that

$$\sqrt{b(-\overline{a}e^{h\theta})} \leq \frac{1}{2}\left(b - \overline{a}e^{h\theta}\right),$$

so that

$$L''(h\theta) = -\frac{abe^{h\theta}}{(b-ae^{h\theta})^2} \leq \frac{1}{4}.$$

It follows that

$$L(h) = \frac{1}{2}h^2 L''(h\theta) \leq \frac{1}{8}h^2.$$

This holds for any $a < b$ such that $a \leq 0 \leq b$, so it follows that

$$\mathbb{E}\left[\exp(\lambda X) \mid \mathcal{F}\right] \leq \exp\left(\frac{\lambda^2(B-A)^2}{8}\right),$$

which is exactly our desired result.

Q.E.D.

Hoeffding's inequality uses a combination of Chernoff bounds and Hoeffding's lemma to obtain a bound for the tail probability of the sample mean of mean zero random variables:

Theorem 8.2 (Hoeffding's Inequality)
Let $\{X_t\}_{t\in\mathbb{N}}$ be a martingale difference sequence with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{N}\}$ on $\mathbb{Z}$. Suppose that there exist $\mathcal{F}$-predictable real valued processes $\{A_t\}_{t\in\mathbb{N}}$ and $\{B_t\}_{t\in\mathbb{N}}$ such that $A_t < B_t$ and $A_t \leq X_t \leq B_t$ almost surely for any $t \in N_+$. Finally, assume that there exists a sequence $\{c_t\}_{t\in\mathbb{N}}$ of non-negative constants such that $B_t - A_t \leq c_t$ for any $t \in N_+$.

Then, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \geq \epsilon\right) \leq \exp\left(-\frac{2T^2\epsilon^2}{\sum_{t=1}^{T} c_t^2}\right)$$

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \leq \epsilon\right) \leq \exp\left(-\frac{2T^2\epsilon^2}{\sum_{t=1}^{T} c_t^2}\right).$$

Proof) We first bound the MGF of $\sum_{t=1}^{T} X_t$ above using Hoeffding's lemma. For any $\lambda > 0$, note that

$$\mathbb{E}\left[\exp(\lambda X_1)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp(\lambda X_1) \mid \mathcal{F}_0\right]\right].$$

Since $A_1 \leq X_1 \leq B_1$ almost surely, where A_1, B_1 are $\mathcal{F}_0$-measurable by the assumption of predictability, and $\mathbb{E}[X_1 \mid \mathcal{F}_0] = 0$ by the definition of an MDS, Hoeffding's lemma tells us that

$$\mathbb{E}\left[\exp(\lambda X_1) \mid \mathcal{F}_0\right] \leq \exp\left(\frac{\lambda^2(B_1 - A_1)^2}{8}\right) \leq \exp\left(\frac{\lambda^2 c_1^2}{8}\right),$$

where the final inequality follows from the assumption that $B_1 - A_1 \leq c_1$.

Now suppose that, for some $T \geq 1$, we have the inequality

$$\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T} X_t\right)\right] \leq \exp\left(\frac{\lambda^2\sum_{t=1}^{T} c_t^2}{8}\right).$$

Note that

$$\mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T+1} X_t\right)\right] = \mathbb{E}\left[\exp\left(\lambda\sum_{t=1}^{T} X_t\right)\cdot\mathbb{E}\left[\exp(\lambda X_{T+1}) \mid \mathcal{F}_T\right]\right]$$

because $X_1, \cdots, X_T$ are $\mathcal{F}_T$ adapted. X_{T+1} is a real random variable such that $A_{T+1} \leq X_{T+1} \leq B_{T+1}$, where A_{T+1} and B_{T+1} are $\mathcal{F}_T$-measurable, and $\mathbb{E}[X_{T+1} \mid \mathcal{F}_T] = 0$ by the

definition of an MDS, so it follows from Hoeffding's lemma that

$$\mathbb{E}[\exp(\lambda X_{T+1}) \mid \mathcal{F}_T] \leq \exp\left(\frac{\lambda^2 (B_{T+1} - A_{T+1})^2}{8}\right) \leq \exp\left(\frac{\lambda^2 c_{T+1}^2}{8}\right).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T} X_t\right)\right] &\leq \exp\left(\frac{\lambda^2 c_{T+1}^2}{8}\right) \times \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T} X_t\right)\right] \\ &\leq \exp\left(\frac{\lambda^2 \sum_{t=1}^{T+1} c_t^2}{8}\right),\end{aligned}$$

where the last inequality follows from the inductive hypothesis. It follows from induction that

$$\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T} X_t\right)\right] \leq \exp\left(\frac{\lambda^2 \sum_{t=1}^{T} c_t^2}{8}\right)$$

for any $T \in N_+$.

Now choose any $\epsilon > 0$ and fix $T \in N_+$. We then have the Chernoff bound

$$\begin{aligned}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} X_t \geq \epsilon\right) &= \mathbb{P}\left(\sum_{t=1}^{T} X_t \geq T\epsilon\right) \\ &\leq \inf_{\lambda \in (0,+\infty)} e^{-\lambda T \epsilon} \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{T} X_t\right)\right] \\ &\leq \inf_{\lambda \in (0,+\infty)} \exp\left(\frac{\lambda^2 \sum_{t=1}^{T} c_t^2}{8} - \lambda T \epsilon\right).\end{aligned}$$

Define $g : (0, +\infty) \to \mathbb{R}$ as

$$g(\lambda) = \frac{\lambda^2 \sum_{t=1}^{T} c_t^2}{8} - \lambda T \epsilon$$

for any $\lambda > 0$. Then, because $g''(\lambda) = \frac{1}{4}\sum_{t=1}^{T} c_t^2 > 0$ for any $\lambda > 0$, g is a strictly convex function on $(0, +\infty)$ that achieves its minimum at $\lambda^* > 0$, where λ^* satisfies the first order condition

$$\frac{1}{4}\lambda^* \sum_{t=1}^{T} c_t^2 = T\epsilon.$$

Solving for λ^* shows us that $\lambda^* = \frac{4T\epsilon}{\sum_{t=1}^T c_t^2}$, and therefore

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^T X_t \geq \epsilon\right) \leq \exp\left(\frac{(\lambda^*)^2 \sum_{t=1}^T c_t^2}{8} - \lambda^* T\epsilon\right).$$

$$= \exp\left(2\frac{T^2\epsilon^2}{\sum_{t=1}^T c_t^2} - 4\frac{T^2\epsilon^2}{\sum_{t=1}^T c_t^2}\right) = \exp\left(-\frac{2T^2\epsilon^2}{\sum_{t=1}^T c_t^2}\right).$$

This completes the proof for the right tail. The proof for the left tail follows almost identically by applying the Chernoff bound for the left tail.

Q.E.D.

Hoeffding's inequality can be easily extended to bounded submartingales and supermartingales via the use of Doob's decomposition theorem: we state this result as a corollary.

Corollary to Theorem 8.2 (Azuma-Hoeffding Inequality)
Let $\mathcal{F}$ be a filtration on $\mathbb{N}$ and $\{X_t\}_{t\in\mathbb{N}}$ an $\mathcal{F}$-adapted process. Suppose that there exist $\mathcal{F}$-predictable processes $\{A_t\}_{t\in\mathbb{N}}$ and $\{B_t\}_{t\in\mathbb{N}}$ such that $A_t < B_t$ and $A_t \leq X_t - X_{t-1} \leq B_t$ almost surely for any $t \in N_+$. Furthermore, let $\{c_t\}_{t\in\mathbb{N}}$ be a sequence of non-negative constants such that $B_t - A_t \leq c_t$ for any $t \in N_+$.

For any $\epsilon > 0$, if $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-supermartingale, then

$$\mathbb{P}(X_T - X_0 \leq -\epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^T c_t^2}\right).$$

If, on the other hand, $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-submartingale, then

$$\mathbb{P}(X_T - X_0 \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^T c_t^2}\right).$$

Finally, if $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-martingale, then

$$\mathbb{P}(|X_T - X_0| \geq \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^T c_t^2}\right).$$

Proof) Suppose initially that $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-supermartingale. Then, by Doob's decomposition theorem, there exists an $\mathcal{F}$-martingale $\{M_t\}_{t\in\mathbb{N}}$ and an $\mathcal{F}$-predictable process $\{F_t\}_{t\in\mathbb{N}}$ with decreasing paths such that

$$X_t = X_0 + M_t + F_t$$

for any $t \in \mathbb{N}$. Define $\{m_t\}_{t\in\mathbb{N}}$ as $m_0 = 0$ and $m_t = M_t - M_{t-1}$ for any $t \in N_+$. Because

$\{M_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-martingale, it follows that $\{m_t\}_{t\in\mathbb{N}}$ is an MDS with respect to $\mathcal{F}$. Furthermore, since

$$X_t - X_{t-1} = M_t - M_{t-1} + F_t - F_{t-1},$$

for any $1 \leq t \leq T$, we have

$$A_t - (F_t - F_{t-1}) \leq m_t \leq B_t - (F_t - F_{t-1})$$

for any $t \in N_+$, where both the lower and upper bounds are $\mathcal{F}_{t-1}$-measurable. Note also that

$$B_t - (F_t - F_{t-1}) - (A_t - (F_t - F_{t-1}) = B_t - A_t \leq c_t$$

for any $t \in N_+$. It follows from Hoeffding's inequality that, for any $T \in N_+$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{T} m_t \geq \epsilon\right) = \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} m_t \geq \frac{1}{T}\epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^{T} c_t^2}\right).$$

Since

$$\sum_{t=1}^{T} m_t = M_T - M_0 = X_T - X_0 - (F_T - F_0) \geq X_T - X_0,$$

where we used the fact that $F_T - F_0 \leq 0$ from the fact that $\{F_t\}_{t\in\mathbb{N}}$ has decreasing paths, we can see that

$$\mathbb{P}\left(X_T - X_0 \geq \epsilon\right) \leq \mathbb{P}\left(\sum_{t=1}^{T} m_t \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^{T} c_t^2}\right).$$

Now suppose that $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-submartingale. Then, because $\{-X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-supermartingale, we can conclude that

$$\mathbb{P}\left(X_T - X_0 \leq -\epsilon\right) \leq \mathbb{P}\left(\sum_{t=1}^{T} m_t \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^{T} c_t^2}\right).$$

Finally, if $\{X_t\}_{t\in\mathbb{N}}$ is an $\mathcal{F}$-martingale, then because it is both an $\mathcal{F}$-supermartingale and an $\mathcal{F}$-submartingale, the claim follows from noting that

$$\mathbb{P}\left(|X_T - X_0| \geq \epsilon\right) \leq \mathbb{P}\left(X_T - X_0 \leq -\epsilon\right) + \mathbb{P}\left(X_T - X_0 \geq \epsilon\right).$$

Q.E.D.

8.1.2 A Consistent Estimator for the CDF: The Glivenko-Cantelli Theorem

In this section we derive a non-parametric estimator for the distribution function constructed from iid data and show that it converges uniformly to the true distribution function. This is a fundamental uniform law of large numbers result that forms the basis of empirical process theory, among other branches of probability theory.

The Empirical Distribution Function

We start by defining the empirical distribution function. For any $n \in N_+$, define the function $\hat{F}_n : \mathbb{R}^k \times \mathbb{R}^{nk} \to [0,1]$ as

$$\hat{F}_n(\tau; x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq \tau\}}$$

for any $\tau, x_1, \cdots, x_n \in \mathbb{R}^k$. Note that $\hat{F}_n(\cdot; x^n)$ is a right continuous and increasing function with left limits for any $x^n \in \mathbb{R}^{nk}$, and that we can interpret it as the distribution function of a discrete random variable that takes on the values $x_1, \cdots, x_n$ with equal probability.

Given an iid sample $X^n = (X_1, \cdots, X_n)$ drawn from the same distribution μ on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. μ can be characterized via its distribution function $F : \mathbb{R}^k \to [0,1]$ defined as

$$F(\tau) = \mu\left(\prod_{j=1}^{k} (-\infty, \tau_j]\right)$$

for any $\tau \in \mathbb{R}^k$. We want to argue that the empirical cdf $\hat{F}_n(\tau; X^n)$ is a consistent estimator of the true distribution function F.

First, we can easily see that $\hat{F}_n(\tau; X^n)$ is an unbiased estimator for $F(\tau)$ for each $\tau \in \mathbb{R}^k$:

$$\mathbb{E}\left[\hat{F}_n(\tau; X^n)\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(X_i \leq \tau) = F(\tau).$$

We can go one step further and assert that $\hat{F}_n(\tau; X^n)$ is a consistent estimator for $F(\tau)$. Defining the function $g_\tau : \mathbb{R}^k \to \{0,1\}$ as $g_\tau(x) = I_{\{x \leq \tau\}}$ for any $x \in \mathbb{R}^k$, $\{g_\tau(X_i)\}_{i \in N_+}$ is a sequence of iid real random variables with mean $F(\tau)$ and finite variance $F(\tau)(1 - F(\tau))$. It follows from the usual WLLN for iid data that

$$\hat{F}_n(\tau; X^n) = \frac{1}{n} \sum_{i=1}^{n} g_\tau(X_i) \xrightarrow{p} F(\tau),$$

and in fact, in light of the SLLN, we can see that this convergence takes place almost surely as well:

$$\hat{F}_n(\tau; X^n) \xrightarrow{a.s.} F(\tau).$$

The Glivenko-Cantelli theorem asserts that a stronger result holds, namely that $\hat{F}_n(\tau; X^n)$

converges to $F(\tau)$ uniformly on $\mathbb{R}^k$. In other words, we want to show that

$$\sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right|$$

converges uniformly to 0 in some sense. Note that, for any $x^n\in\mathbb{R}^{nk}$, both $\hat{F}_n(\cdot;x^n)$ and F are bounded functions on $\mathbb{R}^k$; therefore, the above object can be written as

$$\left\|\hat{F}_n(\cdot;X^n)-F\right\|_\infty,$$

where $\|\cdot\|_\infty$ is the supremum norm on the space $\mathcal{B}(\mathbb{R}^k,\mathbb{R})$ of bounded functions on $\mathbb{R}^k$.

Measurability of the Sup Difference

Our first order of business is to prove that $\left\|\hat{F}_n(\cdot;X^n)-F\right\|_\infty$ is indeed a random variable. Showing this is complicated by the fact that the empirical cdf is not continuous on $\mathbb{R}^k$; it is an increasing right continuous function with left limits that has finitely many discontinuity points. Nevertheless, we can show that the supremum above is measurable by demonstrating that the supremum over the uncountable set $\mathbb{R}^k$ can be reduced to the maximum over a finite number of quantities.

We now present two different treatments of this matter: first, we deal with the case $k=1$, where we can explicitly provide a formula for the sup difference as a maximum over finitely many measurable elements. For the general case, a similar result holds but it is far more difficult to provide an analytical formulation, so we instead show that the sup difference is measurable by showing that the supremum can be taken over the countable set $\mathbb{Q}^k$ instead of the uncountable set $\mathbb{R}^k$.

Lemma 8.3 (Measurability in the Case $k=1$)
If $k=1$, then for any sample $X_1,\cdots,X_n$ of real valued random variables we have

$$\sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right|=\max_{0\le i\le n-1}\left(\left|F(X^{(i)}-)-\frac{i}{n}\right|,\left|F(X^{(i)})-\frac{i-1}{n}\right|\right),$$

where $F(x-)=\sup_{t<x}F(t)$ for any $x\in\mathbb{R}$ and $X^{(i)}$ is the ith order statistic of the sample $X_1,\cdots,X_n$. By implication,

$$\sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right|=\sup_{\tau\in\mathbb{Q}}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right|.$$

Proof) First, choose $x_1,\cdots,x_n\in\mathbb{R}$ ordered as $\tilde{x}_1\le\cdots\le\tilde{x}_n$. We can consider the following

partition of the real line:

$$\underbrace{(-\infty, \tilde{x}_1)}_{P_0}, \quad \underbrace{[\tilde{x}_1, \tilde{x}_2)}_{P_1}, \quad \cdots \quad , \underbrace{[\tilde{x}_{n-1}, \tilde{x}_n)}_{P_{n-1}}, \quad \underbrace{[\tilde{x}_n, +\infty)}_{P_n}.$$

For any $1 \leq i \leq n-1$, note that

$$\sup_{\tau \in P_i} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| = \sup_{\tau \in P_i} \left| \frac{i}{n} - F(\tau) \right| = \max \left(\left| F(\tilde{x}_{i+1}-) - \frac{i}{n} \right|, \left| F(\tilde{x}_i) - \frac{i}{n} \right| \right),$$

where the final equality follows from the fact that F is increasing with left limits:

$$F(\tilde{x}_{i+1}-) = \sup_{\tau < \tilde{x}_{i+1}} F(\tau).$$

Meanwhile,

$$\sup_{\tau \in P_0} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| = \sup_{\tau \in P_0} F(\tau) = F(\tilde{x}_1-)$$

and

$$\sup_{\tau \in P_n} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| = \sup_{\tau \in P_0} |1 - F(\tau)| = 1 - F(\tilde{x}_n).$$

Therefore,

$$\begin{aligned} \sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| &= \max_{0 \leq i \leq n} \sup_{\tau \in P_i} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| \\ &= \max \left[\max_{0 \leq i \leq n-1} \left| F(\tilde{x}_{i+1}-) - \frac{i}{n} \right|, \max_{1 \leq i \leq n} \left| F(\tilde{x}_i) - \frac{i}{n} \right| \right]. \end{aligned}$$

This holds for any $x_1, \cdots, x_n$. The first result follows by simply plugging in the sample observations.

This revelation also allows us to reduce the supremum over the uncountable set $\mathbb{R}$ into one over the countable set $\mathbb{Q}$. Given any $x_1, \cdots, x_n \in X$, where we assume without loss of generality that $x_1 < \cdots < x_n$, choose any $\epsilon > 0$. Suppose initially that

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| = \left| F(x_i) - \frac{i}{n} \right|$$

for some $1 \leq i \leq n$. Then, by the right continuity of F and the denseness of $\mathbb{Q}$ in $\mathbb{R}$, there exists some $q \in \mathbb{Q}$ satisfying $x_i \leq q < x_{i+1}$ that is close enough to x_i such that

$$|F(q) - F(x_i)| < \epsilon.$$

It follows that

$$\left|\hat{F}_n(q;x_1,\cdots,x_n)-F(\tau)\right| = \left|F(q)-\frac{i}{n}\right| > \left|F(x_i)-\frac{i}{n}\right| - \frac{\epsilon}{2}$$

$$= \sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| - \epsilon.$$

On the other hand, suppose

$$\sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| = \left|F(x_{i+1}-)-\frac{i}{n}\right|$$

for some $0 \leq i \leq n-1$. Then, by the definition of the left limit, there exists a $q \in \mathbb{Q}$ satisfying $x_i \leq q < x_{i+1}$ that is close enough to x_{i+1} so that

$$|F(q)-F(x_{i+1}-)| < \epsilon.$$

It then follows that

$$\left|\hat{F}_n(q;x_1,\cdots,x_n)-F(q)\right| = \left|F(q)-\frac{i}{n}\right| > \left|F(x_{i+1}-)-\frac{i}{n}\right| - \frac{\epsilon}{2}$$

$$= \sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| - \epsilon.$$

Therefore, in any case there exists a $q \in \mathbb{Q}$ such that

$$\left|\hat{F}_n(q;x_1,\cdots,x_n)-F(q)\right| > \sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| - \epsilon.$$

By implication,

$$\sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| - \epsilon \leq \sup_{q\in\mathbb{Q}}\left|\hat{F}_n(q;x_1,\cdots,x_n)-F(q)\right|.$$

This holds for any $\epsilon > 0$, so

$$\sup_{\tau\in\mathbb{R}}\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-F(\tau)\right| = \sup_{q\in\mathbb{Q}}\left|\hat{F}_n(q;x_1,\cdots,x_n)-F(q)\right|.$$

Q.E.D.

Lemma 8.4 (Measurability in the General Case)
For any sample $X_1, \cdots, X_n$ of k-dimensional random vectors we have

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| = \sup_{\tau \in \mathbb{Q}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right|.$$

Proof) Fix $x_1, \cdots, x_n \in \mathbb{R}^k$. The supremum

$$\sup_{\tau \in \mathbb{Q}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right|$$

is clearly bounded above by

$$\sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right|$$

because $\mathbb{Q}^k$ is a subset of $\mathbb{R}^k$. Now suppose, for the sake of contradiction, that

$$\sup_{\tau \in \mathbb{Q}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| < \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right|.$$

Then, by the definition of suprema, there exists some $\tau^* \in \mathbb{R}^k$ such that

$$\sup_{\tau \in \mathbb{Q}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| < \left| \hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*) \right|.$$

Choose any $\epsilon > 0$.

To proceed, consider a partition of $\mathbb{R}^k$ into a finite number of half open k-cells on which $\hat{F}_n(\cdot; x_1, \cdots, x_n)$ assumes a fixed value. Specifically, we order the jth elements of $x_1, \cdots, x_n$ as

$$\tilde{x}_{1j} \leq \cdots \leq \tilde{x}_{nj},$$

and define $\tilde{x}_{0j} = -\infty$, $\tilde{x}_{n+1,j} = +\infty$. Having ordered the elements like this for $1 \leq j \leq k$, we can form the following partition of $\mathbb{R}^k$:

$$P = \Big\{ \prod_{j=1}^{k} [\tilde{x}_{i_m,j}, \tilde{x}_{i_m+1,j}) \mid 0 \leq i_m \leq n \quad \forall 1 \leq m \leq k \Big\}.$$

P partitions $\mathbb{R}^k$ into at most $(n+1)^k$ distinct subsets of $\mathbb{R}^k$. It follows that $\tau^* \in \mathcal{C}^*$ for a unique non-empty half-open k-cell $\mathcal{C}^*$ in P. Denote this k-cell by

$$\mathcal{C}^* = \prod_{j=1}^{k} [a_j, b_j),$$

where $a_j < b_j$ for $1 \leq j \leq k$ because $\mathcal{C}^*$ is non-empty.

We now consider two distinct cases. First, suppose $\hat{F}_n(\tau^*; x_1, \cdots, x_n) \leq F(\tau^*)$, so that

$$\left|\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*)\right| = F(\tau^*) - \hat{F}_n(\tau^*; x_1, \cdots, x_n).$$

In this case, for any $1 \leq j \leq k$, by the denseness of $\mathbb{Q}$ in $\mathbb{R}$ there exists some $q_j \in \mathbb{Q}$ such that

$$a_j \leq \tau_j^* < q_j < b_j.$$

Defining $q = (q_1, \cdots, q_k) \in \mathbb{Q}^k$, it follows that $q \in \mathcal{C}^*$, and because $\hat{F}_n(\cdot; x_1, \cdots, x_n)$ takes on a single value on each cell in P,

$$\hat{F}_n(\tau^*; x_1, \cdots, x_n) = \hat{F}_n(q; x_1, \cdots, x_n),$$

while the increasing nature of F indicates that $F(\tau^*) \leq F(q)$. It follows that

$$F(\tau^*) - \hat{F}_n(\tau^*; x_1, \cdots, x_n) \leq F(q) - \hat{F}_n(q; x_1, \cdots, x_n) \leq \sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right|$$

$$< \sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right| + \epsilon.$$

On the other hand, suppose $F(\tau^*) < \hat{F}_n(\tau^*; x_1, \cdots, x_n)$, so that

$$\left|\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*)\right| = \hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*).$$

By the right continuity of F, there exists a $q \in \mathbb{Q}^k$ such that $\tau^* \leq q$ and $|F(\tau^*) - F(q)| < \epsilon$, in particular $-F(\tau^*) < -F(q) + \epsilon$. Meanwhile, the fact that $\hat{F}_n(\cdot; x_1, \cdots, x_n)$ is increasing on $\mathbb{R}^k$ implies that $\hat{F}_n(\tau^*; x_1, \cdots, x_n) \leq \hat{F}_n(q; x_1, \cdots, x_n)$. Therefore,

$$\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*) < \hat{F}_n(q; x_1, \cdots, x_n) - F(q) + \epsilon \leq \sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right| + \epsilon.$$

In any case, we have seen that

$$\left|\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*)\right| < \sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right| + \epsilon.$$

This holds for any choice of $\epsilon > 0$, so it must be the case that

$$\left|\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*)\right| \leq \sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right|.$$

which contradicts our choice of τ^* as the value in $\mathbb{R}^k$ such that

$$\sup_{\tau \in \mathbb{Q}^k} \left|\hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau)\right| < \left|\hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*)\right|.$$

Therefore,

$$\sup_{\tau \in \mathbb{Q}^k} \left| \hat{F}_n(\tau; x_1, \cdots, x_n) - F(\tau) \right| < \left| \hat{F}_n(\tau^*; x_1, \cdots, x_n) - F(\tau^*) \right|,$$

which allows us to establish that

$$\left\| \hat{F}_n(\cdot; X^n) - F \right\|_\infty$$

is a random variable.

Q.E.D.

The Glivenko-Cantelli Theorem

Now we are ready to state and prove our theorem of interest. Note that we are able to argue in the theorem that $\left\| \hat{F}_n(\cdot; X^n) - F \right\|_\infty$ converges almost surely to 0, so that it is a type of strong uniform law of large numbers. The proof makes use of a symmetrization trick and Hoeffding's inequality. We first state a lemma.

Lemma 8.5 (Symmetrization Lemma)
Let $\mathbb{T}$ be a countable index set and $\{Z_t\}_{t \in \mathbb{T}}$ and $\{Z'_t\}_{t \in \mathbb{T}}$ independent sequences of bounded real random variables such that $Z_t \sim Z'_t$ for each $t \in \mathbb{T}$. For any $\epsilon > 0$, suppose there exists an $0 < \alpha < \epsilon$, a $\beta \in (0,1)$ and an index $t^* \in \mathbb{T}$ such that

$$\mathbb{P}\left(|Z'_{t*}| < \alpha \right) \geq \beta.$$

Then,

$$\mathbb{P}\left(\sup_{t \in \mathbb{T}} |Z_t| > \epsilon \right) \leq \frac{1}{\beta} \mathbb{P}\left(\sup_{t \in \mathbb{T}} |Z_t - Z'_t| > \epsilon - \alpha \right).$$

Proof) Note first that the quantities

$$\sup_{t \in \mathbb{T}} |Z_t|, \quad \sup_{t \in \mathbb{T}} |Z_t - Z'_t|, \quad \inf_{t \in \mathbb{T}} |Z'_t|$$

are bounded real random variables, where the boundedness follows from the boundedness of the sequences $\{Z_t\}_{t \in \mathbb{T}}$ and $\{Z'_t\}_{t \in \mathbb{T}}$ and the measurability from the fact that $\mathbb{T}$ is countable. Note that the event $\{\inf_{t \in \mathbb{T}} |Z'_t| < \alpha\}$ contains the event $\{|Z'_{t*}| < \alpha\}$;

therefore,

$$\mathbb{P}\left(\inf_{t\in\mathbb{T}}|Z'_t|<\alpha\right)\geq\mathbb{P}\left(|Z'_{t*}|<\alpha\right)\geq\beta.$$

Now note that

$$\begin{aligned}\beta\cdot\mathbb{P}\left(\sup_{t\in\mathbb{T}}|Z_t|>\epsilon\right)&\leq\mathbb{P}\left(\inf_{t\in\mathbb{T}}|Z'_t|<\alpha\right)\times\mathbb{P}\left(\sup_{t\in\mathbb{T}}|Z_t|>\epsilon\right)\\&=\mathbb{P}\left(\{\inf_{t\in\mathbb{T}}|Z'_t|<\alpha\}\cap\{\sup_{t\in\mathbb{T}}|Z_t|>\epsilon\}\right),\end{aligned}$$

where the equality follows from the fact that $\{Z_t\}_{t\in\mathbb{T}}$ and $\{Z'_t\}_{t\in\mathbb{T}}$ are independent. Now choose some $\omega\in\Omega$ such that

$$\inf_{t\in\mathbb{T}}|Z'_t(\omega)|<\alpha\quad\text{and}\quad\sup_{t\in\mathbb{T}}|Z_t(\omega)|>\epsilon.$$

For any $t\in\mathbb{T}$,

$$\begin{aligned}|Z_t(\omega)-Z'_t(\omega)|&\geq|Z_t(\omega)|-|Z'_t(\omega)|\\&\geq|Z_t(\omega)|-\inf_{s\in\mathbb{T}}|Z'_s(\omega)|\\&\geq|Z_t(\omega)|-\alpha.\end{aligned}$$

This holds for any $t\in\mathbb{T}$, so taking suprema on both sides yields

$$\sup_{t\in\mathbb{T}}|Z_t(\omega)-Z'_t(\omega)|\geq\sup_{t\in\mathbb{T}}|Z_t(\omega)|-\alpha>\epsilon-\alpha.$$

We have shown that

$$\{\inf_{t\in\mathbb{T}}|Z'_t|<\alpha\}\cap\{\sup_{t\in\mathbb{T}}|Z_t|>\epsilon\}\leq\{\sup_{t\in\mathbb{T}}|Z_t-Z'_t|>\epsilon-\alpha\},$$

so the monotonicity of measures implies that

$$\beta\cdot\mathbb{P}\left(\sup_{t\in\mathbb{T}}|Z_t|>\epsilon\right)\leq\mathbb{P}\left(\sup_{t\in\mathbb{T}}|Z_t-Z'_t|>\epsilon-\alpha\right).$$

Multiplying both sides by $\frac{1}{\beta}$ completes the proof.

Q.E.D.

Theorem 8.6 (The Glivenko-Cantelli Theorem)
Let $\{X_i\}_{i\in N_+}$ be a sequence of iid random variables taking values in $\mathbb{R}^k$, and $F:\mathbb{R}^k\to[0,1]$ the distribution function of each X_i. Letting $\hat{F}_n:\mathbb{R}^k\times\mathbb{R}^{nk}\to[0,1]$ be the empirical distribution function defined above,

$$\sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right|\overset{a.s.}{\to}0.$$

Proof) We first define the mapping $G_n:\mathbb{R}^{(n+1)k}\to[-1,1]$ as

$$G_n(\tau;x_1,\cdots,x_n)=\frac{1}{n}\sum_{i=1}^{n}\left(I_{\{x_i\le\tau\}}-F(\tau)\right)$$

for any $\tau,x_1,\cdots,x_n\in\mathbb{R}^k$. In light of our reduction of the supremum over $\mathbb{R}^k$ into a supremum over $\mathbb{Q}^k$, we want to show that

$$\sup_{q\in\mathbb{Q}^k}|G_n(q;X^n)|\overset{a.s.}{\to}0.$$

To this end, we first proceed with a symmetrization step. Choose any $\epsilon>0$. Let $\{X_i'\}_{i\in N_+}$ be a sequence of iid random variables with distribution function F that are independent of $\{X_i\}_{i\in N_+}$. By the WLLN, we know that

$$\lim_{n\to\infty}\mathbb{P}\left(|G_n(O_{k\times1};X'^n)|<\frac{\epsilon}{2}\right)=1;$$

therefore, there exists some $N\in N_+$ such that

$$\mathbb{P}\left(|G_n(O_{k\times1};X'^n)|>\frac{\epsilon}{2}\right)>\frac{1}{2}$$

for any $n\ge N$.

Fix any $n\ge N$. Then, $\{G_n(q;X^n)\}_{q\in\mathbb{Q}^k}$ and $\{G_n(q;X'^n)\}_{q\in\mathbb{Q}^k}$ are identically distributed but independent sequences of random variables taking values in $[-1,1]$ such that

$$\mathbb{P}\left(|G_n(O_{k\times1};X'^n)|>\frac{\epsilon}{2}\right)>\frac{1}{2}.$$

From the symmetrization lemma, it then follows that

$$\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}|G_n(q;X^n)|>\epsilon\right)\le2\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}|G_n(q;X^n)-G_n(q;X'^n)|>\frac{\epsilon}{2}\right).$$

Since

$$\left|G_n(q;X^n)-G_n(q;X'^n)\right| = \left|\left(\hat{F}_n(q;X^n)-F(q)\right)-\left(\hat{F}_n(q;X'^n)-F(q)\right)\right| = \left|\hat{F}_n(q;X^n)-\hat{F}_n(q;X'^n)\right|,$$

we can see that

$$\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}|G_n(q;X^n)|>\epsilon\right) \le 2\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}\left|\hat{F}_n(q;X^n)-\hat{F}_n(q;X'^n)\right|>\frac{\epsilon}{2}\right).$$

This concludes the symmetrization step.

Now note that, for any $x_1,\cdots,x_n\in\mathbb{R}^k$ and $x'_1,\cdots,x'_n\in\mathbb{R}^k$,

$$\hat{F}_n(\tau;x_1,\cdots,x_n)-\hat{F}_n(\tau;x'_1,\cdots,x'_n) = \frac{1}{n}\sum_{i=1}^n\left(I_{\{x_i\le\tau\}}-I_{\{x'_i\le\tau\}}\right)$$

for any $\tau\in\mathbb{R}^k$. This tells us that the function

$$\hat{F}_n(\cdot;x_1,\cdots,x_n)-\hat{F}_n(\cdot;x'_1,\cdots,x'_n)$$

on $\mathbb{R}^k$ assumes at most $(2n+1)^k$ distinct values. Furthermore, for each $\tau\in\mathbb{R}^k$, because $\{I_{\{X_i\le\tau\}}-I_{\{X'_i\le\tau\}}\}_{i\in N_+}$ is a collection of iid random variables (so that it is trivially an MDS with respect to the filtration it generates) that takes values in $[-1,1]$, Hoeffding's inequality tells us that

$$\begin{aligned}\mathbb{P}\left(\left|\hat{F}_n(\tau;x_1,\cdots,x_n)-\hat{F}_n(\tau;x'_1,\cdots,x'_n)\right|>\frac{\epsilon}{2}\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\left(I_{\{X_i\le\tau\}}-I_{\{X'_i\le\tau\}}\right)\right|>\frac{\epsilon}{2}\right)\\ &\le \exp\left(-\frac{2n^2\left(\frac{\epsilon}{2}\right)^2}{4n}\right)=\exp\left(-\frac{n\epsilon^2}{8}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}\left|\hat{F}_n(q;X^n)-\hat{F}_n(q;X'^n)\right|>\frac{\epsilon}{2}\right) &\le \mathbb{P}\left(\bigcup_{q\in\mathbb{Q}^k}\left\{\left|\hat{F}_n(q;X^n)-\hat{F}_n(q;X'^n)\right|>\frac{\epsilon}{2}\right\}\right)\\ &\le (2n+1)^k\exp\left(-\frac{n\epsilon^2}{8}\right),\end{aligned}$$

and for any $n\ge N$ we have

$$\begin{aligned}\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}|G_n(q;X^n)|>\epsilon\right) &\le 2\mathbb{P}\left(\sup_{q\in\mathbb{Q}^k}\left|\hat{F}_n(q;X^n)-\hat{F}_n(q;X'^n)\right|>\frac{\epsilon}{2}\right)\\ &\le 2(2n+1)^k\exp\left(-\frac{n\epsilon^2}{8}\right).\end{aligned}$$

Since

$$\frac{(2n+3)^k \exp\left(-\frac{(n+1)\epsilon^2}{8}\right)}{(2n+1)^k \exp\left(-\frac{n\epsilon^2}{8}\right)} = \frac{(2n+3)^k}{(2n+1)^k} \times \exp\left(-\frac{\epsilon^2}{8}\right) \to \exp\left(-\frac{\epsilon^2}{8}\right) < 1$$

as $n \to \infty$, the ratio test for series convergence reveals that

$$\sum_{n=1}^{\infty} (2n+1)^k \exp\left(-\frac{n\epsilon^2}{8}\right) < +\infty.$$

It follows that

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}\left(\sup_{q\in\mathbb{Q}^k} |G_n(q;X^n)| > \epsilon\right) &= \sum_{n=1}^{N} \mathbb{P}\left(\sup_{q\in\mathbb{Q}^k} |G_n(q;X^n)| > \epsilon\right) + \sum_{n=N+1}^{\infty} \mathbb{P}\left(\sup_{q\in\mathbb{Q}^k} |G_n(q;X^n)| > \epsilon\right) \\ &\leq \sum_{n=1}^{N} \mathbb{P}\left(\sup_{q\in\mathbb{Q}^k} |G_n(q;X^n)| > \epsilon\right) + 2\sum_{n=N+1}^{\infty} (2n+1)^k \exp\left(-\frac{n\epsilon^2}{8}\right) < +\infty.\end{aligned}$$

This holds for any $\epsilon > 0$, so by the Borel-Cantelli lemma we can conclude that

$$\sup_{q\in\mathbb{Q}^k} |G_n(q;X^n)| \overset{a.s.}{\to} 0,$$

which is exactly the result we desired.

Q.E.D.

8.1.3 Rate of Convergence and the Kolmogorov-Smirnov Statistic

Having shown that the sup difference between the empirical cdf and the true distribution function goes to 0 under large samples, we will naturally be interested in the rate of convergence of the sup difference. It turns out that the only possible rate at which the sup difference becomes stable is the familiar rate $\frac{1}{\sqrt{n}}$.

Formally, let $X_1, \cdots, X_n$ be iid observations drawn from a distribution with cdf $F : \mathbb{R}^k \to [0,1]$. The bound obtained in the proof of the Glivenko-Cantelli theorem tells us that

$$\begin{aligned}\mathbb{P}\left(n^{\delta} \cdot \sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| > \epsilon\right) &= \mathbb{P}\left(\sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| > \epsilon \cdot n^{-\delta}\right) \\ &\leq 2(2n+1)^k \exp\left(-\frac{n^{1-2\delta}\epsilon^2}{8}\right)\end{aligned}$$

for any $\delta < \frac{1}{2}$ and $\epsilon > 0$. Since $1 - 2\delta > 0$ in this case, taking $n \to \infty$ on both sides reveals that

$$\lim_{n \to \infty} \mathbb{P}\left(n^{\delta} \cdot \sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| > \epsilon\right) = 0,$$

and because this holds for any $\epsilon > 0$,

$$n^{\delta} \cdot \sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| = o_p(1)$$

as $n \to \infty$.

On the other hand, if $\delta > \frac{1}{2}$, then for any $\tau \in \mathbb{R}^k$

$$n^{\delta}\left[\frac{1}{n}\sum_{i=1}^{n}\left(I_{\{X_i \leq \tau\}} - F(\tau)\right)\right] = n^{\delta - \frac{1}{2}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I_{\{X_i \leq \tau\}} - F(\tau)\right)\right] \xrightarrow{p} +\infty,$$

since $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(I_{\{X_i \leq \tau\}} - F(\tau)\right) = O_p(1)$ by the Lindeberg-Levy CLT. Because

$$n^{\delta} \cdot \sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| \geq n^{\delta}\left[\frac{1}{n}\sum_{i=1}^{n}\left(I_{\{X_i \leq t\}} - F(t)\right)\right]$$

for any $t \in \mathbb{R}^k$, it follows that

$$n^{\delta} \cdot \sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right| \xrightarrow{p} +\infty$$

as well. By a process of elimination, the only rate at which $\sup_{\tau \in \mathbb{R}^k} \left|\hat{F}_n(\tau; X^n) - F(\tau)\right|$ can be stable is $\frac{1}{\sqrt{n}}$.

This leads us to the Kolmogorov-Smirnov test statistic for comparing two distributions. Given a target distribution function $F : \mathbb{R}^k \to [0,1]$ and an iid sample $X_1, \cdots, X_n$ drawn from

some unknown distribution described by the cdf $G : \mathbb{R}^k \to [0,1]$, the Kolmogorov-Smirnov test statistic for testing the hypothesis $H_0 : F = G$ against the hypothesis $H_1 : F \neq G$ is given as

$$KS_n(X^n) = \sqrt{n} \cdot \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right|.$$

Note that the null and alternative hypotheses can be formulated as

$$H_0 : \|F - G\|_\infty = 0, \quad H_1 : \|F - G\|_\infty > 0,$$

since $F = G$ on $\mathbb{R}$ if and only if their supremum distance $\|F - G\|_\infty$ is equal to 0. The KS test is performed by rejecting H_0 if the test statistic is greater than some critical value, since intuitively, the fact that $KS_n(X^n)$ is large under large samples indicates that the distribution from which $X_1, \cdots, X_n$ is drawn is likely not equal to F.

The Glivenko-Cantelli theorem reveals that the KS test is consistent under any alternative; this is the content of the next theorem:

Theorem 8.7 (Consistency of the Kolmogorov-Smirnov Test)
Let $\{X_i\}_{i \in N_+}$ be a sequence of iid random variables taking values in $\mathbb{R}^k$, drawn from the distribution function $G : \mathbb{R}^k \to [0,1]$. Let $F : \mathbb{R}^k \to [0,1]$ be a distribution function such that $\|F - G\|_\infty > 0$. Then, letting $\hat{F}_n : \mathbb{R}^k \times \mathbb{R}^{nk} \to [0,1]$ be the empirical distribution function,

$$\lim_{n \to \infty} \mathbb{P}\left(\sqrt{n} \cdot \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| > M \right) = 1$$

for any $M > 0$.

Proof) First we decompose the KS test statistic. For any $\tau \in \mathbb{R}^k$, by the triangle inequality we have

$$\begin{aligned} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| &\geq |F(\tau) - G(\tau)| - \left| \hat{F}_n(\tau; X^n) - G(\tau) \right| \\ &\geq |F(\tau) - G(\tau)| - \sup_{t \in \mathbb{R}^k} \left| \hat{F}_n(t; X^n) - G(t) \right|. \end{aligned}$$

This holds for any $\tau \in \mathbb{R}^k$, so taking suprema on both sides yields

$$\sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| \geq \|F - G\|_\infty - \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - G(\tau) \right|.$$

Now choose any $M > 0$. We then have

$$\begin{aligned} \mathbb{P}\left(\sqrt{n} \cdot \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| \leq M \right) &\leq \mathbb{P}\left(\|F - G\|_\infty - \sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - G(\tau) \right| \leq \frac{M}{\sqrt{n}} \right) \\ &= \mathbb{P}\left(\sup_{\tau \in \mathbb{R}^k} \left| \hat{F}_n(\tau; X^n) - G(\tau) \right| \geq \|F - G\|_\infty - \frac{M}{\sqrt{n}} \right). \end{aligned}$$

Let $N \in N_+$ be chosen so that $\|F-G\|_\infty - \frac{M}{\sqrt{N}} > 0$. Then, for any $n > N$, we have

$$\|F-G\|_\infty - \frac{M}{\sqrt{n}} > \|F-G\|_\infty - \frac{M}{\sqrt{N}},$$

so that

$$\mathbb{P}\left(\sqrt{n}\cdot \sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right| \le M\right) \le \mathbb{P}\left(\sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-G(\tau)\right| > \|F-G\|_\infty - \frac{M}{\sqrt{N}}\right)$$

for any $n > N$. Define $\epsilon(M) = \|F-G\|_\infty - \frac{M}{\sqrt{N}} > 0$. We saw in the proof of the Glivenko-Cantelli theorem that the tail probability of $\sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-G(\tau)\right|$ can be bounded above as

$$\mathbb{P}\left(\sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-G(\tau)\right| > \epsilon\right) \le 2(2n+1)^k \exp\left(-\frac{n\epsilon^2}{8}\right)$$

for any $\epsilon > 0$ using a combination of symmetrization and Hoeffding's inequality. Therefore, for any $n < N$, we have

$$\mathbb{P}\left(\sqrt{n}\cdot \sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right| \le M\right) \le 2(2n+1)^k \exp\left(-\frac{n\cdot\epsilon(M)^2}{8}\right).$$

The right hand side vanishes as $n\to\infty$, so it follows that

$$\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}\cdot \sup_{\tau\in\mathbb{R}^k}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right| \le M\right) = 0.$$

This holds for any $M > 0$, so the proof is complete.

Q.E.D.

Having shown that the test is consistent, it remains to find the critical value for the test given some significance level. Remarkably, when $k = 1$ and the target distribution is continuous, we can find the exact finite sample distribution of $KS_n(X^n)$ and generate critical values through Monte-Carlo simulations. This is due to the distribution-free property of the KS test statistic for $k = 1$ under continuous target distributions, which we now demonstrate.

Let $X_1, \cdots, X_n$ be iid real random variables drawn from a distribution described by a continuous cdf $F : \mathbb{R} \to [0,1]$. This implies that F is strictly increasing and therefore an invertible function from $\mathbb{R}$ onto $F(\mathbb{R}) \supset (0,1)$. For any $\tau \in \mathbb{R}$, we can write

$$\begin{aligned}\left|\hat{F}_n(\tau;X^n)-F(\tau)\right| &= \left|\frac{1}{n}\sum_{i=1}^n\left(I_{\{X_i\le\tau\}}-F(\tau)\right)\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n\left(I_{\{F(X_i)\le F(\tau)\}}-F(\tau)\right)\right| \le \sup_{u\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n\left(I_{\{F(X_i)\le u\}}-u\right)\right|.\end{aligned}$$

This holds for any $\tau \in \mathbb{R}$, so

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| \leq \sup_{u \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{F(X_i) \leq u\}} - u \right) \right|.$$

Suppose that this inequality is strict. Then, there exists some $u \in [0,1]$ such that

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| < \left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{F(X_i) \leq u\}} - u \right) \right|.$$

We can rule out the case $u = 0, 1$ because in that case the right hand side would equal 0. Now, since $u \in (0,1)$, we can choose $\tau^* \in \mathbb{R}$ such that $F(\tau^*) = u$. By implication,

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{F(X_i) \leq u\}} - u \right) \right| &= \left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{X_i \leq F^{-1}(u)\}} - F(\tau^*) \right) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{X_i \leq \tau^*\}} - F(\tau^*) \right) \right| = \left| \hat{F}_n(\tau^*; X^n) - F(\tau^*) \right| \\
&\leq \sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right|,
\end{aligned}$$

which contradicts the inequality above. Therefore,

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| = \sup_{u \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} \left(I_{\{F(X_i) \leq u\}} - u \right) \right|.$$

Here, note that $\{F(X_1), \cdots, F(X_n)\}$ is an iid sample drawn from the uniform distribution on $(0,1)$, and that $F_U(u) = u$ is precisely its distribution function; therefore, the equation above can be written as

$$\sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| = \sup_{u \in \mathbb{R}} \left| \hat{F}_n(u; F(X_1), \cdots, F(X_n)) - F_U(u) \right|$$

and as such that, under the null $F = G$,

$$KS_n(X^n) = \sqrt{n} \cdot \sup_{\tau \in \mathbb{R}} \left| \hat{F}_n(\tau; X^n) - F(\tau) \right| \sim \sqrt{n} \cdot \sup_{u \in \mathbb{R}} \left| \hat{F}_n(u; U^n) - F_U(u) \right|$$

for some iid sample $U_1, \cdots, U_n$ drawn from the uniform distribution on $[0,1]$. From the closed-form expression for the sup difference we derived earlier, we finally have

$$KS_n(X^n) \sim \sqrt{n} \cdot \max_{1 \leq i \leq n} \left(\left| U^{(i)} - \frac{i}{n} \right|, \left| U^{(i)} - \frac{i-1}{n} \right| \right)$$

for iid random variables $U_1, \cdots, U_n \sim Uniform(0,1)$ and their order statistics $U^{(1)}, \cdots, U^{(n)}$ under the null. This is a remarkable result because it shows us that the KS test statistic in itself has a pivotal exact finite sample distribution under the null that does not depend on any ex-post modifications.

8.2 Non-Parametric Estimation of Density Functions

Suppose there is an iid sample $X_1, \cdots, X_n \sim X$ taking values in $\mathbb{R}^k$, where the k-dimensional distribution of X described by a continuous density $f : \mathbb{R}^k \to \mathbb{R}_+$. Kernel density estimation is a technique employed when we want to estimate the values of $f(x_0)$ for each x_0 in the support E of X.

8.2.1 Motivation: Discrete Approximation to $F'(x)$

To motivate the kernel estimation, or smoothing, approach, let $k = 1$ and note that $f(x_0) = F'(x_0)$ for any point x_0 in $\mathbb{R}$, where $F : \mathbb{R} \to [0,1]$ is the distribution function of X. For small enough $h > 0$, this indicates that we can approximate $f(x_0)$ as

$$f(x_0) \approx \frac{F(x_0 + h) - F(x_0 - h)}{2h}.$$

Since we do not know F, we can think of approximating it via the empirical cdf $\hat{F}_n : \mathbb{R}^{n+1} \to [0,1]$ defined as

$$\hat{F}_n(\tau; x_1, \cdots, x_n) = \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq \tau\}}$$

for any $x_1, \cdots, x_n, \tau \in \mathbb{R}$. For large n, the Glivenko-Cantelli theorem tells us that $\hat{F}_n(\cdot; X^n)$ is uniformly close to F, so that $f(x_0)$ can be approximated as

$$\begin{aligned} f(x_0) \approx \frac{\hat{F}_n(x_0 + h; X^n) - \hat{F}_n(x_0 - h; X^n)}{2h} &= \frac{1}{2nh} \sum_{i=1}^{n} \left(I_{\{X_i \leq x_0 + h\}} - I_{\{X_i \leq x_0 - h\}} \right) \\ &= \frac{1}{2nh} \sum_{i=1}^{n} I_{\{-h < X_i - x_0 \leq h\}}. \end{aligned}$$

Defining the function $K : \mathbb{R} \to \mathbb{R}_+$ as

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } x \in (-1, 1] \\ 0 & \text{otherwise} \end{cases},$$

it follows that

$$f(x_0) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x_0}{h} \right).$$

What we are essentially doing to estimate $f(x_0)$ is taking the weighted average of the points in the data around x_0, where the weight assigned to each observation depends on its proximity to x_0. In this example, the weight is determined by the kernel $K(\cdot)$, so that all points in a neighborhood of radius h around x_0 is assigned equal weight $\frac{1}{2h}$ while the points outside that neighborhood are given weight 0.

As such, in general we can think of the following kinds of estimators of $f(x_0)$ for any $x_0 \in \mathbb{R}^k$. Let $K : \mathbb{R}^k \to \mathbb{R}_+$ be a non-negative valued function that is symmetric in the sense that $K(-x) = K(x)$ for any $x \in \mathbb{R}^k$. We call $K(\cdot)$ our kernel function. Then, our kernel estimator $\hat{f}_n(x_0)$ of $f(x_0)$ can be given as

$$\hat{f}_n(x_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^k} K\left(\frac{X_i - x_0}{h}\right)$$

for some choice $h > 0$.

We call this value of h the bandwidth; the larger it is, the wider the neighborhood around x_0 used to construct the estimator of $f(x_0)$. Thus, a large h has the effect of smoothing our estimate of $\hat{f}_n(x_0)$, similarly to how a wide moving average window produces a smoother trend in time series analysis. The trade-off to a smoother estimate of $f(x_0)$ is that, by using observations that are far from x_0 to inform our estimate of the density at x_0, we obtain biased estimates of $f(x_0)$. To again use an analogy from time series, this is similar to how estimates of the long run variance, or the spectral density at frequency 0, become increasingly biased as more periodogram ordinates, that is, approximations to the spectral density at non-zero frequencies, are used in its construction. This bias-variance tradeoff is central to kernel density estimation, and as we see below, it is key to understanding the asymptotic properties of the kernel density estimator.

8.2.2 Consistency and Asymptotic Normality of Kernel Density Estimators

In order for the kernel density estimator of $f(x_0)$ to be consistent and asymptotically normal, we must impose some smoothness conditions on the density f, as well as some normalizations to the kernel $K(\cdot)$. Formally, we make the following assumptions:

A1 Smoothness Properties of the Density

We assume that $f : \mathbb{R}^k \to \mathbb{R}_+$ is twice continuously differentiable, and that there exists some $M > 0$ such that

$$\left|\frac{\partial f(x)}{\partial x}\right|, \left\|\frac{\partial^2 f(x)}{\partial x \partial x'}\right\| \leq M$$

for Lebesgue-almost every $x \in \mathbb{R}^k$.

A2 Properties of the Kernel

We assume that $K : \mathbb{R}^k \to \mathbb{R}_+$ is a bounded function satisfying the following conditions:

$$K(x_j, x_{-j}) = K(-x_j, x_{-j}) \quad \text{for any } 1 \leq j \leq k, \quad x_j \in \mathbb{R}, \quad x_{-j} \in \mathbb{R}^{k-1}$$

$$\int_{\mathbb{R}^k} K(x)dx = 1, \quad \int_{\mathbb{R}^k} |x|^2 K(x)dx < +\infty.$$

In other words, $K(\cdot)$ is an element-wise even function that integrates to 1 (so that the

values of $K(\cdot)$ can be properly interpreted as weights). Furthermore, $K(\cdot)$ has thin tails, since the nth term test for integrals implies that

$$\lim_{|x|\to+\infty} |x|^2 K(x) = 0.$$

We let the bound of $K(\cdot)$ be $c > 0$, so that $|K(x)| \leq c$ for any $x \in \mathbb{R}^k$.

A3 Rate of Convergence of the Bandwidth

Given a sample of size n, we use the bandwidth $h_n > 0$ to compute the kernel density estimator, where

$$h_n \to 0, \quad nh_n^k \to \infty \quad \text{as } n \to \infty.$$

Under these assumptions, we can show that the kernel density estimator is consistent for the true density $f(x_0)$ and asymptotically normal. First, we present a relevant lemma, which will prove useful not only for the exposition on kernel density exposition but also for kernel regressions later on:

Lemma 8.8 (Rates of Convergence of the Bias and Variance)
Suppose $K : \mathbb{R}^k \to \mathbb{R}_+$ is a kernel function that satisfies the conditions of assumption A2, and let $g : \mathbb{R}^k \to \mathbb{R}$ be a function that is twice continuously differentiable with first and second derivatives that are bounded by some constant $M > 0$. Then, for any $x_0 \in \mathbb{R}^k$ and $h > 0$, there exist constants $c_1(g, K), c_2(g, K) > 0$ such that

$$\left| \int_{\mathbb{R}^k} K(z) g(x_0 + hz) dz - g(x_0) \right| \leq c_1(g, K) h^2$$

$$\left| \int_{\mathbb{R}^k} K(z)^2 g(x_0 + hz) dz - g(x_0) \int_{\mathbb{R}^k} K(z)^2 dz \right| \leq c_2(g, K) h.$$

Proof) **Step 1: Taylor Expansions of g**

To avoid technicalities concerning measurability and integrability, we phrase the Taylor expansions as inequalities rather than exact expansions. The proof relies on two Taylor expansions of g around x_0. For any $z \in \mathbb{R}^k$, note that

$$|g(x_0 + hz) - g(x_0)| \leq hM|z|$$

and

$$\left| g(x_0 + hz) - g(x_0) - h \cdot z' \frac{\partial g(x_0)}{\partial x} \right| \leq \frac{h^2}{2} |z|^2 M,$$

where we used the fact that all derivatives of g are bounded above by $M > 0$.

We touch for a moment on an integrability issue involving the kernel $K(\cdot)$. Define v as the probability measure on $\mathbb{R}^k$ defined as

$$v(A) = \int_A K(z)dz$$

for any $A \in \mathcal{B}(\mathbb{R}^k)$. Then, by Hölder's inequality,

$$\int_{\mathbb{R}^k} |z|K(z)dz = \int_{\mathbb{R}^k} |z|dv(z) \leq \left(\int_{\mathbb{R}^k} |z|^2 dv(z)\right)^{\frac{1}{2}} \left(\int_{\mathbb{R}^k} |z|^2 K(z)dz\right)^{\frac{1}{2}}.$$

The right hand side is finite, so it follows that $\int_{\mathbb{R}^k} |z|K(z)dz < +\infty$ as well, showing us that each

$$z_j \mapsto z_j \cdot K(z_j, z_{-j})$$

is a Borel measurable and integrable mapping on $\mathbb{R}^k$ for any $1 \leq j \leq k$ and $z_{-j} \in \mathbb{R}^{k-1}$.

Step 2: Convolution of K and g

Using the second order expansion above, we can see that

$$\begin{aligned}
&\left|\int_{\mathbb{R}^k} K(z)g(x_0 + hz)dz - g(x_0)\int_{\mathbb{R}^k} K(z)dz - h \cdot \left(\int_{\mathbb{R}^k} z \cdot K(z)dz\right)' \frac{\partial g(x_0)}{\partial x}\right| \\
&\qquad \leq \int_{\mathbb{R}^k} K(z)\left|g(x_0 + hz) - g(x_0) - h \cdot z' \frac{\partial g(x_0)}{\partial x}\right| dz \\
&\qquad \leq \frac{h^2}{2}\left(M\int_{\mathbb{R}^k} |z|^2 K(z)dz\right).
\end{aligned}$$

By assumption, $\int_{\mathbb{R}^k} K(z)dz = 1$. Furthermore, because K is an element-wise even function on $\mathbb{R}^k$, for any $1 \leq j \leq k$ and $z_{-j} \in \mathbb{R}^{k-1}$, the mapping $z_j \mapsto z_j \cdot K(z_j, z_{-j})$ is an odd function. Therefore,

$$\int_{\mathbb{R}^k} z_j \cdot K(z)dz = \int_{\mathbb{R}^{k-1}} \int_{\mathbb{R}} z_j \cdot K(z_j, z_{-j})dz_j dz_{-j} = 0,$$

and it follows that

$$\left|\int_{\mathbb{R}^k} K(z)g(x_0 + hz)dz - g(x_0)\right| \leq \frac{h^2}{2}\left(M\int_{\mathbb{R}^k} |z|^2 K(z)dz\right).$$

By assumption, $\int_{\mathbb{R}^k} |z|^2 K(z)dz < +\infty$, so defining

$$c_1(g,K) = \frac{M}{2}\left(\int_{\mathbb{R}^k} |z|^2 K(z)dz\right) < +\infty,$$

we have our desired result

$$\left|\int_{\mathbb{R}^k} K(z)g(x_0+hz)dz - g(x_0)\right| \leq c_1(g,K)h^2.$$

Step 3: Convolution of K^2 and g

Using the first-order expansion this time around, we have

$$\left|\int_{\mathbb{R}^k} K(z)^2 g(x_0+hz)dz - g(x_0)\int_{\mathbb{R}^k} K(z)^2 dz\right|$$

$$\leq \int_{\mathbb{R}^k} K(z)^2 |g(x_0+hz)-g(x_0)|dz \leq h\left(Mc\int_{\mathbb{R}^k} |z|K(z)dz\right).$$

Defining $c_2(f,K) = Mc\int_{\mathbb{R}^k} |z|K(z)dz < +\infty$, it follows that

$$\left|\int_{\mathbb{R}^k} K(z)^2 g(x_0+hz)dz - g(x_0)\int_{\mathbb{R}^k} K(z)^2 dz\right| \leq c_2(f,K)h,$$

as we desired.

Q.E.D.

Theorem 8.9 (Rates of Convergence of the Bias and Variance)
Let $\{X_i\}_{i\in N_+}$ be an iid sequence of k-dimensional real random vectors drawn from a distribution with density $f : \mathbb{R}^k \to \mathbb{R}_+$. Suppose assumptions A1 and A2 hold true, and define the n-sample kernel density estimator at the point $x_0 \in \mathbb{R}^k$ using the bandwidth $h > 0$ as

$$\hat{f}_n(x_0; X^n, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^k} K\left(\frac{X_i - x_0}{h}\right).$$

Then, there exist $c_1, c_2 > 0$ such that

$$\left|\mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] - f(x_0)\right| \leq c_1 h^2$$

and

$$\left|\mathrm{Var}\left[\hat{f}_n(x_0; X^n, h)\right] - \frac{1}{nh^k} f(x_0)\int_{\mathbb{R}^k} K(z)^2 dz\right| \leq \frac{c_2 h^{1-k} + (f(x_0)+c_1h^2)^2}{n}.$$

Proof) Since the density f and the kernel function $K(\cdot)$ satisfy the conditions of lemma 8.8, there exist constants $c_1, c_2 > 0$ depending on f and $K(\cdot)$ such that, for any $x_0 \in \mathbb{R}^k$,

$$\left| \int_{\mathbb{R}^k} K(z) f(x_0 + hz) dz - f(x_0) \right| \leq c_1 h^2$$

$$\left| \int_{\mathbb{R}^k} K(z)^2 f(x_0 + hz) dz - f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz \right| \leq c_2 h.$$

We now study the bias and variance terms in turn to obtain the bounds of interest.

Starting with the bias term, the iid nature of the collection $K\left(\frac{X_1 - x_0}{h}\right), \cdots, K\left(\frac{X_n - x_0}{h}\right)$ implies that

$$\mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] = \frac{1}{h^k} \mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)\right] = \frac{1}{h^k} \int_{\mathbb{R}^k} K\left(\frac{x - x_0}{h}\right) f(x) dx.$$

By a simple linear change of variables, and noting that $|\det(h \cdot I_k)| = h^k$, we have

$$\mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] = \int_{\mathbb{R}^k} K(z) f(x_0 + hz) dz,$$

so that we immediately have

$$\left| \mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] - f(x_0) \right| \leq c_1 h^2$$

from lemma 9.1.

Moving onto the variance term, the iid-ness of $K\left(\frac{X_1 - x_0}{h}\right), \cdots, K\left(\frac{X_n - x_0}{h}\right)$ once again implies that

$$\begin{aligned}
\mathrm{Var}\left[\hat{f}_n(x_0; X^n, h)\right] &= \mathrm{Var}\left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^k} K\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \frac{1}{n} \mathrm{Var}\left[\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)\right] \\
&= \frac{1}{n}\left[\mathbb{E}\left|\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)\right|^2 - \left(\mathbb{E}\left[\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)\right]\right)^2\right].
\end{aligned}$$

We saw above that

$$\mathbb{E}\left[\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)\right] = \mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right],$$

so that

$$\left| \mathbb{E}\left[\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)\right] - f(x_0) \right| \leq c_1 h^2,$$

and in particular

$$\left(\mathbb{E}\left[\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)\right]\right)^2 \leq (f(x_0)+c_1h^2)^2.$$

On the other hand, using the same linear change of variables as in the proof for the bias term, we can see that

$$\begin{aligned}\mathbb{E}\left|\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)\right|^2 &= \frac{1}{h^{2k}}\int_{\mathbb{R}^k} K\left(\frac{x-x_0}{h}\right)^2 f(x)dx \\ &= \frac{1}{h^k}\int_{\mathbb{R}^k} K(z)^2 f(x_0+hz)dz.\end{aligned}$$

As such, the result from lemma 8.8 implies that

$$\left|\mathbb{E}\left|\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)\right|^2 - \frac{1}{h^k}f(x_0)\int_{\mathbb{R}^k} K(z)^2dz\right| \leq c_2h^{1-k},$$

and therefore

$$\begin{aligned}&\left|\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h)\right] - \frac{1}{nh^k}f(x_0)\int_{\mathbb{R}^k} K(z)^2dz\right| \\ &= \left|\frac{1}{n}\left(\mathbb{E}\left|\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)\right|^2 - \frac{1}{h^k}f(x_0)\int_{\mathbb{R}^k} K(z)^2\right) - \frac{1}{n}\left(\mathbb{E}\left[\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)\right]\right)^2\right| \\ &\leq \frac{c_2h^{1-k}}{n} + \frac{(f(x_0)+c_1h^2)^2}{n}.\end{aligned}$$

This completes the proof.

Q.E.D.

The result above shows us that the bias is increasing in the bandwidth h, while the variance is decreasing in h. This corresponds to our intuition that averaging over a wider neighborhood of each point would smooth the kernel density estimator, that is, it would make the estimator less volatile, but would push it farther away from the true density at each point. This trade-off between the bias and variance when it comes to the size of h actually allows us to pin down the asymptotically optimal bandwidth; this task will be taken up in later sections. For now, we show that the preceding lemma allows us to almost immediately establish the consistency of the kernel density estimator.

Theorem 8.10 (Consistency of Kernel Density Estimator)
Let $\{X_i\}_{i\in N_+}$ be an iid sequence of k-dimensional real random vectors drawn from a distribution with density $f:\mathbb{R}^k \to \mathbb{R}_+$. Suppose assumptions A1 to A3 hold true, and define the n-sample

kernel density estimator at the point $x_0 \in \mathbb{R}^k$ using the bandwidth h_n as

$$\hat{f}_n(x_0; X^n, h_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n^k} K\left(\frac{X_i - x_0}{h_n}\right).$$

Then, the kernel density estimator is consistent, that is,

$$\hat{f}_n(x_0; X^n, h_n) \xrightarrow{p} f(x_0)$$

for any $x_0 \in \mathbb{R}^k$.

Proof) We saw above that there exist $c_1, c_2 > 0$ such that

$$\left|\mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0)\right| \leq c_1 h_n^2$$

$$\left|\mathrm{Var}\left[\hat{f}_n(x_0; X^n, h_n)\right] - \frac{1}{nh_n^k} f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right| \leq \frac{c_2 h_n^{1-k} + (f(x_0) + c_1 h_n^2)^2}{n}.$$

for any $n \in N_+$. Note that

$$\left|\hat{f}_n(x_0; X^n, h_n) - f(x_0)\right| \leq \left|\mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0)\right| + \left|\hat{f}_n(x_0; X^n, h_n) - \mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right]\right|,$$

so that

$$\mathbb{E}\left|\hat{f}_n(x_0; X^n, h_n) - f(x_0)\right|^2 \leq 2\left[\left|\mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0)\right|^2 + \mathbb{E}\left|\hat{f}_n(x_0; X^n, h_n) - \mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right]\right|^2\right]$$

$$= 2\left[\left|\mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0)\right|^2 + \mathrm{Var}\left[\hat{f}_n(x_0; X^n, h_n)\right]\right].$$

The first term goes to 0 as $n \to \infty$ because $c_1 h_n^2 \to 0$ as $n \to \infty$. Meanwhile,

$$\mathrm{Var}\left[\hat{f}_n(x_0; X^n, h_n)\right] \leq \frac{c_2 h_n + (f(x_0) + c_1 h_n^2)^2 h_n^k}{nh_n^k} + \frac{1}{nh_n^k} f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz,$$

and since $h_n \to 0$ and $nh_n^k \to \infty$ as $n \to \infty$, taking $n \to \infty$ on both sides shows us that

$$\lim_{n\to\infty} \mathrm{Var}\left[\hat{f}_n(x_0; X^n, h_n)\right] = 0.$$

Therefore,

$$\lim_{n\to\infty} \mathbb{E}\left|\hat{f}_n(x_0; X^n, h_n) - f(x_0)\right|^2 = 0$$

and we have

$$\hat{f}_n(x_0; X^n, h_n) \xrightarrow{L^2} f(x_0),$$

which establishes the consistency result we desired.

Q.E.D.

The asymptotic normality of the kernel density estimator can also be easily established. However, it turns out that the kernel density estimator is not asymptotically unbiased, since each $\hat{f}_n(x_0; X^n, h_n)$ is centered not around the true density $f(x_0)$ but the true density plus a bias term $b_n(x_0; h_n)$ defined as

$$b_n(x_0; h_n) = \mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0).$$

This result is stated and proved below:

Theorem 8.11 (Asymptotic Normality of the Kernel Density Estimator) Let $\{X_i\}_{i \in N_+}$ be an iid sequence of k-dimensional real random vectors drawn from a distribution with density $f : \mathbb{R}^k \to \mathbb{R}_+$. Suppose assumptions A1 to A3 hold true, and define the n-sample kernel density estimator at the point $x_0 \in \mathbb{R}^k$ under the bandwidth h_n as

$$\hat{f}_n(x_0; X^n, h_n) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_n^k} K\left(\frac{X_i - x_0}{h_n}\right).$$

Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$, the kernel density estimator has the following asymptotic distribution:

$$\sqrt{nh_n^k}\left[\hat{f}_n(x_0; X^n, h_n) - (f(x_0) + b_n(x_0; h_n))\right] \xrightarrow{d} \mathcal{N}\left[0,\ f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right]$$

as $n \to \infty$, where

$$b_n(x_0; h_n) = \mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right] - f(x_0).$$

Proof) The proof is a straightforward application of the Lindeberg-Feller CLT for triangular arrays of independent random variables. Define $\{Z_{n,i}\}_{1 \le i \le n, n \in N_+}$ as

$$Z_{n,i} = \frac{1}{\sqrt{nh_n^k}} \left[K\left(\frac{X_i - x_0}{h_n}\right) - \mathbb{E}\left[K\left(\frac{X_i - x_0}{h_n}\right)\right]\right]$$

for any $n \in N_+$ and $1 \le i \le n$. Note that $\{Z_{n,i}\}_{1 \le i \le n, n \in N_+}$ is a triangular array of independent real random variables with common mean 0. To apply the Lindeberg-Feller CLT, we need to show that:

i) $\sum_{i=1}^{n} \mathrm{Var}[Z_{n,i}]$ converges to some positive constant.

ii) The Lindeberg condition holds, that is,

$$\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\left[|Z_{n,i}|^2\cdot I_{|Z_{n,i}|>\epsilon}\right]=0$$

for any $\epsilon>0$.

We show that each condition is satisfied one by one. First, define

$$s_n^2=\sum_{i=1}^{n}\mathrm{Var}\left[Z_{n,i}\right]=\frac{1}{h_n^k}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h_n}\right)\right].$$

Since

$$\frac{1}{h_n^k}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h_n}\right)\right]=nh_n^k\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h_n)\right],$$

and we know from lemma 8.8 that there exist $c_1,c_2>0$ such that

$$\left|\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h_n)\right]-\frac{1}{nh_n}f(x_0)\int_{\mathbb{R}^k}K(z)^2dz\right|\leq\frac{c_2h_n^{1-k}+(f(x_0)+c_1h_n^2)^2}{n}$$

for any $n\in N_+$, it follows that

$$\begin{aligned}&\left|\frac{1}{h_n^k}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h_n}\right)\right]-f(x_0)\int_{\mathbb{R}^k}K(z)^2dz\right|\\&\qquad=nh_n^k\left|\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h_n)\right]-\frac{1}{nh_n^k}f(x_0)\int_{\mathbb{R}^k}K(z)^2dz\right|\\&\qquad\leq h_n(c_2+(f(x_0)+c_1h_n^2)^2h_n^{k-1})\end{aligned}$$

for any $n\in N_+$. Since $h_n\to0$ as $n\to\infty$, this shows us that

$$\lim_{n\to\infty}\frac{1}{h_n}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h_n}\right)\right]=f(x_0)\int_{\mathbb{R}^k}K(z)^2dz,$$

and as such that

$$\lim_{n\to\infty}s_n^2=f(x_0)\int_{\mathbb{R}^k}K(z)^2dz>0.$$

It remains to verify the Lindeberg condition. Fix $\epsilon>0$, and note that

$$\mathbb{P}\left(|Z_{n,i}|>\epsilon\right)\leq\frac{1}{\epsilon^2}\mathrm{Var}\left[Z_{n,i}\right]=\frac{1}{\epsilon^2n}s_n^2$$

by Chebyshev's inequality. Since the function $K(\cdot)$ is bounded above by $c>0$, we can

see that $|Z_{n,i}| \leq \frac{2c}{\sqrt{nh_n^k}}$ for $1 \leq i \leq n$ and therefore

$$\sum_{i=1}^{n} \mathbb{E}\left[|Z_{n,i}|^2 \cdot I_{|Z_{n,i}|>\epsilon}\right] \leq \frac{4c^2}{h_n^k} \cdot \mathbb{P}\left(|Z_{n,1}| > \epsilon\right) \leq \frac{4c^2}{\epsilon^2} \times \frac{1}{nh_n^k} \times s_n^2.$$

We saw above that $s_n^2 = O(1)$ and $nh_n^k \to \infty$ as $n \to \infty$, so

$$\lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}\left[|Z_{n,i}|^2 \cdot I_{|Z_{n,i}|>\epsilon}\right] = 0.$$

Since this holds for any $\epsilon > 0$, the Lindeberg condition holds, and the Lindeberg-Feller CLT implies that

$$\sum_{i=1}^{n} Z_{n,i} \xrightarrow{d} \mathcal{N}\left[0,\ f(x_0)\int_{\mathbb{R}^k} K(z)^2 dz\right].$$

Finally, note that

$$\begin{aligned}
\sqrt{nh_n^k}\left[\hat{f}_n(x_0; X^n, h_n) - (f(x_0) + b_n(x_0; h_n))\right] &= \sqrt{nh_n^k}\left[\hat{f}_n(x_0; X^n, h_n) - \mathbb{E}\left[\hat{f}_n(x_0; X^n, h_n)\right]\right] \\
&= \frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} \left[K\left(\frac{X_i - x_0}{h_n}\right) - \mathbb{E}\left[K\left(\frac{X_i - x_0}{h_n}\right)\right]\right] \\
&= \sum_{i=1}^{n} Z_{n,i}
\end{aligned}$$

for any $n \in N_+$. This completes the proof.

Q.E.D.

8.2.3 Asymptotically Optimal Bandwidth

The kernel density estimator turns out to be very sensitive to the choice of bandwidth h_n. Under additional smoothness assumptions on the density f, it is possible to analytically obtain an expression for the bandwidth h_n that is optimal in an asymptotic integrated mean square error sense. First, we make the following assumption:

A4 Additional Smoothness Assumptions

Assume that the density $f : \mathbb{R}^k \to \mathbb{R}_+$ is four times continuously differentiable, and that there exists some $M > 0$ such that, for any $1 \leq m \leq 4$ and $1 \leq j_1, \cdots, j_m \leq k$,

$$\left|\frac{\partial^m f(x)}{\partial x_{j_1} \cdots \partial x_{j_m}}\right| \leq M$$

for any $x \in \mathbb{R}^k$. Furthermore,

$$\int_{\mathbb{R}^k} |z|^4 K(z)dz < +\infty,$$

which implies, through a construction using Hölder's inequality and the indefinite integral of $K(\cdot)$ with respect to the Lebesgue measure, that $\int_{\mathbb{R}^k} |z|^m K(z)dz < +\infty$ for any $1 \leq m \leq 4$.

Under this additional assumption, we are now able to obtain a sharper characterization of the bias of the kernel density estimator.

Lemma 8.12 (Modified Rate of Convergence of the Bias)
Let $\{X_i\}_{i \in N_+}$ be an iid sequence of k-dimensional real random vectors drawn from a distribution with density $f : \mathbb{R}^k \to \mathbb{R}_+$. Suppose assumptions A1, A2 and A4 hold true, and define the n-sample kernel density estimator at the point $x_0 \in \mathbb{R}^k$ under the bandwidth $h > 0$ as

$$\hat{f}_n(x_0; X^n, h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^k} K\left(\frac{X_i - x_0}{h}\right).$$

Then, there exists some $c_4 > 0$ such that

$$\left| \mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] - f(x_0) - \frac{h^2}{2} \sum_{j=1}^{k} \frac{\partial^2 f(x_0)}{\partial x_j^2} \left(\int_{\mathbb{R}^k} z_j^2 K(z)dz \right) \right| \leq c_4 h^4$$

for any $n \in N_+$.

Proof) We now rely on the following fourth-order Taylor expansion of f around x_0, again posed as an inequality:

$$\left| f(x_0 + hz) - f(x_0) - h \cdot z' \frac{\partial f(x_0)}{\partial x} + \frac{h^2}{2} \cdot z' \frac{\partial^2 f(x_0)}{\partial x \partial x'} z - \frac{h^3}{3!} \sum_{j_1=1}^{k} \cdots \sum_{j_3=1}^{k} \frac{\partial^3 f(x_0)}{\partial x_{j_1} \cdots \partial x_{j_3}} \left(\prod_{i=1}^{3} z_{j_i} \right) \right|$$
$$\leq \frac{h^4}{4!} M k^4 |z|^4$$

for any $z \in \mathbb{R}^k$ and $h > 0$.

As in lemma 8.8, we begin by noting that

$$\mathbb{E}\left[\hat{f}_n(x_0; X^n, h)\right] = \frac{1}{h^k} \mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)\right] = \frac{1}{h^k} \int_{\mathbb{R}^k} K\left(\frac{x - x_0}{h}\right) f(x)dx$$
$$= \int_{\mathbb{R}^k} K(z) f(x_0 + hz)dz,$$

where we have used the familiar linear change of variables for the last equality. Then,

$$\left| \int_{\mathbb{R}^k} K(z) f(x_0 + hz) dz - f(x_0) \int_{\mathbb{R}^k} K(z) dz - h \cdot \sum_{j=1}^{k} \frac{\partial f(x_0)}{\partial x_j} \left(\int_{\mathbb{R}^k} z_j K(z) dz \right) \right.$$

$$\left. - \frac{h^3}{2} \sum_{j_1=1}^{k} \cdots \sum_{j_3=1}^{k} \frac{\partial^3 f(x_0)}{\partial x_{j_1} \cdots \partial x_{j_3}} \left[\int_{\mathbb{R}^k} \left(\prod_{i=1}^{3} z_{j_i} \right) K(z) dz \right] \right|$$

$$\leq \frac{h^4}{4!} \left(M k^4 \int_{\mathbb{R}^k} |z|^4 K(z) dz \right).$$

By assumption, $\int_{\mathbb{R}^k} K(z) dz = 1$. Furthermore, we saw above that

$$\int_{\mathbb{R}^k} z_j K(z) dz = 0$$

for any $1 \leq j \leq k$. This indicates that, for any $1 \leq j_1, j_2 \leq k$, if $j_1 \neq j_2$ then

$$\int_{\mathbb{R}^k} z_{j_1} z_{j_2} K(z) dz = \int_{\mathbb{R}^{k-1}} z_{j_2} \left(\int_{\mathbb{R}} z_{j_1} K(z_{j_1}, z_{-j_1}) dz_{j_1} \right) dz_{-j_1} = 0.$$

Similarly, choose any $1 \leq j_1, j_2, j_3 \leq k$. We now consider the following cases:

Case 1 j_1, j_2, j_3 not all the same

In this case, letting $j_1 \neq j_2, j_3$ without loss of generality, we have

$$\int_{\mathbb{R}^k} \left(\prod_{i=1}^{3} z_{j_i} \right) K(z) dz = \int_{\mathbb{R}^{k-1}} z_{j_2} z_{j_3} \left(\int_{\mathbb{R}} z_{j_1} K(z_{j_1}, z_{-j_1}) dz_{j_1} \right) dz_{-j_1} = 0,$$

since $\int_{\mathbb{R}} z_{j_1} K(z_{j_1}, z_{-j_1}) dz_{j_1} = 0$ for any $z_{-j_1} \in \mathbb{R}^{k-1}$.

Case 2 $j_1 = j_2 = j_3$

In this case, denoting $j = j_1 = j_2 = j_3$, the fact that $K(\cdot)$ is an element-wise even function tells us that the mapping $z_j \mapsto z_j^3 K(z, z_{-j})$ is an integrable odd function for any $z_{-j} \in \mathbb{R}^{k-1}$. Therefore,

$$\int_{\mathbb{R}^k} z_j^3 K(z) dz = \int_{\mathbb{R}^{k-1}} \int_{\mathbb{R}} z_j^3 K(z_j, z_{-j}) dz_j dz_{-j} = 0.$$

In any case,

$$\int_{\mathbb{R}^k} \left(\prod_{i=1}^{3} z_{j_i} \right) K(z) dz = 0,$$

so that we can rewrite the bias term as

$$\left|\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)-\frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\right|\leq\frac{h^4}{4!}\left(Mk^4\int_{\mathbb{R}^k}|z|^4K(z)dz\right).$$

Defining $c_4=\frac{M}{4!}k^4\left(\int_{\mathbb{R}^k}|z|^4K(z)dz\right)<+\infty$, we now have the desired result:

$$\left|\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)-\frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\right|\leq c_4h^4.$$

Q.E.D.

The new formula for the bias term can be written as

$$b_n(x_0;h):=\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)=\frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)+O(h^4).$$

In other words, the bias is not only proportional to the bandwidth h, but is also proportional to the magnitude of the curvature, that is, the size of the second derivatives $\frac{\partial^2 f(x_0)}{\partial x_j^2}$.

The mean squared error at x_0 under the bandwidth $h>0$ can now be defined as

$$\begin{aligned}MSE_n(x_0;h)&=\mathbb{E}\left|\hat{f}_n(x_0;X^n,h)-f(x_0)\right|^2\\&=\mathbb{E}\left[\left(\hat{f}_n(x_0;X^n,h)-\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]\right)+\left(\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)\right)\right]^2\\&=\mathbb{E}\left|\hat{f}_n(x_0;X^n,h)-\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]\right|^2+\left(\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)\right)^2\\&=\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h)\right]+b_n(x_0)^2,\end{aligned}$$

so that it depends on both the variance and the bias of the estimator. The integrated mean squared error is defined as

$$IMSE_n(h)=\int_{\mathbb{R}^k}MSE_n(x_0;h)dx_0,$$

which can be viewed as an equal-weight sum of the mean squared errors at each point on $\mathbb{R}^k$. Note that $MSE_n:\mathbb{R}^k\to\mathbb{R}_+$ is a non-negative measurable function, so that $IMSE_n$ takes values in $[0,+\infty]$.

Recall from lemmas 8.8 and 8.12 that there exist constants $c_1,c_2,c_4>0$ such that

$$\left|\mathrm{Var}\left[\hat{f}_n(x_0;X^n,h)\right]-\frac{1}{nh^k}f(x_0)\int_{\mathbb{R}^k}K(z)^2dz\right|\leq\frac{c_2h^{1-k}+(f(x_0)+c_1h^2)^2}{n}$$

$$\left|\mathbb{E}\left[\hat{f}_n(x_0;X^n,h)\right]-f(x_0)-\frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\right|\leq c_4h^4.$$

Since these differences all vanish when $n\to\infty$, $h\to 0$ and $nh^k\to\infty$, we can define the asymptotically relevant part of the IMSE, called the asymptotic integrated mean squared error (AIMSE), as

$$AIMSE_n(h)=\int_{\mathbb{R}^k}\left[\frac{1}{nh^k}f(x_0)\int_{\mathbb{R}^k}K(z)^2dz+\left(\frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\right)^2\right]dx_0$$

$$=\frac{1}{nh^k}\int_{\mathbb{R}^k}K(z)^2dz+\frac{h^4}{4}\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right).$$

The second term can be written as a quadratic form of a positive semidefinite matrix, so it is non-negative. The first and second derivatives of $AIMSE_n(h)$ with respect to h are given as

$$AIMSE_n'(h)=-\frac{k}{nh^{k+1}}\int_{\mathbb{R}^k}K(z)^2dz+h^3\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right)$$

$$AIMSE_n''(h)=\frac{k(k+1)}{nh^{k+2}}\int_{\mathbb{R}^k}K(z)^2dz+3h^2\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right)>0$$

for any $h>0$, so the mapping $AIMSE_n:(0,+\infty)\to\mathbb{R}$ is strictly convex. This indicates that the bandwidth h_n^* that minimizes $AIMSE_n$ satisfies the first order condition

$$k\frac{1}{n(h_n^*)^{k+1}}\int_{\mathbb{R}^k}K(z)^2dz=(h_n^*)^3\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right),$$

solving which we obtain

$$h_n^*=\left(\frac{k\int_{\mathbb{R}^k}K(z)^2dz}{\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right)}\right)^{\frac{1}{k+4}}\times n^{-1/(k+4)}.$$

This is the optimal bandwidth in the asymptotic integrated mean squared error sense. Defining

$$c^*=\left(\frac{k\int_{\mathbb{R}^k}K(z)^2dz}{\sum_{j=1}^{k}\sum_{m=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\left(\int_{\mathbb{R}^k}z_m^2K(z)dz\right)\left(\int_{\mathbb{R}^k}\frac{\partial^2 f(x_0)}{\partial x_j^2}\frac{\partial^2 f(x_0)}{\partial x_m^2}dx_0\right)}\right)^{\frac{1}{k+4}},$$

the AIMSE under the optimal bandwidth $h_n^*=\frac{c^*}{n^{1/(k+4)}}$ is given as

$$AIMSE_n(h_n^*)=\frac{1}{n^{4/(k+4)}}\frac{(c^*)^4(k+4)}{4k}$$

$$\times \sum_{j=1}^{k} \sum_{m=1}^{k} \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} z_m^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} \frac{\partial^2 f(x_0)}{\partial x_j^2} \frac{\partial^2 f(x_0)}{\partial x_m^2} dx_0 \right).$$

The best rate at which the AIMSE converges to 0 is $n^{-4/(k+4)}$. In other words, the larger the dimension k of the data, the slower the rate of convergence. Intuitively, we would also expect that the larger the dimension of the data, the less informative it is about the joint density. As such, this result is a mathematical illustration of the curse of dimensionality that plagues multivariate analysis.

8.3 Kernel Regression Models

We can easily extend the kernel density estimation framework studied in the previous section to non-parametric regression models. Formally, we consider the following model. Suppose there is an integrable real random variable Y and a k-dimensional real random vector X, and that we are interested in the conditional expectation $\mathbb{E}[Y \mid X]$. By the Doob-Dynkin lemma, there exists some Borel-measurable function $m : \mathbb{R}^k \to \mathbb{R}$ such that

$$\mathbb{E}[Y \mid X] = m(X).$$

Letting u be some error term that is independent of X, we can express the above equation in as the following regression model:

$$Y = m(X) + u.$$

Given an iid sample $\{(Y_i, X_i)\}_{i \in N_+}$ from (Y, X), we want to find the values of the function $m(\cdot)$ on $\mathbb{R}^k$, that is, we want to estimate $m(x_0)$ for each $x_0 \in \mathbb{R}^k$.

8.3.1 Motivating the Kernel Approach

To motivate the kernel approach we will study below, suppose X has full support on $\mathbb{R}^k$ and let $f_X : \mathbb{R}^k \to \mathbb{R}_{++}$ be its density function. Similarly, let $f_{Y,X} : \mathbb{R}^{k+1} \to \mathbb{R}_+$ be the joint density of Y and X. It follows that, for any $x_0 \in \mathbb{R}^k$, the following relationship holds:

$$m(x_0) = \mathbb{E}[Y \mid X = x_0] = \int_{\mathbb{R}} y \frac{f_{Y,X}(y, x_0)}{f_X(x_0)} dy.$$

Thus, we can reduce the problem of estimating $m(x_0)$ to the problem of estimating the densities $f_X(x_0)$ and $f_{Y,X}(y, x_0)$ for any $y \in \mathbb{R}$. We have already studied one method of estimating densities above, so let us apply that method. Specifically, let $K_X : \mathbb{R}^k \to \mathbb{R}_+$ and $K_Y : \mathbb{R} \to \mathbb{R}_+$ be kernel functions of the type used in kernel density estimation, and we propose estimating $f_X(x_0)$ and $f_{Y,X}(y, x_0)$ by

$$\hat{f}_{X,n}(x_0) = \frac{1}{nh^k} \sum_{i=1}^{n} K_X\left(\frac{X_i - x_0}{h}\right)$$

$$\hat{f}_{Y,X,n}(y, x_0) = \frac{1}{nh^{k+1}} \sum_{i=1}^{n} K_X\left(\frac{X_i - x_0}{h}\right) K_Y\left(\frac{Y_i - y}{h}\right)$$

for any $y \in \mathbb{R}$, where $h > 0$ is some bandwidth parameter. Note how we used a kernel for the joint density of Y and X that is separable across the dependent and independent variables. Now, a natural estimator of $m(x_0)$ is

$$\hat{m}_n(x_0) = \int_{\mathbb{R}} y \frac{\hat{f}_{Y,X,n}(y, x_0)}{\hat{f}_{X,n}(x_0)} dy$$

$$= \frac{\sum_{i=1}^n K_X\left(\frac{X_i - x_0}{h}\right)\left[\frac{1}{h}\int_{\mathbb{R}} y \cdot K_Y\left(\frac{Y_i - y}{h}\right) dy\right]}{\sum_{i=1}^n K_X\left(\frac{X_i - x_0}{h}\right)}.$$

By a linear change of variables that we are surely familiar with at this point,

$$\frac{1}{h}\int_{\mathbb{R}} y \cdot K_Y\left(\frac{Y_i - y}{h}\right) dy = \int_{\mathbb{R}} (Y_i + hz) K_Y(-z) dz.$$

If $K_Y(\cdot)$ is an even non-negative function that integrates to 1, as we have continued to assume in the last section, the integral on the right hand side can be simplified to

$$\begin{aligned}\int_{\mathbb{R}} (Y_i + hz) K_Y(-z) dz &= \int_{\mathbb{R}} (Y_i + hz) K_Y(z) dz \\ &= Y_i \int_{\mathbb{R}} K_Y(z) dz + h \cdot \int_{\mathbb{R}} z K_Y(z) dz = Y_i,\end{aligned}$$

since $z \mapsto z K_Y(z)$ is an odd function and $\int_{\mathbb{R}} K_Y(z) dz = 1$. Therefore, our estimator for $m(x_0)$ can be written as

$$\hat{m}_n(x_0) = \frac{\sum_{i=1}^n K_X\left(\frac{X_i - x_0}{h}\right) Y_i}{\sum_{i=1}^n K_X\left(\frac{X_i - x_0}{h}\right)}.$$

This is the Nadaraya-Watson estimator of $m(x_0)$.

Although it follows directly from kernel density estimation like we have just shown, the estimator has a more straightforward and intuitive interpretation. If we define

$$w_n(i; X^n, h) = \frac{K_X\left(\frac{X_i - x_0}{h}\right)}{\sum_{j=1}^n K_X\left(\frac{X_j - x_0}{h}\right)},$$

then the estimator can be written as

$$\hat{m}_n(x_0) = \sum_{i=1}^n w_n(i; X^n, h) Y_i,$$

where $w_n(1; X^n, h), \cdots, w_n(n; X^n, h)$ are non-negative weights that sum to 1. In other words, the Nadaraya-Watson estimator of $m(x_0)$ can be interpreted as a weighted average of the sample observations of the dependent variables, where the weight on the ith observation Y_i depends on how close the corresponding regressors X_i are to the point x_0. This makes complete sense; the first thing anyone who was interested in the conditional expectation of Y given $X = x_0$ would be to take the sample average of the observations of Y whose corresponding regressors are close to x_0.

In the following sections we take a formal look at the Nadaraya-Watson estimator and derive its asymptotic properties, including consistency and asymptotic normality.

8.3.2 Model Assumptions and Preliminary Asymptotic Results

We make the following assumptions:

R1 The DGP and Regressor Exogeneity

Let $\{(Y_i, X_i)\}_{i \in N_+}$ be an iid sample, where each Y_i is a real random variable and X_i a k-dimensional real random vector. We assume that there exists a Borel measurable function $m : \mathbb{R}^k \to \mathbb{R}$ such that

$$\mathbb{E}[Y_i \mid X_i] = m(X_i)$$

holds for each $i \in N_+$.

Defining $u_i = Y_i - m(X_i)$, the following relationship holds:

$$Y_i = m(X_i) + u_i.$$

Mean independence follows immediately: $\mathbb{E}[u_i \mid X_i] = 0$. By the Doob-Dynkin lemma, there exists a Borel-measurable function $\sigma_0 : \mathbb{R}^k \to \mathbb{R}_+$ such that

$$\mathbb{E}\left[u_i^2 \mid X_i\right] = \sigma_0(X_i),$$

so that the errors are potentially heteroskedastic.

R2 Smoothness of the Conditional Mean and Density of Regressors

Let $f : \mathbb{R}^k \to \mathbb{R}_+$ be the density of X_i with respect to the Lebesgue measure. We assume that f, m and σ_0 are twice continuously differentiable on $\mathbb{R}^k$, and that there exists some $M > 0$ such that

$$\left|\frac{\partial g(x)}{\partial x}\right|, \left\|\frac{\partial^2 g(x)}{\partial x \partial x'}\right\| \leq M$$

for any $x \in \mathbb{R}^k$ and $g = m, f, \sigma_0$.

R3 Properties of Kernel Function

We assume that $K : \mathbb{R}^k \to \mathbb{R}_+$ is a bounded function satisfying the following conditions:

$$K(x_j, x_{-j}) = K(-x_j, x_{-j}) \quad \text{for any } 1 \leq j \leq k, \quad x_j \in \mathbb{R}, \quad x_{-j} \in \mathbb{R}^{k-1}$$

$$\int_{\mathbb{R}^k} K(x)dx = 1, \quad \int_{\mathbb{R}^k} |x|^2 K(x)dx < +\infty.$$

In other words, $K(\cdot)$ is an element-wise even function that integrates to 1 (so that the values of $K(\cdot)$ can be properly interpreted as weights). We let the bound of $K(\cdot)$ be $c > 0$, so that $|K(x)| \leq c$ for any $x \in \mathbb{R}^k$.

R4 Rate of Convergence of the Bandwidth

Given a sample of size n, we use the bandwidth $h_n > 0$ to compute the Nadaraya-Watson estimator, where

$$h_n \to 0, \quad nh_n^k \to \infty \quad \text{as } n \to \infty.$$

Note that assumptions R3 and R4 are the exact same as those for kernel density estimation. Assumption R2 simply requires the conditional mean function m to satisfy the same differentiability and boundedness conditions as the density f of the regressors. Under assumptions R2 to R4, the results we derived for kernel density estimation shows us that

$$\hat{f}_n(x_0; X^n, h_n) = \frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) \xrightarrow{p} f(x_0)$$

for any $x_0 \in \mathbb{R}^k$. As such, if $f(x_0) > 0$ for some $x_0 \in \mathbb{R}^k$, then $\hat{f}_n(x_0; X^n, h_n) > 0$ with probability approaching 1 as well, ensuring that the Nadaraya-Watson estimator is well-defined in the limit.

Under our assumptions, we can immediately establish the consistency and asymptotic normality of the Nadaraya-Watson estimator for $m(x_0)$. These results rely on the following asymptotic results, which are our "building blocks", so to say.

Lemma 8.13 (Asymptotic Building Blocks of Kernel Regression)
Suppose assumptions R1 to R3 hold true, and choose any $x_0 \in \mathbb{R}^k$. Then, there exist $c_1, c_2 > 0$ such that

$$\mathbb{E}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right] = 0$$

$$\left|\text{Var}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right] - \frac{1}{nh^k} \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right| \leq \frac{c_2 h^{1-k}}{n}$$

$$\left|\mathbb{E}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) m(X_i)\right] - m(x_0) f(x_0)\right| \leq c_1 h^2$$

$$\left|\text{Var}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) m(X_i)\right] - \frac{1}{nh^k} f(x_0) m(x_0)^2 \int_{\mathbb{R}^k} K(z)^2 dz\right| \leq \frac{c_2 h^{1-k} + (m(x_0) f(x_0) + c_1 h^2)^2}{n}$$

for any $n \in N_+$ and $h > 0$. In particular, if $\{h_n\}_{n \in N_+}$ is a sequence of positive reals that satisfies assumption R4, then

$$\frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \xrightarrow{p} 0$$

$$\frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) m(X_i) \xrightarrow{p} m(x_0) f(x_0)$$

as $n \to \infty$ and

$$\lim_{n\to\infty} \text{Var}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i\right] = \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz.$$

Proof) By lemma 8.8 and the regularity assumptions imposed on f, m and σ_0, we can see that there exist constants $c_1, c_2 > 0$ such that

$$\left|\int_{\mathbb{R}^k} K(z) g(x_0 + hz) dz - g(x_0)\right| \leq c_1 h^2$$

$$\left|\int_{\mathbb{R}^k} K(z)^2 g(x_0 + hz) dz - g(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right| \leq c_2 h$$

for $g = f, m, \sigma_0$ and their products $f \times \sigma_0$, $f \times m$ and $f \times m^2$.

Starting first with the sum involving the error terms,

$$\mathbb{E}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right] = \mathbb{E}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) \mathbb{E}\left[u_i \mid X_i\right]\right] = 0$$

by mean independence, and

$$\begin{aligned}
\text{Var}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right] &= \mathbb{E}\left|\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right|^2 \\
&= \frac{1}{nh^{2k}} \mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)^2 u_1^2\right] \\
&= \frac{1}{nh^{2k}} \mathbb{E}\left[\sigma_0(X_1) K\left(\frac{X_1 - x_0}{h}\right)^2\right]
\end{aligned}$$

by the iid nature of $K\left(\frac{X_n - x_0}{h}\right) u_n, \cdots, K\left(\frac{X_n - x_0}{h}\right) u_n$. Expanding the expectation as an integral shows us that

$$\begin{aligned}
\frac{1}{nh^{2k}} \mathbb{E}\left[\sigma_0(X_1) K\left(\frac{X_1 - x_0}{h}\right)^2\right] &= \frac{1}{nh^{2k}} \int_{\mathbb{R}^k} K\left(\frac{x - x_0}{h}\right)^2 \sigma_0(x) f(x) dx \\
&= \frac{1}{nh^k} \int_{\mathbb{R}^k} K(z)^2 (f \times \sigma_0)(x_0 + hz) dz,
\end{aligned}$$

where the final equality follows from the now-familiar change of variables formula. As such,

$$\left|\text{Var}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i\right] - \frac{1}{nh^k} \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right|$$

$$= \frac{1}{nh^k}\left|\int_{\mathbb{R}^k} K(z)^2(f\times\sigma_0)(x_0+hz)dz - \sigma_0(x_0)f(x_0)\int_{\mathbb{R}^k} K(z)^2dz\right| \leq \frac{c_2h^{1-k}}{n}.$$

Moving onto the terms involving only the regressors and the conditional mean equation, we have

$$\begin{aligned}\mathbb{E}\left[\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)m(X_i)\right] &= \frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right] \\ &= \frac{1}{h^k}\int_{\mathbb{R}^k} K\left(\frac{x-x_0}{h}\right)m(x)f(x)dx \\ &= \int_{\mathbb{R}^k} K(z)(f\times m)(x_0+hz)dz.\end{aligned}$$

Applying the results of lemma 9.1 to the pointwise product $f\times m$, it follows that

$$\begin{aligned}&\left|\mathbb{E}\left[\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)m(X_i)\right] - m(x_0)f(x_0)\right| \\ &\qquad = \left|\int_{\mathbb{R}^k} K(z)(f\times m)(x_0+hz)dz - (f\times m)(x_0)\right| \leq c_1h^2.\end{aligned}$$

Moving onto the variance term,

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)m(X_i)\right] &= \frac{1}{nh^{2k}}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right] \\ &= \frac{1}{n}\left[\frac{1}{h^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right|^2 - \left(\frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right]\right)^2\right].\end{aligned}$$

Note that

$$\begin{aligned}\frac{1}{h^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right|^2 &= \frac{1}{h^{2k}}\int_{\mathbb{R}^k} K\left(\frac{x-x_0}{h}\right)^2 m(x)^2f(x)dx \\ &= \frac{1}{h^k}\int_{\mathbb{R}^k} K(z)^2(f\times m^2)(x_0+hz)dz.\end{aligned}$$

Therefore, using lemma 8.8 applied to the function $f\times m^2$, we have

$$\begin{aligned}&\left|\frac{1}{h^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right|^2 - \frac{1}{h^k}f(x_0)m(x_0)^2\int_{\mathbb{R}^k} K(z)^2dz\right| \\ &\qquad = \frac{1}{h^k}\left|\int_{\mathbb{R}^k} K(z)^2(f\times m^2)(x_0+hz)dz - f(x_0)m(x_0)^2\int_{\mathbb{R}^k} K(z)^2dz\right| \leq c_2h^{1-k}.\end{aligned}$$

Furthermore,

$$\left(\frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right]\right)^2 \leq (m(x_0)f(x_0)+c_1h^2)^2,$$

so we have

$$\begin{aligned}
&\left|\mathrm{Var}\left[\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)m(X_i)\right] - \frac{1}{nh^k}f(x_0)m(x_0)^2\int_{\mathbb{R}^k}K(z)^2dz\right| \\
&\qquad \leq \frac{1}{n}\left|\frac{1}{h^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right|^2 - \frac{1}{h^k}f(x_0)m(x_0)^2\int_{\mathbb{R}^k}K(z)^2dz\right| \\
&\qquad\qquad + \frac{1}{n}\left(\frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1-x_0}{h}\right)m(X_1)\right]\right)^2 \\
&\qquad\qquad \leq \frac{c_2h^{1-k}+(m(x_0)f(x_0)+c_1h^2)^2}{n}.
\end{aligned}$$

It remains to prove the convergence in probability results. Let $\{h_n\}_{n\in N_+}$ be a sequence of positive reals satisfying assumption R4. Then, the previous result tells us that

$$\mathbb{E}\left|\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)u_i\right|^2 \leq \frac{1}{nh_n^k}\sigma_0(x_0)f(x_0)\int_{\mathbb{R}^k}K(z)^2dz + \frac{c_2h_n}{nh_n^k}$$

for any $n\in N_+$, and sending $n\to\infty$ on both sides shows us that

$$\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)u_i \xrightarrow{L^2} 0,$$

which implies $\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)u_i = o_p(1)$.

Likewise, for any $n\in N_+$ we have

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i) - m(x_0)f(x_0)\right|^2 \\
&\qquad = \mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i) - \mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i)\right]\right. \\
&\qquad\qquad \left. + \mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i)\right] - m(x_0)f(x_0)\right]^2 \\
&\leq 2\cdot\mathrm{Var}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i)\right] + 2\left(\mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)m(X_i)\right] - m(x_0)f(x_0)\right)^2
\end{aligned}$$

$$\leq \frac{c_2 h_n + (m(x_0)f(x_0) + c_1 h_n^2)^2 h_n^k}{n h_n^k} + c_1 h_n^2.$$

Taking $n \to \infty$ on both sides leads to the right hand side converging to 0, so we have

$$\frac{1}{n h_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) m(X_i) \xrightarrow{L^2} m(x_0) f(x_0).$$

Finally,

$$\left| \text{Var}\left[\frac{1}{\sqrt{n h_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \right] - \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz \right|$$

$$= n h_n^k \left| \text{Var}\left[\frac{1}{n h_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \right] - \frac{1}{n h_n^k} \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz \right|$$

$$\leq n h_n^k \frac{c_2 h_n^{1-k}}{n} = c_2 h_n$$

for any $n \in N_+$, so taking $n \to \infty$ on both sides shows us that

$$\lim_{n \to \infty} \text{Var}\left[\frac{1}{\sqrt{n h_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \right] = \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz.$$

This completes the proof.

Q.E.D.

8.3.3 Consistency and Asymptotic Normality

Under the assumptions made and asymptotic results derived in the previous section, we can now prove that the Nadaraya-Watson estimator is consistent without imposing any further regularity conditions.

Theorem 8.14 (Consistency of Kernel Regression Estimator)
Suppose assumptions R1 to R4 hold true, and choose any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$. Letting the Nadaraya-Watson estimator for $m(x_0)$ be defined as

$$\hat{m}_n(x_0; h_n) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)},$$

$\hat{m}_n(x_0; h_n)$ is consistent for $m(x_0)$, that is,

$$\hat{m}_n(x_0; h_n) \xrightarrow{p} m(x_0.).$$

Proof) As noted before the statement of the theorem, under assumptions R2 to R4

$$\hat{f}_n(x_0; X^n, h_n) = \frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) \xrightarrow{p} f(x_0) > 0,$$

so that $\hat{m}_n(x_0; h_n)$ is well defined with probability approaching 1. Using the fact that $Y_i = m(X_i) + u_i$, we can decompose $\hat{m}_n(x_0; h_n)$ as follows:

$$\hat{m}_n(x_0; h_n) = \frac{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) m(X_i)}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)} + \frac{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)}.$$

From the results of lemma 8.13, we know that

$$\frac{1}{nh^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right) u_i \xrightarrow{p} 0$$

$$\frac{1}{nh^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right) m(X_i) \xrightarrow{p} m(x_0) f(x_0).$$

Therefore,

$$\hat{m}_n(x_0; h_n) \xrightarrow{d} \frac{m(x_0) f(x_0)}{f(x_0)} + \frac{0}{f(x_0)} = m(x_0),$$

which is our desired result.

Q.E.D.

For asymptotic normality, we need further bounds on higher moments of the error term. To this end, we introduce the following assumptions:

R5 Boundedness of Moments

We assume that there exists a $\delta > 0$ such that the conditional $2+\delta$-moments of the error term u_i is well-behaved. Specifically, we assume that there exists a twice continuously differentiable function $b : \mathbb{R}^k \to \mathbb{R}_+$ such that

$$\mathbb{E}\left[|u_i|^{2+\delta} \mid X_i\right] = b(X_i).$$

Furthermore, we assume that there exists some $M_b > 0$ such that

$$\left|\frac{\partial b(x)}{\partial x}\right| \leq M_b$$

for any $x \in \mathbb{R}^k$.

If u_i and X_i are independent, then this assumption can be formulated in terms of the boundedness of the unconditional moments of u_i.

R6 Additional Smoothness Assumptions

We assume that the conditional mean function m and the density f of the regressors are four times continuously differentiable, with derivatives of all orders bounded above by some constant $M > 0$. Furthermore, we assume

$$\int_{\mathbb{R}^k} |z|^4 K(z) dz < +\infty,$$

which implies that $\int_{\mathbb{R}^k} |z|^r K(z) dz < +\infty$ for $1 \leq r \leq 4$.

Under the additional assumption, we have the following boundedness result:

Lemma 8.15 (Boundedness of Product of Kernel and Error)
Suppose assumptions R1 to R5 hold true. Then, for any $x_0 \in \mathbb{R}^k$,

$$\lim_{n\to\infty} n \cdot \mathbb{E}\left|\frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_1 - x_0}{h_n}\right) u_1\right|^{2+\delta} = 0.$$

Proof) Fix $x_0 \in \mathbb{R}^k$ and $n \in N_+$. Under assumption R2, we can Taylor expand the product $g = f \times b$ as follows: for any $z \in \mathbb{R}^k$,

$$|g(x + h_n z) - g(x_0)| \leq h_n M_{prod} |z|$$

by the boundedness of the derivatives of g, where $M_{prod} = \max(M, M_b) > 0$. Further-

more, by assumption R3

$$\int_{\mathbb{R}^k} |z||K(z)|^{2+\delta} dz \leq c^{1+\delta} \int_{\mathbb{R}^k} |z| K(z) dz < +\infty.$$

Now note that

$$n \cdot \mathbb{E}\left| \frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_1 - x_0}{h_n}\right) u_1 \right|^{2+\delta} = \frac{1}{n^{\frac{\delta}{2}} (h_n^k)^{1+\frac{\delta}{2}}} \mathbb{E}\left| K\left(\frac{X_1 - x_0}{h_n}\right) u_1 \right|^{2+\delta}.$$

By assumption R5,

$$\begin{aligned}
\mathbb{E}\left| K\left(\frac{X_1 - x_0}{h_n}\right) u_1 \right|^{2+\delta} &= \mathbb{E}\left[\left| K\left(\frac{X_1 - x_0}{h_n}\right) \right|^{2+\delta} \mathbb{E}\left[|u_1|^{2+\delta} \mid X_1 \right] \right] \\
&= \mathbb{E}\left[\left| K\left(\frac{X_1 - x_0}{h_n}\right) \right|^{2+\delta} b(X_1) \right] \\
&= \int_{\mathbb{R}^k} \left| K\left(\frac{x - x_0}{h_n}\right) \right|^{2+\delta} b(x) f(x) dx \\
&= h_n^k \int_{\mathbb{R}^k} |K(z)|^{2+\delta} g(x_0 + h_n z) dz,
\end{aligned}$$

where the last equality is derived from the usual linear change of variables formula. Therefore,

$$\begin{aligned}
&\left| n \cdot \mathbb{E}\left| \frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_1 - x_0}{h_n}\right) u_1 \right|^{2+\delta} - \frac{1}{(nh_n^k)^{\frac{\delta}{2}}} g(x_0) \int_{\mathbb{R}^k} |K(z)|^{2+\delta} dz \right| \\
&\quad = \left| \frac{1}{(nh_n^k)^{\frac{\delta}{2}}} \int_{\mathbb{R}^k} |K(z)|^{2+\delta} g(x_0 + h_n z) dz - \frac{1}{(nh_n^k)^{\frac{\delta}{2}}} g(x_0) \int_{\mathbb{R}^k} |K(z)|^{2+\delta} dz \right| \\
&\quad \leq \frac{h_n}{(nh_n^k)^{\frac{\delta}{2}}} \left(M_{prod} c^{1+\delta} \int_{\mathbb{R}^k} |z| K(z) dz \right).
\end{aligned}$$

Since

$$\frac{1}{(nh_n^k)^{\frac{\delta}{2}}} \to 0, \quad h_n \to 0$$

as $n \to \infty$ by assumption R4, so putting all these results together shows us that

$$\lim_{n \to \infty} n \cdot \mathbb{E}\left| \frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_1 - x_0}{h_n}\right) u_1 \right|^{2+\delta} = 0.$$

In particular, the sequence $\{n \cdot \mathbb{E}\left|\frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_1 - x_0}{h_n}\right) u_1\right|^{2+\delta}\}_{n \in N_+}$ is bounded.

Q.E.D.

Using this result, we can now prove a CLT for the cross product of the kernel and error term.

Lemma 8.16 (CLT for Cross Product of Kernel and Error)
Suppose assumptions R1 to R5 hold true. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$,

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right].$$

Proof) Define the triangular array $\{Z_{n,i}\}_{1 \leq i \leq n, n \in N_+}$ as

$$Z_{n,i} = \frac{1}{\sqrt{nh_n^k}} K\left(\frac{X_i - x_0}{h_n}\right) u_i$$

for $1 \leq i \leq n$. As before, we want to verify that the conditions for the Lindeberg-Feller CLT for triangular arrays hold, that is,

1. As $n \to \infty$,

$$s_n^2 = \text{Var}\left[\sum_{i=1}^{n} Z_{n,i}\right]$$

converges to some positive constant.

2. The Lindberg condition is satisfied:

$$\lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{E}\left[|Z_{n,i}|^2 \cdot I_{\{|Z_{n,i}| > \epsilon\}}\right] = 0$$

for any $\epsilon > 0$.

We first verify the condition for the variance; by lemma 8.13, we immediately have

$$s_n^2 = \text{Var}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i\right] \to \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz > 0,$$

where the final inequality follows because σ_0 is positive valued on $\mathbb{R}^k$, $f(x_0) > 0$ by design and $\int_{\mathbb{R}^k} K(z)^2 dz > 0$ from the fact that $K(\cdot)$ is non-negative and integrates to 1. Moving onto the Lindeberg condition, note first that the preceding result can be stated

as

$$\lim_{n\to\infty} n \cdot \mathbb{E}|Z_{n,1}|^2 = \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz > 0.$$

It follows that $\{n \cdot \mathbb{E}|Z_{n,1}|^2\}_{n\in N_+}$ is a bounded sequence and therefore that $n \cdot Z_{n,1}^2 = O_p(1)$ as $n \to \infty$.

Choose any $\epsilon > 0$. For any $n \in N_+$,

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[|Z_{n,i}|^2 \cdot I_{\{|Z_{n,i}|>\epsilon\}}\right] &= n \cdot \mathbb{E}\left[|Z_{n,1}|^2 \cdot I_{\{|Z_{n,1}|>\epsilon\}}\right] \\ &= n \cdot \mathbb{E}\left[|Z_{n,1}|^2 \cdot I_{\{n\cdot Z_{n,1}^2 > n\cdot\epsilon^2\}}\right].\end{aligned}$$

By Hölder's inequality,

$$\mathbb{E}\left[|Z_{n,1}|^2 \cdot I_{\{n\cdot Z_{n,1}^2 > n\cdot\epsilon^2\}}\right] \leq \left[\mathbb{E}|Z_{n,1}|^{2+\delta}\right]^{\frac{2}{2+\delta}} \left[\mathbb{P}\left(n \cdot Z_{n,1}^2 > n\epsilon^2\right)\right]^{\frac{\delta}{2+\delta}},$$

so we have

$$n \cdot \mathbb{E}\left[|Z_{n,1}|^2 \cdot I_{\{n\cdot Z_{n,1}^2 > n\cdot\epsilon^2\}}\right] \leq \left[n \cdot \mathbb{E}|Z_{n,1}|^{2+\delta}\right]^{\frac{2}{2+\delta}} \left[n \cdot \mathbb{P}\left(n \cdot Z_{n,1}^2 > n\epsilon^2\right)\right]^{\frac{\delta}{2+\delta}}.$$

The first term goes to 0 as $n \to \infty$ by the preceding lemma. The second term can be bounded above using Markov's inequality as follows:

$$n \cdot \mathbb{P}\left(n \cdot Z_{n,1}^2 > n\epsilon^2\right) \leq n \cdot \frac{1}{n\epsilon^2} \mathbb{E}\left[n \cdot Z_{n,1}^2\right] = \frac{s_n^2}{\epsilon^2}.$$

Since $s_n^2 = O(1)$ as $n \to \infty$, it follows that $n \cdot \mathbb{P}\left(n \cdot Z_{n,1}^2 > n\epsilon^2\right) = O(1)$ as well. Therefore,

$$\lim_{n\to\infty} n \cdot \mathbb{E}\left[|Z_{n,1}|^2 \cdot I_{\{n\cdot Z_{n,1}^2 > n\cdot\epsilon^2\}}\right] = 0,$$

and since this holds for any $\epsilon > 0$, the Lindeberg condition holds. By the Lindeberg-Feller CLT for triangular arrays, we can now conclude that

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i = \sum_{i=1}^{n} Z_{n,i} \xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right].$$

Q.E.D.

Finally, for kernel regressions we can obtain a closed form expression for the asymptotic bias, which relies on the following result:

Lemma 8.17 (Asymptotic Bias of Nadaraya-Watson Estimator)
Suppose assumptions R1 to R6 hold true, and that $nh_n^{k+8} \to 0$ as $n \to \infty$. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$,

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot f(x_0) b(x_0; h_n) \xrightarrow{p} 0,$$

where

$$b(x_0; h) = \frac{1}{2} h^2 \sum_{j=1}^{k} \left[\frac{\partial^2 m(x_0)}{\partial x_j^2} + 2\frac{1}{f(x_0)} \frac{\partial m(x_0)}{\partial x_j} \frac{\partial f(x_0)}{\partial x_j}\right] \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz\right).$$

for any $h > 0$.

Proof) Define the functions $g_1, g_2 : \mathbb{R}^k \to \mathbb{R}$ as

$$g_1(x) = (m(x) - m(x_0)) f(x), \quad g_2(x) = (m(x) - m(x_0))^2 f(x)$$

for any $x \in \mathbb{R}^k$. By assumption R2, g_2 satisfies the requisite smoothness and boundedness conditions for lemma 8.8 to hold: that is, there exists a $c_2 > 0$ such that

$$\left|\int_{\mathbb{R}^k} K(z)^2 g_2(x_0 + hz) dz\right| \leq c_2 h$$

for any $n \in N_+$ and $h > 0$, where we use the fact that $g_2(x_0) = 0$.

In contrast to g_2, which requires only a first order expansion around x_0, we employ a fourth order expansion of g_1 around x_0, once again stated as an inequality: for any $h > 0$ and $z \in \mathbb{R}^k$,

$$\left|g_1(x_0 + hz) - h \cdot z' \frac{\partial g_1(x_0)}{\partial x} - \frac{h^2}{2} \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \frac{\partial^2 g_1(x_0)}{\partial x_{j_1} \partial x_{j_2}} z_{j_1} z_{j_2} - \frac{h^3}{3!} \sum_{j_1=1}^{k} \cdots \sum_{j_3=1}^{k} \frac{\partial^3 g_1(x_0)}{\partial x_{j_1} \cdots \partial x_{j_3}} \left(\prod_{i=1}^{3} z_{j_i}\right)\right|$$
$$\leq \frac{h^4}{4!} M k^4 |z|^4.$$

Now it follows that

$$\left|\int_{\mathbb{R}^k} K(z) g_1(x_0 + hz) dz - h \sum_{j=1}^{k} \frac{\partial g_1(x_0)}{\partial x_j} \left(\int_{\mathbb{R}^k} z_j K(z) dz\right) - \frac{h^2}{2} \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \frac{\partial^2 g_1(x_0)}{\partial x_{j_1} \partial x_{j_2}} \left(\int_{\mathbb{R}^k} z_{j_1} z_{j_2} K(z) dz\right)\right.$$
$$\left. - \frac{h^3}{3!} \sum_{j_1=1}^{k} \cdots \sum_{j_3=1}^{k} \frac{\partial^3 g_1(x_0)}{\partial x_{j_1} \cdots \partial x_{j_3}} \left(\int_{\mathbb{R}^k} \left(\prod_{i=1}^{3} z_{j_i}\right) K(z) dz\right)\right|$$

$$\leq \frac{h^4}{4!}\left(Mk^4\int_{\mathbb{R}^k}|z|^4K(z)dz\right).$$

Again, the fact that $K(\cdot)$ is an elementwise even function implies that

$$\int_{\mathbb{R}^k} z_{j_1}z_{j_2}K(z)dz = 0 \quad \text{if } j_1 \neq j_2$$
$$\int_{\mathbb{R}^k}\left(\prod_{i=1}^{3} z_{j_i}\right)K(z)dz = 0 \quad \text{for any } 1 \leq j_1, j_2, j_3 \leq k,$$

so that

$$\left|\int_{\mathbb{R}^k} K(z)g_1(x_0+hz)dz - \frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 g_1(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2K(z)dz\right)\right|$$
$$\leq \frac{h^4}{4!}\left(Mk^4\cdot\int_{\mathbb{R}^k}|z|^4K(z)dz\right).$$

Defining $\frac{Mk^4}{4!}\cdot\int_{\mathbb{R}^k}|z|^4K(z)dz = c_1 > 0$, it follows that

$$\left|\int_{\mathbb{R}^k} K(z)g_1(x_0+hz)dz - \frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 g_1(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2K(z)dz\right)\right| \leq c_1h^4$$

for any $h > 0$ and $n \in N_+$. Define

$$b(x_0;h) = \frac{h^2}{2f(x_0)}\sum_{j=1}^{k}\frac{\partial^2 g_1(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2K(z)dz\right),$$

so that

$$\left|\int_{\mathbb{R}^k} K(z)g_1(x_0+hz)dz - b(x_0;h)f(x_0)\right| \leq c_1h^4$$

for any $h > 0$ and $n \in N_+$.

For any $n \in N_+$, we have

$$\mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] = \frac{1}{h_n^k}\mathbb{E}\left[K\left(\frac{X_1-x_0}{h_n}\right)(m(X_1)-m(x_0))\right]$$
$$= \frac{1}{h_n^k}\int_{\mathbb{R}^k}K\left(\frac{x-x_0}{h_n}\right)g_1(x)dx$$
$$= \int_{\mathbb{R}^k}K(z)g_1(x_0+h_nz)dz,$$

so that the preceding result reveals that

$$\left|\mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] - b(x_0;h_n)f(x_0)\right| \leq c_1h_n^4.$$

We can expand the variance as

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] \\
&\qquad = \frac{1}{nh_n^{2k}}\mathrm{Var}\left[K\left(\frac{X_1-x_0}{h_n}\right)(m(X_1)-m(x_0))\right] \\
&\qquad \le \frac{1}{n}\times\frac{1}{h_n^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h_n}\right)(m(X_1)-m(x_0))\right|^2,
\end{aligned}$$

where

$$\begin{aligned}
\frac{1}{h_n^{2k}}\mathbb{E}\left|K\left(\frac{X_1-x_0}{h_n}\right)(m(X_1)-m(x_0))\right|^2 &= \frac{1}{h_n^{2k}}\int_{\mathbb{R}^k} K\left(\frac{x-x_0}{h_n}\right)^2 g_2(x)dx \\
&= \frac{1}{h_n^k}\int_{\mathbb{R}^k} K(z)^2 g_2(x_0+h_n z)dz \le c_2 h_n^{1-k}.
\end{aligned}$$

It follows that

$$\mathrm{Var}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] \le \frac{c_2 h_n^{1-k}}{n} = \frac{c_2 h_n}{nh_n^k}.$$

Finally, since

$$\begin{aligned}
&\mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] \\
&\qquad = nh_n^k\mathrm{Var}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] \le c_2 h_n,
\end{aligned}$$

taking $n\to\infty$ on both sides shows us that

$$\lim_{n\to\infty}\mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] = 0.$$

On the other hand,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] - \sqrt{nh_n^k}\times b(x_0;h_n)f(x_0)\right| \\
&= \sqrt{nh_n^k}\cdot\left|\mathbb{E}\left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))\right] - b(x_0;h_n)f(x_0)\right| \le c_1\sqrt{nh_n^{k+8}}.
\end{aligned}$$

By assumption, $nh_n^{k+8} \to 0$ as $n \to \infty$, so

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) - \sqrt{nh_n^k} \times b(x_0; h_n) f(x_0)\right] = 0$$

as well.

Putting the pieces together, we can decompose the mean squared error into the sum of the variance plus the squared bias:

$$\begin{aligned}
&\mathbb{E}\left|\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) - \sqrt{nh_n^k} \times b(x_0; h_n) f(x_0)\right|^2 \\
&\qquad = \mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0))\right] \\
&\qquad + \left(\mathbb{E}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0))\right] - \sqrt{nh_n^k} \times b(x_0; h_n) f(x_0)\right)^2.
\end{aligned}$$

Therefore, taking $n \to \infty$ on both sides yields

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) - \sqrt{nh_n^k} \times b(x_0; h_n) f(x_0) = o_p(1)$$

as $n \to \infty$, which is exactly the result we want.

Finally, note that

$$\frac{\partial^2 g_1(x_0)}{\partial x_j^2} = \frac{\partial^2 m(x_0)}{\partial x_j^2} f(x_0) + 2\frac{\partial m(x_0)}{\partial x_j} \times \frac{\partial f(x_0)}{\partial x_j},$$

so that the expression for the bias term $b(x_0; h)$ is exactly as stated in the claim of the lemma.

Q.E.D.

We can now state the asymptotic normality result:

Theorem 8.18 (Asymptotic Normality of Kernel Regression Estimator) Suppose assumptions R1 to R6 hold true, and choose any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$. Letting the Nadaraya-Watson estimator for $m(x_0)$ be defined as

$$\hat{m}_n(x_0; h_n) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)},$$

if $nh_n^{k+8} \to 0$ as $n \to \infty$, then $\hat{m}_n(x_0; h_n)$ has the following asymptotic distribution:

$$\sqrt{nh_n^k}\Big[\hat{m}_n(x_0; h_n) - m(x_0) - b(x_0; h_n)\Big] \xrightarrow{d} \mathcal{N}\left[0,\ \frac{\sigma_0(x_0)}{f(x_0)} \int_{\mathbb{R}^k} K(z)^2 dz\right],$$

where

$$b(x_0; h) = h^2 \sum_{j=1}^k \left[\frac{1}{2} \cdot \frac{\partial^2 m(x_0)}{\partial x_j^2} + \frac{1}{f(x_0)} \frac{\partial m(x_0)}{\partial x_j} \frac{\partial f(x_0)}{\partial x_j}\right] \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz\right)$$

for any $h > 0$.

Proof) Again, we rely on the expansion

$$\hat{m}_n(x_0; h_n) = \frac{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) m(X_i)}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)} + \frac{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)}.$$

We can see that

$$\sqrt{nh_n^k}\Big[\hat{m}_n(x_0; h_n) - m(x_0)\Big] = \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) m(X_i)}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)} - \sqrt{nh_n^k} m(x_0)$$
$$+ \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)}.$$

We can further rearrange the terms above, so that

$$\sqrt{nh_n^k}\Big[\hat{m}_n(x_0; h_n) - m(x_0) - b(x_0; h_n)\Big]$$
$$= \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) (m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot b(x_0; h_n) \hat{f}_n(x_0; h_n)}{\hat{f}_n(x_0; h_n)} + \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\hat{f}_n(x_0; h_n)},$$

where

$$\hat{f}_n(x_0; h_n) = \frac{1}{nh_n^k} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) \xrightarrow{p} f(x_0).$$

and $\sqrt{nh_n^k}\left[\hat{f}_n(x_0;h_n) - \mathbb{E}\left[\hat{f}_n(x_0;h_n)\right]\right] = O_p(1)$ under the maintained assumptions. We study each term in turn.

Cross Product of Kernel and Error

From lemma 8.16, we know that

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i \xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0) f(x_0) \int_{\mathbb{R}^k} K(z)^2 dz\right].$$

Therefore, by Slutsky's theorem,

$$\frac{\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\hat{f}_n(x_0;h_n)} \xrightarrow{d} \mathcal{N}\left[0,\ \frac{\sigma_0(x_0)}{f(x_0)} \int_{\mathbb{R}^k} K(z)^2 dz\right].$$

Cross Product of Kernel and Conditional Mean Equation

From lemma 8.17, we have

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot f(x_0) b(x_0;h_n) \xrightarrow{p} 0.$$

Meanwhile, from lemma 9.5 we can see that there exists a $c_4 > 0$ such that

$$\left|\mathbb{E}\left[\hat{f}_n(x_0;h_n)\right] - f(x_0) - \frac{h_n^2}{2}\sum_{j=1}^{k} \frac{\partial^2 f(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2 K(z) dz\right)\right| \leq c_4 h_n^4$$

for any $n \in N_+$. Since

$$\begin{aligned}\left|\sqrt{nh_n^k}\left[\hat{f}_n(x_0;h_n) - f(x_0)\right]\right| &\leq \left|\sqrt{nh_n^k}\left[\hat{f}_n(x_0;h_n) - \mathbb{E}\left[\hat{f}_n(x_0;h_n)\right]\right]\right| + \left|\sqrt{nh_n^k} \times \left[\mathbb{E}\left[\hat{f}_n(x_0;h_n)\right] - f(x_0)\right]\right| \\ &\leq O_p(1) + \frac{M}{2}\sqrt{nh_n^{k+4}}\left(\int_{\mathbb{R}^k} |z|^2 K(z) dz\right) + c_4\sqrt{nh_n^{k+8}},\end{aligned}$$

and

$$b(x_0;h_n) = C(x_0) h_n^2,$$

where

$$C(x_0) = \sum_{j=1}^{k} \left[\frac{1}{2} \cdot \frac{\partial^2 m(x_0)}{\partial x_j^2} + \frac{1}{f(x_0)} \frac{\partial m(x_0)}{\partial x_j} \frac{\partial f(x_0)}{\partial x_j} \right] \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right),$$

it follows that

$$\left| b_n(x_0; h_n) \cdot \sqrt{nh_n^k} \left[\hat{f}_n(x_0; h_n) - f(x_0) \right] \right| \leq O_p(1) h_n^2 + \frac{M}{2} \sqrt{nh_n^{k+8}} \left(\int_{\mathbb{R}^k} |z|^2 K(z) dz \right) + c_4 h_n^2 \sqrt{nh_n^{k+8}}.$$

Under our assumptions, taking $n \to \infty$ on the right hand side sends it to 0, so

$$b_n(x_0; h_n) \cdot \frac{\sqrt{nh_n^k} \left[f(x_0) - \hat{f}_n(x_0; h_n) \right]}{\hat{f}_n(x_0; h_n)} \xrightarrow{p} \frac{0}{f(x_0)} = 0.$$

Putting these results together, we obtain

$$\frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) (m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot b(x_0; h_n) \hat{f}_n(x_0; h_n)}{\hat{f}_n(x_0; h_n)}$$

$$= \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) (m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot f(x_0) b(x_0; h_n)}{\hat{f}_n(x_0; h_n)} + b_n(x_0; h_n) \cdot \frac{\sqrt{nh_n^k} \left[f(x_0) - \hat{f}_n(x_0; h_n) \right]}{\hat{f}_n(x_0; h_n)},$$

and because both terms on the right hand side are $o_p(1)$,

$$\frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) (m(X_i) - m(x_0)) - \sqrt{nh_n^k} \cdot b(x_0; h_n) \hat{f}_n(x_0; h_n)}{\hat{f}_n(x_0; h_n)} = o_p(1)$$

as well.

Asymptotic Normality

We have shown above that

$$\sqrt{nh_n^k} \left[\hat{m}_n(x_0; h_n) - m(x_0) - b(x_0; h_n) \right] = \frac{\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i}{\frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)} + o_p(1),$$

so by Slutsky's theorem,

$$\sqrt{nh_n^k} \left[\hat{m}_n(x_0; h_n) - m(x_0) - b(x_0; h_n) \right] \xrightarrow{d} \mathcal{N}\left[0, \frac{\sigma_0(x_0)}{f(x_0)} \int_{\mathbb{R}^k} K(z)^2 dz \right].$$

This completes the proof.

Q.E.D.

Note that the asymptotic normality of the Nadaraya-Watson estimator in the above form holds only if the bandwidth h_n converges to 0 not too fast nor too slow. Specifically, it must be the case that

$$nh_n^k \to +\infty, \quad nh_n^{k+8} \to 0$$

as $n \to \infty$, so that h_n^k converges to 0 slower than $\frac{1}{n}$ but faster than $\frac{1}{n}h_n^{-8}$. One choice of h_n that satisfies these two conditions is a bandwidth of the form

$$h_n = C \cdot n^r$$

where C is some positive constant and $r \in \mathbb{R}$ satisfies

$$-\frac{1}{k} < r < -\frac{1}{k+8}.$$

8.3.4 The Optimal Bandwidth and the Cross Validation Criterion

As with kernel density estimation, the Nadaraya-Watson estimator is sensitive to the choice of bandwidth h_n. There are two approaches to bandwidth selection that we can take at this point. First, as in the previous section, we can choose the bandwidth that minimizes some asymptotic mean squared error criterion. However, it turns out that the resulting bandwidth is much more computationally intensive than that of kernel density estimation, so a feasible, data-driven alternative is to find the bandwidth that minimizes the cross validation criterion. We introduce both methods below.

The Mean Squared Error Approach

As with kernel density estimation, we can choose the optimal bandwidth h_n by minimizing the weighted integrated mean squared error (WIMSE)

$$WIMSE_n(h) = \int_{\mathbb{R}^k} \left(\mathbb{E}|\hat{m}_n(x_0;h) - m(x_0)|^2 \right) f(x_0) dx_0.$$

Instead of just integrating the mean squared error over all values in $\mathbb{R}^k$, this time we work with a criterion that places greater weight on areas of $\mathbb{R}^k$ at which the regressors are more likely to take values. Once again, the mean squared error at $x_0 \in \mathbb{R}^k$ can be decomposed into the sum of the variance and the squared bias:

$$\begin{aligned} MSE_n(x_0;h) &:= \mathbb{E}|\hat{m}_n(x_0;h) - m(x_0)|^2 \\ &= \mathrm{Var}\left[\hat{m}_n(x_0;h)\right] + \left(\mathbb{E}\left[\hat{m}_n(x_0;h)\right] - m(x_0)\right)^2. \end{aligned}$$

The asymptotic variance of $\hat{m}_n(x_0; h_n)$ when h_n satisfies assumption R4 and $nh_n^{k+8} \to 0$ is given as

$$\frac{\sigma_0(x_0)}{f(x_0)} \left(\int_{\mathbb{R}^k} K(z)^2 dz \right),$$

while the asymptotic bias has been obtained as

$$b(x_0; h_n) = h_n^2 \sum_{j=1}^{k} \underbrace{\left[\frac{1}{2} \cdot \frac{\partial^2 m(x_0)}{\partial x_j^2} + \frac{1}{f(x_0)} \frac{\partial m(x_0)}{\partial x_j} \frac{\partial f(x_0)}{\partial x_j} \right]}_{\frac{1}{2f(x_0)} \frac{\partial g_1(x_0)}{\partial x_j^2}} \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right).$$

Therefore, the asymptotic weighted integrated mean squared error (AWIMSE) under the bandwidth $h > 0$ is defined as

$$\begin{aligned} AWIMSE_n(h) = & \frac{1}{nh^k} \left(\int_{\mathbb{R}^k} \sigma_0(x_0) dx_0 \right) \left(\int_{\mathbb{R}^k} K(z)^2 dz \right) \\ & + \frac{h^4}{4} \sum_{j=1}^{k} \sum_{m=1}^{k} \left(\int_{\mathbb{R}^k} \frac{\partial g_1(x_0)}{\partial x_j^2} \cdot \frac{\partial g_1(x_0)}{\partial x_m^2} \cdot \frac{1}{f(x_0)} dx_0 \right) \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} z_m^2 K(z) dz \right). \end{aligned}$$

The second derivative of $AWIMSE_n(h)$ with respect to h is always positive, so $AWIMSE_n(\cdot)$ is strictly convex on $(0, +\infty)$. This tells us that the optimal bandwidth h_n^* satisfies the first order condition

$$\begin{aligned} 0 = AWIMSE_n'(h_n^*) = & -k \frac{1}{n(h_n^*)^{k+1}} \left(\int_{\mathbb{R}^k} \sigma_0(x_0) dx_0 \right) \left(\int_{\mathbb{R}^k} K(z)^2 dz \right) \\ & + (h_n^*)^3 \sum_{j=1}^{k} \sum_{m=1}^{k} \left(\int_{\mathbb{R}^k} \frac{\partial g_1(x_0)}{\partial x_j^2} \cdot \frac{\partial g_1(x_0)}{\partial x_m^2} \cdot \frac{1}{f(x_0)} dx_0 \right) \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} z_m^2 K(z) dz \right). \end{aligned}$$

Solving this equation for h_n^* yields

$$h_n^* = \left(\frac{k \left(\int_{\mathbb{R}^k} \sigma_0(x_0) dx_0 \right) \left(\int_{\mathbb{R}^k} K(z)^2 dz \right)}{\sum_{j=1}^{k} \sum_{m=1}^{k} \left(\int_{\mathbb{R}^k} \frac{\partial g_1(x_0)}{\partial x_j^2} \cdot \frac{\partial g_1(x_0)}{\partial x_m^2} \cdot \frac{1}{f(x_0)} dx_0 \right) \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} z_m^2 K(z) dz \right)} \right) \times n^{-1/(k+4)}.$$

Defining

$$c^* = \frac{k \left(\int_{\mathbb{R}^k} \sigma_0(x_0) dx_0 \right) \left(\int_{\mathbb{R}^k} K(z)^2 dz \right)}{\sum_{j=1}^{k} \sum_{m=1}^{k} \left(\int_{\mathbb{R}^k} \frac{\partial g_1(x_0)}{\partial x_j^2} \cdot \frac{\partial g_1(x_0)}{\partial x_m^2} \cdot \frac{1}{f(x_0)} dx_0 \right) \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) \left(\int_{\mathbb{R}^k} z_m^2 K(z) dz \right)},$$

we can write $h_n^* = c^* n^{-1/(k+4)}$; note that $h_n^* \to 0$, $n(h_n^*)^{k+8} \to 0$ and $n(h_n^*)^k \to +\infty$ as $n \to \infty$, as we required. In practice, computing the optimal bandwidth requires knowledge not only of the kernel $K(\cdot)$, but also of the conditional variance function $\sigma_0(\cdot)$, the derivatives of the conditional mean function $m(\cdot)$, and the derivatives of the density f. Therefore, we require an alternative means of obtaining the optimal bandwidth that allows us to calculate the bandwidth using only

the data at hand.

The Cross Validation Criterion

In kernel regressions it is customary to rely on the cross validation criterion to find the optimal bandwidth, primarily because it has a closed-form expression that is very easy to compute.

The leave-one-out cross validation criterion is defined as

$$CV_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{m}_{n,-i}(X_i;h)\right)^2,$$

where $\hat{m}_{n,-i}(x_0;h)$ is the kernel estimator of $m(x_0)$ that is calculated without the ith observations:

$$\hat{m}_{n,-i}(x_0;h) = \frac{\sum_{j\neq i} K\left(\frac{X_j - x_0}{h}\right) Y_j}{\sum_{j\neq i} K\left(\frac{X_j - x_0}{h}\right)}$$

for any $x_0 \in \mathbb{R}^k$. Heuristically, $Y_i - \hat{m}_{n,-i}(X_i;h)$ is the prediction error for observation i under the bandwidth $h > 0$, given that we use all observations except for the ith observation to construct our prediction of Y_i. It follows that $CV_n(h)$ is the average prediction error for each observation under the bandwidth h. As such, a smaller $CV_n(h)$ can be interpreted as higher predictive power.

The cross validation criterion is also a sensible criterion for bandwidth selection because it turns out to have an interpretation as an estimator of the WIMSE. This can be seen as follows. Using the fact that $Y_1 = m(X_1) + u_1$, we have

$$\begin{aligned}\mathbb{E}|Y_1 - \hat{m}_{n,-1}(X_1;h)|^2 &= \mathbb{E}\left[\left[(m(X_1) - \hat{m}_{n,-1}(X_1;h)) + u_1\right]^2\right] \\ &= \mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)^2\right] + 2\mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)u_1\right] + \mathbb{E}\left[u_1^2\right].\end{aligned}$$

The law of iterated expectations and mean independence of the error term shows us that the second term equals 0:

$$\mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)u_1\right] = \mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)\mathbb{E}\left[u_1 \mid X^n\right]\right] = 0.$$

Meanwhile, note that

$$\mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)^2 \mid X_2, \cdots, X_n\right] = \int_{\mathbb{R}^k}\left(m(x) - \hat{m}_{n,-1}(x;h)\right)^2 f(x)dx$$

by the independence of X_1 and $X_2, \cdots, X_n$. Therefore,

$$\mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left(m(X_1) - \hat{m}_{n,-1}(X_1;h)\right)^2 \mid X_2, \cdots, X_n\right]\right]$$

$$= \mathbb{E}\left[\int_{\mathbb{R}^k} \left(m(x) - \hat{m}_{n,-1}(x;h)\right)^2 f(x)dx\right]$$

$$= \int_{\mathbb{R}^k} \mathbb{E}\left[\left(m(x) - \hat{m}_{n,-1}(x;h)\right)^2\right] f(x)dx,$$

where the last equality follows from Fubini's theorem. Finally, the iid nature of the sample $X_1, \cdots, X_n$ tells us that

$$\mathbb{E}\left[\left(m(x) - \hat{m}_{n,-1}(x;h)\right)^2\right] = \mathbb{E}\left[\left(m(x) - \hat{m}_{n-1}(x;h)\right)^2\right] = MSE_{n-1}(x;h),$$

so that we have

$$\mathbb{E}|Y_1 - \hat{m}_{n,-1}(X_1;h)|^2 = \underbrace{\int_{\mathbb{R}^k} MSE_{n-1}(x_0;h) f(x_0)dx_0}_{WIMSE_{n-1}(h)} + \int_{\mathbb{R}^k} \sigma_0(x_0) f(x_0)dx_0.$$

In other words, $CV_n(h)$, being the sample analogue of the population moment $\mathbb{E}|Y_1 - \hat{m}_{n,-1}(X_1;h)|^2$, consistently estimates the WIMSE plus some noise term involving the conditional variance of the errors. This demonstrates that the bandwidth minimizing the cross validation criterion is likely, in large samples, to be close to the bandwidth that minimizes the WIMSE.

For the Nadaraya-Watson estimator the cross validation criterion can be written in a particularly simple form. For any $1 \leq i \leq n$, note that

$$Y_i - \hat{m}_{n,-i}(X_i;h) = Y_i - \frac{\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right) Y_j}{\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right)}$$

$$= Y_i - \frac{\sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right) Y_j - K(0) Y_i}{\sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right) - K(0)}$$

$$= \frac{\sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - X_i}{h}\right) - K(0)} \left(Y_i - \hat{m}_n(X_i;h)\right).$$

Defining the weights $\hat{w}_{n,1}(x_0;h), \cdots, \hat{w}_{n,n}(x_0;h)$ as

$$\hat{w}_{n,i}(x_0;h) = \frac{K\left(\frac{X_i - x_0}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_j - x_0}{h}\right)},$$

we can write

$$\hat{m}_n(x_0;h) = \sum_{i=1}^n \hat{w}_{n,i}(x_0;h) Y_i$$

for any $x_0 \in \mathbb{R}^k$ and

$$Y_i - \hat{m}_{n,-i}(X_i; h) = \frac{Y_i - \hat{m}_n(X_i; h)}{1 - \hat{w}_{n,i}(X_i; h)}$$

for each $1 \leq i \leq n$. Therefore, the cross validation criterion becomes

$$CV_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \hat{m}_n(X_i; h)}{1 - \hat{w}_{n,i}(X_i; h)} \right)^2,$$

which requires us to only compute the Nadaraya-Watson estimator once instead of having to compute it n times with one observation left out each time.

8.4 Local Linear Regression

Local linear regression is a kind of middle ground between linearly fitting the conditional mean and non-parametrically estimating the conditional mean function à la the Nadaraya-Watson estimator.

8.4.1 Motivation: The Boundary Bias Problem

To motivate local linear regression, we return to the Nadaraya-Watson estimator; specifically, under the optimal bandwidth $h_n^* = c^* n^{-1/(k+4)}$, the NW estimator has the following asymptotic distribution:

$$n^{\frac{2}{k+4}}(c^*)^k\left[\hat{m}_n(x_0;h_n^*) - m(x_0) - b(x_0,h_n^*)\right] \xrightarrow{d} \mathcal{N}\left[0,\ \frac{\sigma_0(x_0)}{f(x_0)}\int_{\mathbb{R}^k} K(z)^2 dz\right],$$

where

$$n^{\frac{2}{k+4}}(c^*)^k b(x_0,h_n^*) = (c^*)^{k+2}\sum_{j=1}^{k}\left[\frac{1}{2}\cdot\frac{\partial^2 m(x_0)}{\partial x_j^2} + \frac{1}{f(x_0)}\frac{\partial m(x_0)}{\partial x_j}\frac{\partial f(x_0)}{\partial x_j}\right]\left(\int_{\mathbb{R}^k} z_j^2 K(z)dz\right).$$

As such, under the optimal bandwidth the asymptotic bias is constant and therefore does not disappear in the limit. The bias depends on many things, but we want to focus our attention on the term $\frac{1}{f(x_0)}$. If $f(x_0)$ is small, then $\hat{m}_n(x_0;h_n^*)$ has a large upward bias when $\frac{\partial m(x_0)}{\partial x_j}\frac{\partial f(x_0)}{\partial x_j} > 0$ and a large downward bias when $\frac{\partial m(x_0)}{\partial x_j}\frac{\partial f(x_0)}{\partial x_j} < 0$.

To get a sense of how such large biases can manifest in practice, consider the following example. Suppose $k = 1$ and the true conditional mean function is an increasing linear function. Consider a point x_0 close to the lower boundary of the support of the regressor X_i, so that $f(x_0)$ is close to 0. Then, since $\frac{\partial f(x_0)}{\partial x}$ is positive and the increasing nature of the conditional mean function implies that $\frac{\partial m(x_0)}{\partial x} > 0$ as well, the Nadaraya-Watson estimator has a large upward bias at x_0, that is,

$$\hat{m}_n(x_0;h_n^*) > m(x_0).$$

As such, instead of producing a linear estimate of the conditional mean function, the NW method is likely to produce a conditional mean function that flattens out at the lower boundary of the support of X_i. This is called the boundary bias problem, and it turns out that local linear regression is largely free of this problem, hence its popularity in the literature.

Mathematically, local linear regression is motivated by the following first order Taylor approximation of $m(x)$ around x_0:

$$m(x) \approx m(x_0) + \left(\frac{\partial m(x_0)}{\partial x}\right)'(x - x_0).$$

If the first derivative of m does not exhibit much variation, then the conditional mean function can be well approximated as a linear model, and $m(x_0)$ can be estimated by the OLS estimator

of the intercept from a regression of Y_i on a constant and $X_i - x_0$.

8.4.2 The Local Linear Estimator and Asymptotic Preliminaries

Formally, we can obtain the local linear estimator using a simple weighted least squares scheme. For any $x_0 \in \mathbb{R}^k$ at which $f(x_0) > 0$, the local linear estimator of $m(x_0)$ using the bandwidth $h > 0$ is constructed in two steps:

Step 1 Minimizing the Kernel-weighted Sum of Squared Errors

First, we solve the following minimization problem:

$$\min_{(a_0,b_0)\in\mathbb{R}^{k+1}} \quad \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(Y_i - a_0 - (X_i - x_0)'b_0)^2.$$

Step 2 Obtaining the Estimator

Letting $\hat{a}_n$ and $\hat{b}_n$ be the minimizers obtained from the above problem, we let the estimator of $m(x_0)$ be given as

$$\hat{m}_n(x_0; h) = \hat{a}_n.$$

The squared deviations in the minimization step are weighted using the kernel $K(\cdot)$ to account for possible non-linearities in the data; if the kernel is uniform, then this reduces to an OLS problem, while if $b_0 = O_{k\times 1}$, then this procedure yields the Nadaraya-Watson estimator. In this step, local linear regression can be seen as a middle ground between the linear model and the completely non-parametric model.

The analytical expression for $\hat{m}_n(x_0; h)$ can be found as follows. Defining $Z_i = \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix}$ for any $1 \le i \le n$, $\hat{\beta}_n = (\hat{a}_n, \hat{b}_n')'$ satisfies the following the first order conditions:

$$\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) Z_i \left(Y_i - Z_i'\hat{\beta}_n\right) = O_{(k+1)\times 1},$$

solving which we obtain

$$\hat{\beta}_n = \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) Z_i Z_i'\right]^{-1} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) Z_i Y_i\right],$$

so that $\hat{a}_n = e_1'\hat{\beta}_n$, where $e_1 \in \mathbb{R}^{k+1}$ is the first standard basis vector of $\mathbb{R}^{k+1}$. Now, the formula for the block matrix inversion shows us that the closed form expression for $\hat{m}_n(x_0; h)$ is given as

$$\left(\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right) - \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right)(X_i - x_0)\right]' \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1} \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right)(X_i - x_0)\right]\right)^{-1}$$

$$\times \left(\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right) Y_i - \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right) (X_i - x_0) \right]' \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right) (X_i - x_0)(X_i - x_0)' \right]^{-1} \left[\sum_{i=1}^{n} K\left(\tfrac{X_i - x_0}{h}\right) (X_i - x_0) Y_i \right] \right).$$

This is the familiar expression for the intercept of an OLS regression of Y_i on a constant and $X_i - x_0$ we would obtain from an application of the FWL theorem, except this time the observations are weighted according to the kernel $K(\cdot)$. For notational brevity, we denote the XX part of the estimator above by $\hat{B}_n(x_0; h)$ and the XY part as $\hat{A}_n(x_0; h)$, so that

$$\hat{m}_n(x_0; h) = \left(\hat{B}_n(x_0; h) \right)^{-1} \hat{A}_n(x_0; h).$$

In addition, we can further decompose $\hat{A}_n(x_0; h)$ as

$$\begin{aligned}
\hat{A}_n(x_0; h) = & \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) m(X_i) - \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0) \right]' \\
& \times \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0)(X_i - x_0)' \right]^{-1} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0) m(X_i) \right] \\
& + \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) u_i - \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0) \right]' \\
& \times \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0)(X_i - x_0)' \right]^{-1} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right) (X_i - x_0) u_i \right],
\end{aligned}$$

using the fact that $Y_i = m(X_i) + u_i$. We denote the first part of this sum as $\hat{A}_n^{(m)}(x_0; h)$, and the second part as $\hat{A}_n^{(u)}(x_0; h)$; as such,

$$\hat{m}_n(x_0; h) = \left(\hat{B}_n(x_0; h) \right)^{-1} \left[\hat{A}_n^{(m)}(x_0; h) + \hat{A}_n^{(u)}(x_0; h) \right].$$

We now investigate the asymptotic properties of the local linear estimator under the exact same assumptions as the ones we maintained for the kernel regression model. The following are some preliminary asymptotic results that will prove useful later on:

Lemma 8.19 (Asymptotic Building Blocks of Local Linear Regressions) Suppose assumptions R1 to R6 hold true. Then, for any $x_0 \in \mathbb{R}^k$, we have

$$\frac{1}{nh_n^{k+2}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)' \xrightarrow{p} f(x_0) \int_{\mathbb{R}^k} zz'K(z)dz$$

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i \xrightarrow{p} O_{k\times 1}.$$

Furthermore, if $nh_n^{k+8} \to 0$ as $n \to \infty$, then for any function $\theta : \mathbb{R}^k \to \mathbb{R}$ that is three times continuously differentiable with derivatives of all orders bounded by $M > 0$, we have

$$\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\theta(X_i) - \sqrt{nh_n^k} \times h_n^2 \left(\int_{\mathbb{R}^k} zz'K(z)dz\right) \frac{\partial(\theta \times f)(x_0)}{\partial x} \xrightarrow{p} O_{k\times 1}.$$

$$\frac{1}{nh^{k+2}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i) \xrightarrow{p} \left(\int_{\mathbb{R}^k} zz'K(z)dz\right) \frac{\partial(\theta \times f)(x_0)}{\partial x}.$$

Proof) **Cross Product of Kernel and Squared Regressors**

To prove the first claim, note that, for any $x_0 \in \mathbb{R}^k$ and $h > 0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{nh^{k+2}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)(X_i - x_0)'\right] &= \frac{1}{h^{k+2}} \mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)(X_1 - x_0)(X_1 - x_0)'\right] \\
&= \frac{1}{h^{k+2}} \int_{\mathbb{R}^k} K\left(\frac{x_- x_0}{h}\right)(x - x_0)(x - x_0)' f(x)dx \\
&= \int_{\mathbb{R}^k} zz'K(z)f(x_0 + hz)dz.
\end{aligned}$$

Using the mean value inequality

$$|f(x_0 + hz) - f(x_0)| \leq hM|z|,$$

we can see that

$$\left|\int_{\mathbb{R}^k} zz'K(z)f(x_0 + hz)dz - f(x_0)\int_{\mathbb{R}^k} zz'K(z)dz\right| \leq h \underbrace{\left(M \int_{\mathbb{Z}} |z|^3 K(z)dz\right)}_{c_1}.$$

Now fix $1 \leq j, m \leq k$. We can immediately see that

$$\mathbb{E}\left|\frac{1}{h^k} K\left(\frac{X_1 - x_0}{h}\right)(X_{ij} - x_{0,j})(X_{im} - x_{0,m})\right|^2 = \frac{h^4}{h^k} \int_{\mathbb{R}^k} K(z)^2 z_j^2 z_m^2 f(x_0 + hz)dz,$$

so that

$$\frac{1}{n}\mathbb{E}\left|\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)(X_{1j}-x_{0,j})(X_{1m}-x_{0,m})\right|^2 \leq \frac{h^4}{nh^k}\underbrace{\left(\sup_{x\in\mathbb{R}^k} f(x)\right)\left(\int_{\mathbb{R}^k}|z|^4K(z)dz\right)}_{c_2}$$

where $\sup_{x\in\mathbb{R}^k} f(x) < +\infty$ because f is a continuous function that integrates to 1. Since

$$\begin{aligned}\operatorname{Var}&\left[\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)(X_{ij}-x_{0,j})(X_{im}-x_{0,m})\right] \\ &= \frac{1}{n}\mathbb{E}\left|\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)(X_{ij}-x_{0,j})(X_{im}-x_{0,m})\right|^2 \\ &\quad - \frac{1}{n}\left(\mathbb{E}\left[\frac{1}{h^k}K\left(\frac{X_1-x_0}{h}\right)(X_{ij}-x_{0,j})(X_{im}-x_{0,m})\right]\right)^2,\end{aligned}$$

we have

$$\begin{aligned}\operatorname{Var}&\left[\frac{1}{nh^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)(X_{ij}-x_{0,j})(X_{im}-x_{0,m})\right] \\ &\leq c_2\frac{h^3}{nh^k} + \frac{2}{nh}\left[f(x_0)^2\left(\int_{\mathbb{R}^k}|z|^2K(z)dz\right)^2 + c_1^2h^2\right].\end{aligned}$$

As such, under our assumptions,

$$\lim_{n\to\infty}\operatorname{Var}\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_{ij}-x_{0,j})(X_{im}-x_{0,m})\right] = 0$$

for $1 \leq j, m \leq k$ and

$$\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{nh_nb^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right] = f(x_0)\left(\int_{\mathbb{R}^k} zz'K(z)dz\right).$$

By the decomposition of mean squared error into the variance and squared bias terms, it now follows that

$$\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)' \xrightarrow{L^2} f(x_0)\left(\int_{\mathbb{R}^k} zz'K(z)dz\right).$$

Cross Product of Kernel, Regressors and Error

For any $x_0 \in \mathbb{R}^k$ and $h > 0$, we can consider the following mean value theorem for σ_0:

$$|\sigma_0(x_0+hz) - \sigma_0(x_0)| \leq hM|z|,$$

where we used the fact that the first derivative of σ_0 is bounded above by $M > 0$.

Now consider the term

$$\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i.$$

Its expected value is 0 from mean independence, so we focus on its variance, which is given as

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i\right] &= \frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)^2 (X_1 - x_0)(X_1 - x_0)' u_1^2\right] \\
&= \frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)^2 (X_1 - x_0)(X_1 - x_0)' \sigma_0(X_1)\right] \\
&= \frac{1}{h^k}\int_{\mathbb{R}^k} K\left(\frac{x - x_0}{h}\right)^2 (x - x_0)(x - x_0)' \sigma_0(x) dx \\
&= h^2 \int_{\mathbb{R}^k} zz' K(z)^2 \sigma_0(x_0 + hz) dz.
\end{aligned}$$

From the mean value theorem obtained above, we can see that

$$\left\| \mathrm{Var}\left[\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i\right] - h^2 \sigma_0(x_0)\left(\int_{\mathbb{R}^k} zz' K(z)^2 dz\right)\right\| \leq h^3 \cdot Mc\left(\int_{\mathbb{R}^k} |z|^3 K(z) dz\right).$$

It follows that

$$\begin{aligned}
&\left\| \mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right]\right\| \\
&\qquad \leq h_n^2 \cdot \sigma_0(x_0) c\left(\int_{\mathbb{R}^k} |z|^2 K(z) dz\right) + h_n^3 \cdot Mc\left(\int_{\mathbb{R}^k} |z|^3 K(z) dz\right)
\end{aligned}$$

for any $n \in N_+$, and because the right hand side goes to 0 as $n \to \infty$, we have

$$\lim_{n\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right] = O_{k\times k}.$$

Finally, using the fact that $\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i$ has mean 0, we can write

$$\mathbb{E}\left|\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i\right|^2 = \mathrm{tr}\left(\mathrm{Var}\left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right]\right).$$

Therefore,

$$\frac{1}{\sqrt{nh^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)u_i \xrightarrow{L^2} O_{k\times 1}.$$

Cross Product of Kernel, Regressors and Function

Now suppose that $nh_n^{k+8} \to 0$ as $n \to \infty$, and choose any $x_0 \in \mathbb{R}^k$. Define the functions $g_1, g_2 : \mathbb{R}^k \to \mathbb{R}$ as

$$g_1(x) = \theta(x)f(x), \quad g_2(x) = \theta(x)^2 f(x)$$

for any $x \in \mathbb{R}^k$. Note that g_2 admits the following mean value inequality:

$$|g_2(x_0 + hz) - g_2(x_0)| \le hM|z|$$

for any $h > 0$ and $z \in \mathbb{R}^k$.

In contrast to g_2, which requires only a first order expansion around x_0, we employ a third order expansion of g_1 around x_0: for any $h > 0$ and $z \in \mathbb{R}^k$,

$$g_1(x_0 + hz) = h \cdot z' \frac{\partial g_1(x_0)}{\partial x} + \frac{h^2}{2}\sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\frac{\partial^2 g_1(x_0)}{\partial x_{j_1}\partial x_{j_2}} z_{j_1} z_{j_2}$$

$$+ \frac{h^3}{3!}\sum_{j_1=1}^{k}\cdots\sum_{j_3=1}^{k}\frac{\partial^3 g_1(x_1^*(z))}{\partial x_{j_1}\cdots\partial x_{j_3}}\left(\prod_{i=1}^{3} z_{j_i}\right),$$

where $x_1^*(z)$ is some convex combination of $x_0 + hz$ and x_0. Since the third derivative is bounded above by $M > 0$, we can see that

$$\left| g_1(x_0 + hz) - h \cdot z' \frac{\partial g_1(x_0)}{\partial x} - \frac{h^2}{2}\sum_{j_1=1}^{k}\sum_{j_2=1}^{k}\frac{\partial^2 g_1(x_0)}{\partial x_{j_1}\partial x_{j_2}} z_{j_1} z_{j_2} \right| \le \frac{h^3}{3!} M k^3 |z|^3$$

for any $z \in \mathbb{R}^k$.

We use the familiar bias-variance computations to obtain the desired results. Starting with the bias,

$$\mathbb{E}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] = \frac{1}{h^k}\mathbb{E}\left[K\left(\frac{X_1 - x_0}{h}\right)(X_1 - x_0)\theta(X_1)\right]$$

$$= \frac{1}{h^k} \int_{\mathbb{R}^k} K\left(\frac{x-x_0}{h}\right)(x-x_0) g_1(x) dx$$

$$= h \int_{\mathbb{R}^k} zK(z) g_1(x_0+hz) dz.$$

Using the above decomposition, we have

$$\left| \int_{\mathbb{R}^k} zK(z) g_1(x_0+hz) dz - g(x_0) \int_{\mathbb{R}^k} zK(z) dz + h \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right.$$

$$\left. - \frac{h^2}{2} \sum_{j_1=1}^{k} \sum_{j_2=1}^{k} \frac{\partial^2 g_1(x_0)}{\partial x_{j_1} \partial x_{j_2}} \left(\int_{\mathbb{R}^k} z \cdot z_{j_1} z_{j_2} K(z) dz \right) \right|$$

$$\leq \frac{h^3}{3!} Mk^3 \left(\int_{\mathbb{R}^k} |z|^4 K(z) dz \right).$$

Because $K(\cdot)$ is an elementwise even function,

$$\int_{\mathbb{R}^k} zK(z) dz = O_{k \times 1},$$

$$\int_{\mathbb{R}^k} z \cdot z_{j_1} z_{j_2} K(z) dz = O_{k \times 1}$$

for any $1 \leq j_1, j_2 \leq k$, so that

$$\left| h \int_{\mathbb{R}^k} zK(z) g_1(x_0+hz) dz - h^2 \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right| \leq \frac{h^4}{3!} \left(Mk^4 \cdot \int_{\mathbb{R}^k} |z|^4 K(z) dz \right)$$

Defining $c_3 = \frac{Mk^4}{3!} \left(\int_{\mathbb{R}^k} |z|^4 K(z) dz \right) > 0$, we have

$$\left| h \int_{\mathbb{R}^k} zK(z) g_1(x_0+hz) dz - h^2 \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right| \leq c_3 h^4$$

and therefore

$$\left| \mathbb{E}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i) \right] - h^2 \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right| \leq c_3 h^4.$$

It follows that

$$\left| \mathbb{E}\left[\frac{1}{\sqrt{nh^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i) \right] - \sqrt{nh^{k+4}} \times \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right|$$

$$= \sqrt{nh^k} \left| \mathbb{E}\left[\frac{1}{nh^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i) \right] - h^2 \left(\int_{\mathbb{R}^k} zz' K(z) dz \right) \frac{\partial g_1(x_0)}{\partial x} \right| \leq c_3 \sqrt{nh^{k+8}}.$$

Moving onto the variance term, we have

$$\begin{aligned}
&\operatorname{Var}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] \\
&\qquad = \frac{1}{nh^{2k}}\operatorname{Var}\left[K\left(\frac{X_1 - x_0}{h}\right)(X_1 - x_0)\theta(X_1)\right] \\
&\qquad = \frac{1}{n}\mathbb{E}\left[\frac{1}{h^{2k}}K\left(\frac{X_1 - x_0}{h}\right)^2 (X_1 - x_0)(X_1 - x_0)'\theta(X_1)^2\right] \\
&\quad - \mathbb{E}\left[\frac{1}{h^k}K\left(\frac{X_1 - x_0}{h}\right)(X_1 - x_0)\theta(X_1)\right]\mathbb{E}\left[\frac{1}{h^k}K\left(\frac{X_1 - x_0}{h}\right)(X_1 - x_0)\theta(X_1)\right]'.
\end{aligned}$$

The squared term can be expanded as

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{h^{2k}}K\left(\frac{X_1 - x_0}{h}\right)^2 (X_1 - x_0)(X_1 - x_0)'\theta(X_1)^2\right] \\
&\qquad = \frac{1}{h^{2k}}\int_{\mathbb{R}^k} K\left(\frac{x - x_0}{h}\right)^2 (x - x_0)(x - x_0)' g_2(x)dx \\
&\qquad = \frac{h^2}{h^k}\int_{\mathbb{R}^k} K(z)^2 zz' g_2(x_0 + hz)dz.
\end{aligned}$$

The mean value inequality for g_2 shows us that

$$\left\|\int_{\mathbb{R}^k} K(z)^2 zz' g_2(x_0 + hz)dz - g_2(x_0)\int_{\mathbb{R}^k} zz'K(z)dz\right\| \leq h\underbrace{\left(Mc\int_{\mathbb{R}^k}|z|^3 K(z)dz\right)}_{c_2}.$$

As such,

$$\begin{aligned}
&\left\|\operatorname{Var}\left[\frac{1}{nh^k}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] - \frac{1}{nh^{k-2}}g_2(x_0)\int_{\mathbb{R}^k} zz'K(z)dz\right\| \\
&\qquad \leq c_2\frac{h^3}{nh^k} + \frac{1}{n}\left[Mh^2\left(\int_{\mathbb{R}^k}|z|^2 K(z)dz\right) + c_3 h^4\right]^2,
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\operatorname{Var}\left[\frac{1}{\sqrt{nh^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] - h^2 g_2(x_0)\int_{\mathbb{R}^k} zz'K(z)dz\right\| \\
&\qquad \leq c_2 h^3 + h^k\left[Mh^2\left(\int_{\mathbb{R}^k}|z|^2 K(z)dz\right) + c_3 h^4\right].
\end{aligned}$$

The above results hold for any $h > 0$ and $n \in N_+$, so

$$\mathbb{E}\left|\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\theta(X_i) - \sqrt{nh_n^{k+4}} \times \left(\int_{\mathbb{R}^k} zz'K(z)dz\right)\frac{\partial g_1(x_0)}{\partial x}\right|^2$$

$$= \text{Var}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\theta(X_i)\right]$$

$$+ \left(\mathbb{E}\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\theta(X_i)\right] - \sqrt{nh_n^{k+4}} \times \left(\int_{\mathbb{R}^k} zz' K(z)dz\right)\frac{\partial g_1(x_0)}{\partial x}\right)^2$$

$$\leq c_2 h_n^3 + h_n^k\left[Mh_n^2\left(\int_{\mathbb{R}^k}|z|^2 K(z)dz\right) + c_3 h_n^4\right] + c_3\sqrt{nh_n^{k+8}}.$$

By our assumptions, taking $n \to \infty$ on both sides shows us that

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\theta(X_i) - \sqrt{nh_n^{k+4}} \times \left(\int_{\mathbb{R}^k} zz' K(z)dz\right)\frac{\partial g_1(x_0)}{\partial x} = o_p(1).$$

Similarly, we can see that

$$\left|\mathbb{E}\left[\frac{1}{nh^{k+2}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] - \left(\int_{\mathbb{R}^k} zz' K(z)dz\right)\frac{\partial g_1(x_0)}{\partial x}\right| \leq c_3 h^2$$

and

$$\left\|\text{Var}\left[\frac{1}{nh^{k+2}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i)\right] - \frac{1}{nh^k} g_2(x_0)\int_{\mathbb{R}^k} zz' K(z)dz\right\|$$

$$\leq c_2\frac{h}{nh^k} + \frac{1}{n}\left[Mh\left(\int_{\mathbb{R}^k}|z|^2 K(z)dz\right) + c_3 h^3\right]^2$$

for any $h > 0$ and $n \in N_+$, so the same decomposition of mean squared error into the variance and squared bias shows us that

$$\frac{1}{nh^{k+2}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h}\right)(X_i - x_0)\theta(X_i) - \left(\int_{\mathbb{R}^k} zz' K(z)dz\right)\frac{\partial g_1(x_0)}{\partial x} = o_p(1).$$

Q.E.D.

8.4.3 Asymptotic Normality of the Local Linear Estimator

We now use the results from lemma 8.19 to prove an asymptotic normality theorem for the local linear estimator. It turns out that the local linear estimator has exactly the same asymptotic distribution as the Nadaraya-Watson estimator, but with a simplified bias term that does not involve the density f in any way. This is key to how the local linear estimator overcomes the boundary bias problem. We proceed by proving a series of lemmas, and then bringing them together at the end.

Lemma 8.20 (Probability Limit of the Denominator)
Suppose assumptions R1 to R6 hold true. Assume in addition that $nh_n^{k+8} \to 0$ as $n \to \infty$ and $\int_{\mathbb{R}^k} zz'K(z)dz$ is positive definite. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$,

$$\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n) = \frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right) - \frac{1}{nh_n^k}\left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]'$$
$$\times \left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1}\left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]$$

converges in probability to $f(x_0)$.

Proof) This follows easily; for any $n \in N_+$,

$$\frac{1}{nh_n^k}\left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]'\left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1}\left[\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]$$
$$= \left[\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]'$$
$$\times \left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1}\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right].$$

From the previous lemma, the positivity of $f(x_0)$ and the positive definiteness of $\int_{\mathbb{R}^k} zz'K(z)dz$, we can see that the last two terms here are $O_p(1)$. Meanwhile, we also know that

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0) - \sqrt{nh_n^k} \times h_n^2\left(\int_{\mathbb{R}^k} zz'K(z)dz\right)\frac{\partial f(x_0)}{\partial x} = o_p(1).$$

Premultiplying both sides by $\frac{1}{\sqrt{nh_n^k}}$ reveals that

$$\frac{1}{nh_n^k}\sum_{i=1}^n K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0) - h_n^2\left(\int_{\mathbb{R}^k} zz'K(z)dz\right)\frac{\partial f(x_0)}{\partial x} = o_p(1)$$

as well, since $nh_n^k \to +\infty$ as $n \to \infty$, and since $h_n \to 0$ as $n \to \infty$, we can conclude that

$$\frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0) = o_p(1).$$

Therefore,

$$\frac{1}{nh_n^k}\hat{B}_n(x_0; h_n) = \frac{1}{nh_n^k} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) + o_p(1),$$

and from what we already studied in the section on kernel density estimation,

$$\frac{1}{nh_n^k}\hat{B}_n(x_0; h_n) \xrightarrow{p} f(x_0).$$

Q.E.D.

Lemma 8.21 (Weak Limit of the Numerator)
Suppose assumptions R1 to R6 hold true. Assume in addition that $nh_n^{k+8} \to 0$ as $n \to \infty$ and $\int_{\mathbb{R}^k} zz'K(z)dz$ is positive definite. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$, we have

$$\begin{aligned}\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(u)}(x_0; h_n) &= \frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right) u_i - \frac{1}{\sqrt{nh_n^k}} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]' \\ &\quad \times \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right] \\ &\xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0)f(x_0)\left(\int_{\mathbb{R}^k} K(z)^2 dz\right)\right].\end{aligned}$$

Proof) This can be proven in a manner very similar to the result for the denominator. For any $n \in N_+$,

$$\begin{aligned}&\frac{1}{\sqrt{nh_n^k}} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]' \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1} \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right] \\ &\qquad = \left[\frac{1}{nh_n^{k+2}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]' \\ &\quad \times \left[\frac{1}{nh_n^{k+2}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1} \left[\frac{1}{\sqrt{nh_n^k}} \sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i\right].\end{aligned}$$

From lemma 8.19, the positivity of $f(x_0)$ and the positive definiteness of $\int_{\mathbb{R}^k} zz'K(z)dz$, we can see that the first two terms here are $O_p(1)$. Meanwhile, we also saw from lemma

8.19 that

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)u_i = o_p(1)$$

as $n \to \infty$. Therefore, the entire term above is asymptotically negligible, and we can write

$$\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(u)}(x_0; h_n) = \frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)u_i + o_p(1).$$

An application of lemma 8.16 now allows us to conclude from Slutsky's theorem that

$$\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(u)}(x_0; h_n) \xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0)f(x_0)\left(\int_{\mathbb{R}^k} K(z)^2 dz\right)\right].$$

Q.E.D.

Lemma 8.22 (Asymptotic Bias for Local Linear Regression)
Suppose assumptions R1 to R6 hold true. Assume in addition that $nh_n^{k+4} = O(1)$ as $n \to \infty$ and $\int_{\mathbb{R}^k} zz'K(z)dz$ is positive definite. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$,

$$\begin{aligned}
&\sqrt{nh_n^k}\left[\frac{1}{nh_n^k}\hat{A}_n^{(m)}(x_0; h_n) - m(x_0)\frac{1}{nh_n^k}\hat{B}_n(x_0; h_n) - f(x_0)b_n(x_0; h_n)\right] \\
&\qquad = \frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(m(X_i) - m(x_0)) \\
&\qquad - \frac{1}{\sqrt{nh_n^k}}\left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)\right]' \\
&\qquad \times \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(X_i - x_0)'\right]^{-1} \\
&\qquad \times \left[\sum_{i=1}^{n} K\left(\frac{X_i - x_0}{h_n}\right)(X_i - x_0)(m(X_i) - m(x_0))\right] \\
&\qquad - \sqrt{nh_n^k} \times f(x_0)b_n(x_0; h_n) \xrightarrow{p} O_{k\times 1},
\end{aligned}$$

where

$$b_n(x_0; h) = \frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 m(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2 K(z)dz\right)$$

for any $h > 0$.

Proof) Note first that

$$\frac{1}{\sqrt{nh_n^k}}\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\right]'\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]$$

$$=\left[\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right)(X_i-x_0)\right]'\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}$$
$$\times\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right].$$

From the preceding lemma, we showed that, for any function $\theta:\mathbb{R}^k\to\mathbb{R}$ that is thrice continuously differentiable with bounded derivative,

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\theta(X_i)-\sqrt{nh_n^k}\times h_n^2\left(\int_{\mathbb{R}^k} zz'K(z)dz\right)\frac{\partial(\theta\times f)(x_0)}{\partial x}\xrightarrow{p}O_{k\times 1}$$

$$\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\theta(X_i)\xrightarrow{p}\left(\int_{\mathbb{R}^k} zz'K(z)dz\right)\frac{\partial(\theta\times f)(x_0)}{\partial x}.$$

Both the function $\theta(x)=1$ and $\theta(x)=m(x)-m(x_0)$ satisfy the boundedness conditions. Furthermore, we also saw that

$$\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\xrightarrow{p}f(x_0)\left(\int_{\mathbb{R}^k} zz'K(z)dz\right).$$

Putting all this together, we have

$$\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)m(X_i)=O_p(1)$$

$$\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}=O_p(1),$$

so that

$$\frac{1}{\sqrt{nh_n^k}}\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\right]'\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]$$

$$-\sqrt{nh_n^k}\times h_n^2\left(\frac{\partial f(x_0)}{\partial x}\right)'\left(\int_{\mathbb{R}^k} zz'K(z)dz\right)'$$
$$\times\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^n K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]=o_p(1)$$

and

$$\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\frac{1}{nh_n^{k+2}}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]\xrightarrow{p}\frac{\partial m(x_0)}{\partial x}.$$

Under the maintained assumptions, $\sqrt{nh_n^k}\times h_n^2=\sqrt{nh_n^{k+4}}=O(1)$, so that we can finally conclude

$$\frac{1}{\sqrt{nh_n^k}}\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\right]'\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]$$
$$-\sqrt{nh_n^k}\times h_n^2\left(\frac{\partial f(x_0)}{\partial x}\right)'\left(\int_{\mathbb{R}^k}zz'K(z)dz\right)\left(\frac{\partial m(x_0)}{\partial x}\right)\xrightarrow{p}0.$$

Since

$$\int_{\mathbb{R}^k}z_jz_mK(z)dz=0$$

for any $1\leq j\neq m\leq k$, it follows that

$$\left(\frac{\partial f(x_0)}{\partial x}\right)'\left(\int_{\mathbb{R}^k}zz'K(z)dz\right)\left(\frac{\partial m(x_0)}{\partial x}f(x_0)\right)=\sum_{j=1}^{k}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)\frac{\partial f(x_0)}{\partial x_j}\cdot\frac{\partial m(x_0)}{\partial x_j}.$$

From the result we derived above,

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))$$
$$-\sqrt{nh_n^k}\times h_n^2\sum_{j=1}^{k}\left[\frac{f(x_0)}{2}\frac{\partial^2m(x_0)}{\partial x_j^2}+\frac{\partial m(x_0)}{\partial x_j}\frac{\partial f(x_0)}{\partial x_j}\right]\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)=o_p(1).$$

As such,

$$\frac{1}{\sqrt{nh_n^k}}\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(m(X_i)-m(x_0))$$
$$-\frac{1}{\sqrt{nh_n^k}}\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)\right]'\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(X_i-x_0)'\right]^{-1}\left[\sum_{i=1}^{n}K\left(\frac{X_i-x_0}{h_n}\right)(X_i-x_0)(m(X_i)-m(x_0))\right]$$
$$-\sqrt{nh_n^k}\times h_n^2\frac{f(x_0)}{2}\sum_{j=1}^{k}\frac{\partial^2m(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k}z_j^2K(z)dz\right)$$

conveges to 0 in probability.

Q.E.D.

Theorem 8.23 (Asymptotic Distribution of Local Linear Estimator)
Suppose assumptions R1 to R6 hold true. Assume in addition that $nh_n^{k+4} = O(1)$ as $n \to \infty$ and $\int_{\mathbb{R}^k} zz' K(z)dz$ is positive definite. Then, for any $x_0 \in \mathbb{R}^k$ such that $f(x_0) > 0$,

$$\sqrt{nh_n^k}\Big[\hat{m}_n(x_0;h_n) - m(x_0) - b_n(x_0;h_n)\Big] \xrightarrow{d} \mathcal{N}\left[0,\ \frac{\sigma_0(x_0)}{f(x_0)}\left(\int_{\mathbb{R}^k} K(z)^2 dz\right)\right],$$

where

$$b_n(x_0;h) = \frac{h^2}{2}\sum_{j=1}^{k}\frac{\partial^2 m(x_0)}{\partial x_j^2}\left(\int_{\mathbb{R}^k} z_j^2 K(z)dz\right)$$

for any $h > 0$.

Proof) Recall that we can decompose $\hat{m}_n(x_0;h_n)$ as

$$\begin{aligned}\hat{m}_n(x_0;h_n) &= \left(\hat{B}_n(x_0;h_n)\right)^{-1}\left[\hat{A}_n^{(m)}(x_0;h) + \hat{A}_n^{(u)}(x_0;h_n)\right] \\ &= \left(\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n)\right)^{-1}\left[\frac{1}{nh_n^k}\hat{A}_n^{(m)}(x_0;h_n) + \frac{1}{nh_n^k}\hat{A}_n^{(u)}(x_0;h_n)\right],\end{aligned}$$

so that

$$\begin{aligned}&\sqrt{nh_n^k}\Big[\hat{m}_n(x_0;h_n) - m(x_0) - b_n(x_0;h_n)\Big] \\ &= \left(\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n)\right)^{-1}\left[\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(m)}(x_0;h_n) - \sqrt{nh_n^k}\times f(x_0)b_n(x_0;h_n)\right] \\ &\quad + \left(\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n)\right)^{-1}\left[f(x_0) - \frac{1}{nh_n^k}\hat{B}_n(x_0;h_n)\right]\sqrt{nh_n^k}\times b_n(x_0;h_n) \\ &\quad + \left(\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n)\right)^{-1}\left(\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(u)}(x_0;h_n)\right).\end{aligned}$$

We have already established the following results:

$$\frac{1}{nh_n^k}\hat{B}_n(x_0;h_n) \xrightarrow{p} f(x_0)$$

$$\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(m)}(x_0;h_n) - \sqrt{nh_n^k}\times f(x_0)b_n(x_0;h_n) \xrightarrow{p} 0$$

$$\frac{1}{\sqrt{nh_n^k}}\hat{A}_n^{(u)}(x_0;h_n) \xrightarrow{d} \mathcal{N}\left[0,\ \sigma_0(x_0)f(x_0)\left(\int_{\mathbb{R}^k} K(z)^2 dz\right)\right].$$

Furthermore,

$$\sqrt{nh_n^k} \times b_n(x_0; h_n) = \sqrt{nh_n^{k+4}} \times \frac{1}{2} \sum_{j=1}^{k} \frac{\partial^2 m(x_0)}{\partial x_j^2} \left(\int_{\mathbb{R}^k} z_j^2 K(z) dz \right) = O(1)$$

as $n \to \infty$ because we assumed that $nh_n^{k+4} = O(1)$ as $n \to \infty$. Therefore,

$$\begin{aligned}
\sqrt{nh_n^k} \Big[\hat{m}_n(x_0; h_n) - m(x_0) - b_n(x_0; h_n) \Big] &= \left(\frac{1}{nh_n^k} \hat{B}_n(x_0; h_n) \right)^{-1} \left(\frac{1}{\sqrt{nh_n^k}} \hat{A}_n^{(u)}(x_0; h_n) \right) + o_p(1) \\
&\xrightarrow{d} \frac{1}{f(x_0)} \mathcal{N} \left[0, \, \sigma_0(x_0) f(x_0) \left(\int_{\mathbb{R}^k} K(z)^2 dz \right) \right] \\
&= \mathcal{N} \left[0, \, \frac{\sigma_0(x_0)}{f(x_0)} \left(\int_{\mathbb{R}^k} K(z)^2 dz \right) \right]
\end{aligned}$$

by Slutsky's theorem.

Q.E.D.

Because $b_n(x_0; h_n) \to 0$ and $nh_n^k \to +\infty$ as $n \to \infty$, the consistency of the local linear estimator follows immediately from the above result:

$$\hat{m}_n(x_0; h_n) \xrightarrow{p} m(x_0).$$

Notice how, under slightly stronger assumptions on the rate of convergence of nh_n^{k+4} (namely, assuming that it is stable instead of just assuming that $nh_n^{k+8} \to 0$ as $n \to \infty$), we have shown that the local linear estimator has the same asymptotic distribution as the Nadaraya-Watson estimator, but with a simpler asymptotic bias term. Indeed, the convergence rate condition that $nh_n^{k+4} = O(1)$ is satisfied by the Nadaraya-Watson estimator when h_n is equal to the asymptotically optimal bandwidth, so that the local linear estimator has a simpler asymptotic bias than even the asymptotically optimal Nadaraya-Watson estimator.

Note that the asymptotic bias of the local linear estimator does not involve the density f in any way. In particular, the term $\frac{1}{f(x_0)}$, which causes the bias of the Nadaraya-Watson estimator to explode as $f(x_0) \to 0$ and is the source of its boundary bias, is not present in the asymptotic bias for the local linear estimator. This is precisely the reason why the local linear estimator is an alternative to the Nadaraya-Watson estimator that does not suffer from boundary bias.

Chapter 9

Local Asymptotic Normality and Asymptotic Efficiency

So far, we have either dealt with completely parametric models, in which our object of interest is a finite-dimensional parameter, or non-parametric models in which we are interested in estimating the entire distribution function or density as infinite-dimensional parameters. In many cases we are interested in semiparametric models, where we want to estimate a finite-dimensional subset of an infinite-dimensional parameter. We have already encountered examples of semiparametric models and methods to estimate them; these include any linear regression model where we are agnostic to the error distribution, or GMM models where we impose restrictions on only a few moments.

The advantage of working explicitly with a semiparametric setup is that it allows us to obtain efficient estimators of the parameters of interest, while remaining robust to local misspecification because we do not fully parametrize the model. Recall that, in a fully parametric model, we showed that the MLE is the unqualified asymptotically efficient estimator, whereas the efficient GMM estimator was only shown to be efficient among the class of GMM estimators.

Below we first introduce contiguity and local asymptotic normality, which are preliminary concepts that are required to formalize the notion of semiparametric efficiency. Then, we derive semiparametrically efficient estimators for iid data and also for more general sequences of random variables that may be dependent.

9.1 Contiguity

We first introduce the concept of contiguity, and Le Cam's first and third lemmas. To motivate our interest in contiguity, suppose that we are interested in conducting a test with null $H_0 : \theta = \theta_0$, where the asymptotic distribution of the test statistic T_n is asymptotically normal and unbiased under the null. To evaluate this test, we need to compute its power against various alternative hypotheses. If T_n is constructed using a consistent estimator $\hat{\theta}_n$ of the parameter of interest, then we know that the test has good power against fixed alternative hypotheses $H_1 : \theta = \theta_1 \neq \theta_0$. Our interest, then, is in alternative hypotheses whose distance to θ_0 reduces as $n \to \infty$; such

alternative hypotheses are called local alternatives.

To evaluate how the test fares against local alternatives, we compute the asymptotic power of the test under these alternatives. Under each local alternative, we obtain a different sequence of distributions for the data compared to what we get under the null. We want to find the appropriate rate at which the local alternative converges to the true parameter θ_0 in order to be able to infer the asymptotic distribution of the test statistic T_n under the local alternative using the asymptotic properties of T_n under the null. This is where contiguity comes in. If there are two sequences of probability measures that are contiguous, then we can find the asymptotic distribution of a test statistic of interest under one sequence using the asymptotic properties of the statistic under the other. Since it is best viewed mathematically as an asymptotic version of the absolute continuity of measures, we first review absolute continuity, with a focus on likelihood ratios.

9.1.1 Absolute Continuity and Likelihood Ratios

Let $(E,\mathcal{E})$ be a measurable space. Recall that two measures μ and v on $(E,\mathcal{E})$ are said to be mutually singular (or orthogonal), denoted $\mu \perp v$, if they are concentrated on disjoint measurable subsets of E, that is, if there exist $E_1, E_2 \in \mathcal{E}$ such that $E_1 \cap E_2 = \emptyset$ and $\mu(A \cap E_1) = \mu(A)$ and $v(A \cap E_2) = v(A)$ for any $A \in \mathcal{E}$. If μ, v are probability measures, then $\mu \perp v$ if and only if there exist $E_1, E_2 \in \mathcal{E}$ such that $E_1 \cap E_2 = \emptyset$ and $\mu(E_1) = v(E_2) = 1$.

We first study the simple case when P,Q are dominated by some common σ-finite measure on $(E,\mathcal{E})$.

Lemma 9.1 (Properties of the Likelihood Ratio)
Let $(E,\mathcal{E})$ be a measurable space and P,Q two probability measures on $(E,\mathcal{E})$. Suppose that there exists a σ-finite measure μ on $(E,\mathcal{E})$ such that $Q,P << \mu$. Letting $p,q \in \mathcal{E}_+$ be the Radon-Nikodym derivatives of P and Q with respect to μ, the following hold true:

i) Defining $Q_a, Q_\perp : \mathcal{E} \to [0,1]$ as

$$Q_a(A) = Q(A \cap \{p > 0\}), \quad Q_\perp(A) = Q(A \cap \{p = 0\})$$

for any $A \in \mathcal{E}$,

$$Q = Q_a + Q_\perp, \quad Q_a << P, \quad Q_\perp \perp P.$$

ii) For any $A \in \mathcal{E}$,

$$Q_a(A) = \int_E h dP$$

for some $h \in \mathcal{E}_+$ such that $h = \frac{q}{p}$ P-a.e. on E.

iii) $Q << P$ if and only if $Q(\{p=0\})=0$, or equivalently, $Q_\perp(A)=0$ for any $A\in\mathcal{E}$.

Proof) i) For any $A\in\mathcal{E}$,

$$Q(A)=Q(A\cap\{p>0\})+Q(A\cap\{p=0\})=Q_a(A)+Q_\perp(A),$$

so that $Q=Q_a+Q_\perp$. It remains to show that $Q_a << P$ and $Q_\perp \perp P$.
If $P(A)=0$, then

$$P(A)=\int_A p d\mu=\int_E (p\cdot I_A)d\mu=0$$

and $p\cdot I_A=0$ μ-a.e. by the vanishing property for non-negative functions. Letting $p(x)I_A(x)=0$ for any $x\in E_0$, where $\mu(E_0^c)=0$, we have $A\cap\{p>0\}\subset E_0^c$. Since $\mu(E_0^c)=0$ and $Q<<\mu$, it follows that

$$Q_a(A)=Q(A\cap\{p>0\})\le Q(E_0^c)=0,$$

and we can see that $Q_a << P$. Now note that

$$P(\{p=0\})=\int_{\{p=0\}} p d\mu=0,$$

so that P is concentrated on $\{p>0\}$. Meanwhile, $Q_\perp$ is concentrated on $\{p=0\}$ by definition, so $Q_\perp\perp P$.

ii) Defining $h:E\to\mathbb{R}_+$ as

$$h(x)=\begin{cases}\frac{q(x)}{p(x)} & \text{if } p(x)>0\\ 0 & \text{if } p(x)=0\end{cases},$$

we can now see that

$$\int_A h dP=\int_{A\cap\{p>0\}} h dP=\int_{A\cap\{p>0\}} q d\mu$$

$$=Q(A\cap\{p>0\})=Q_a(A)$$

for any $A\in\mathcal{E}$. In particular,

$$h=\frac{q}{p}\quad P\text{-almost surely}$$

because $P(\{p>0\})=1$.

iii) Suppose that $Q << P$. Then, for any $A \in \mathcal{E}$ such that $P(A) = 0$,

$$Q(A) = Q_a(A) + Q_\perp(A) = 0,$$

so that $Q_\perp(A) = Q(A \cap \{p = 0\}) = 0$ as well. Since $P(\{p = 0\}) = 0$, it follows that $Q(\{p = 0\}) = 0$.
Conversely, suppose that $Q(\{p = 0\}) = 0$. Then, $Q_\perp(A) = Q(A \cap \{p = 0\}) \leq Q(\{p = 0\}) = 0$ for any $A \in \mathcal{E}$, so that $Q = Q_a << P$ on $\mathcal{E}$.

Q.E.D.

The Radon-Nikodym derivative h is called the likelihood ratio, since it is literally the ratio of the densities q and p whenever the quotient is well-defined. We denote it by $\frac{dQ}{dP}$, but we emphasize that it is actually the Radon-Nikodym derivative of Q_a with respect to P, or $\frac{\partial Q_a}{\partial P}$.

The Lebesgue decomposition is a clear generalization of the above results. Thus, the likelihood ratio can be defined even when there does not exist a dominating measure μ for P and Q.

Lemma 9.2 (The Lebesgue Decomposition and the Generalized Likelihood Ratio) Let $(E, \mathcal{E})$ be a measurable space and P, Q two probability measures on $(E, \mathcal{E})$. Then, the following hold true:

i) There exist a unique pair of measures Q_a and $Q_\perp$ on $(E, \mathcal{E})$ such that

$$Q = Q_a + Q_\perp, \quad Q_a << P, \quad Q_\perp \perp P.$$

ii) For any $A \in \mathcal{E}$,

$$Q_a(A) = \int_E h dP$$

for some $h \in \mathcal{E}_+$. This h is unique up to P-almost sure equivalence.

iii) $Q << P$ if and only if $Q = Q_a$.

Proof) Probability measures are σ-finite measures, so the Lebesgue-Radon-Nikodym theorem tells us that there exist a unique pair of measures Q_a and $Q_\perp$ on $(E, \mathcal{E})$ such that

$$Q = Q_a + Q_\perp, \quad Q_a << P, \quad Q_\perp \perp P,$$

and that there exists a non-negative $\mathcal{E}$-measurable function h such that

$$Q_a(A) = \int_A h dP$$

for any $A \in \mathcal{E}$. This h is clearly unique up to P-almost sure equivalence.

Now suppose that $Q << P$. Since $Q_\perp \perp P$, there exist $E_1, E_2 \in \mathcal{E}$ such that $E_1 \cap E_2 = \emptyset$ and, for any $A \in \mathcal{E}$, $Q_\perp(A \cap E_1) = Q_\perp(A)$ and $P(A \cap E_2) = P(A)$. Since $P(E_2^c) = P(E_2^c \cap E_2) = P(\emptyset) = 0$, it follows that

$$Q(E_2^c) = Q_a(E_2^c) + Q_\perp(E_2^c) = 0,$$

so that $Q_\perp(E_2^c) = 0$. By implication, $Q_\perp(E_2^c \cap E_1) = 0$, but since

$$Q_\perp(E_2^c \cap E_1) = Q_\perp(E_2 \setminus E_2) = Q_\perp(E_1) = Q_\perp(E),$$

it follows that $Q_\perp(E) = 0$ as well. This shows us that $Q_\perp(A) = 0$ for any $A \in \mathcal{E}$, so that $Q = Q_a$.

On the other hand, suppose that $Q = Q_a$ on $\mathcal{E}$. Then, $Q = Q_a << P$ trivially by what we proved above.

Q.E.D.

We denote the representative element of the P-almost sure equivalence class

$$\{h \in \mathcal{E}_+ \mid Q_a(A) = \int_A h dP \quad \forall A \in \mathcal{E}\}$$

as $\frac{dQ}{dP}$, and call it the generalized likelihood ratio of Q to P. We can now furnish conditions that are equivalent to absolute continuity in terms of generalized likelihood ratios:

Corollary to Lemma 9.2 (Characterization via the Generalized Likelihood Ratio) Let $(E, \mathcal{E})$ be a measurable space and P, Q two probability measures on $(E, \mathcal{E})$. Let $Q = Q_a + Q_\perp$ be the Lebesgue decomposition of Q with respect to P, and $P = P_a + P_\perp$ the Lebesgue decomposition of P with respect to Q. Then,

$$Q_a(A) = Q\left[A \cap \left\{\frac{dP}{dQ} > 0\right\}\right], \quad Q_\perp(A) = Q\left[A \cap \left\{\frac{dP}{dQ} = 0\right\}\right]$$

for any $A \in \mathcal{E}$. By implication,

$$Q << P \quad \text{if and only if} \quad Q\left[\left\{\frac{dP}{dQ} = 0\right\}\right] = 0$$

$$\text{if and only if} \quad \int_E \frac{dQ}{dP} dP = 1.$$

Proof) Let $\mu = \frac{1}{2}(P+Q)$. Then, μ is a probability measure on $(E,\mathcal{E})$ that dominates both P and Q; by the Radon-Nikodym theorem, there exist $\mathcal{E}$-measurable non-negative functions p,q such that $p = \frac{\partial P}{\partial \mu}$ and $q = \frac{\partial Q}{\partial \mu}$. By lemma 9.1 and the uniqueness of the Lebesgue decomposition,

$$Q_a(A) = Q(A \cap \{p = 0\}), \quad Q_\perp(A) = Q(A \cap \{p > 0\})$$

for any $A \in \mathcal{E}$, and

$$\frac{dP}{dQ} = \frac{p}{q} \quad Q\text{- a.e. on } E.$$

Note that $\{p > 0\} \cap \{q > 0\} = \{p/q > 0\}$; therefore, for any $A \in \mathcal{E}$,

$$Q_a(A) = Q(A \cap \{p > 0\}) = Q(A \cap \{p > 0\} \cap \{q > 0\}) = Q(A \cap \{p/q > 0\}) = Q\left[A \cap \left\{\frac{dP}{dQ} > 0\right\}\right].$$

It follows immediately that

$$Q_\perp(A) = Q(A) - Q_a(A) = Q(A) - Q\left[A \cap \left\{\frac{dP}{dQ} > 0\right\}\right] = Q\left[A \cap \left\{\frac{dP}{dQ} = 0\right\}\right]$$

for any $A \in \mathcal{E}$.

If $Q << P$ then $Q = Q_a$, so that $\frac{dQ}{dP}$ is the Radon-Nikodym derivative of Q with respect to P and thus

$$\int_E \frac{dQ}{dP} dP = Q(E) = 1.$$

On the other hand, if $\int_E \frac{dQ}{dP} dP = 1$, then because this implies $Q_a(E) = 1$, we have $Q_\perp(E) = 0$ and therefore $Q = Q_a$, which implies $Q << P$.

Furthermore, $Q << P$ if and only if $Q = Q_a$, which holds if and only if

$$Q_\perp(E) = Q\left[\left\{\frac{dP}{dQ} = 0\right\}\right] = 0.$$

Q.E.D.

Our interest lies in the case where $Q << P$ and we know the distribution of the statistic (read: $\mathcal{E}$-measurable function) $\mathcal{T} : E \to \mathbb{R}^k$ under the probability measure P, that is, we know the probability measure $P \circ \mathcal{T}^{-1}$. Can we use this information to infer the distribution of $\mathcal{T}$ under Q? Note that knowledge of the distribution $Q \circ \mathcal{T}^{-1}$ of $\mathcal{T}$ under Q is equivalent to knowledge of

the integral

$$\int_E (f \circ \mathcal{T}) dQ := \int_E f d(Q \circ \mathcal{T}^{-1})$$

for any non-negative Borel measurable function $f(\cdot)$. Letting $\frac{dQ}{dP}$ be the likelihood ratio of Q to P, since $Q = Q_a$ due to the fact that $Q << P$, $\frac{dQ}{dP}$ is the Radon-Nikodym derivative of Q with respect to P. As such, it follows that

$$\int_E (f \circ \mathcal{T}) dQ = \int_E (f \circ \mathcal{T}) \frac{dQ}{dP} dP,$$

where we used the fact that $f \circ \mathcal{T}$ is a $\mathcal{E}$-measurable non-negative function.

Suppose that we know the joint distribution of $\left(\mathcal{T}, \frac{dQ}{dP}\right)$ under P, that is, we know the pushforward measure

$$v = P \circ \left(\mathcal{T}, \frac{dQ}{dP}\right)^{-1}$$

taking values in $\mathbb{R}^{k+1}$. Then, for any $B \in \mathcal{B}(\mathbb{R}^k)$,

$$\begin{aligned} Q(\mathcal{T} \in B) &= \int_E (I_B \circ \mathcal{T}) \, dQ \\ &= \int_E (I_B \circ \mathcal{T}) \, \frac{dQ}{dP} dP \\ &= \int_{\mathbb{R}^{k+1}} (y \cdot I_B(x)) \, dv(x, y) \\ &= \int_{B \times \mathbb{R}} y \, dv(x, y), \end{aligned}$$

where we used the formula for integrals of pushforward measures to justify the third equality. In other words, knowledge of the distribution of $\left(\mathcal{T}, \frac{dQ}{dP}\right)$ under P is sufficient to recover the distribution of $\mathcal{T}$ under Q in the case that $Q << P$.

In practice we are more interested in recovering the weak limit of a sequence of distributions corresponding to a sequence of statistics rather than the distribution of a single statistic. This requires an asymptotic notion of absolute continuity to hold between two sequences of distributions, similarly to how we required the absolute continuity relationship $Q << P$ to hold between the two distributions Q and P. This asymptotic relationship is called contiguity, and we define it below.

9.1.2 Contiguity and Le Cam's First Lemma

Let $\{(E_n, \mathcal{E}_n)\}_{n \in N_+}$ be a sequence of measurable functions and $\{P_n\}_{n \in N_+}$, $\{Q_n\}_{n \in N_+}$ sequences such that P_n, Q_n are probability measures on $(E_n, \mathcal{E}_n)$ for each $n \in N_+$. We say that Q_n is contiguous with respect to P_n, which we denote by $Q_n \lhd P_n$, if:

For any sequence $\{A_n\}_{n\in N_+}$ such that $A_n \in \mathcal{E}_n$ for each $n \in N_+$,

$$P_n(A_n) \to 0 \quad \text{implies} \quad Q_n(A_n) \to 0$$

as $n \to \infty$.

How would we begin to characterize contiguity? Seeing as how it is an asymptotic version of absolute continuity, a natural approach is to furnish a characterization based on the limit of the sequences $\{P_n\}_{n\in N_+}$ and $\{Q_n\}_{n\in N_+}$. The choice of the limit becomes clear in light of the following lemma.

Lemma 9.3 (Weak Limits of Likelihood Ratios)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces and $\{P_n\}_{n\in N_+}$ and $\{Q_n\}_{n\in N_+}$ sequences such that P_n, Q_n are probability measures on $(E_n, \mathcal{E}_n)$. Then, every subsequence of the sequence

$$\left\{Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}_{n\in N_+}$$

of probability measures on $\mathbb{R}$ has a further subsequence that converges weakly to some probability measure μ on $\mathbb{R}$ that assigns probability 1 to $[0, +\infty)$.

Likewise, every subsequence of

$$\left\{P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n\in N_+}$$

has a further subsequence that converges weakly to some probability measure v on $\mathbb{R}$ that assigns probability 1 to $[0, +\infty)$.

Proof) We will prove that any subsequence of

$$\left\{\mu_n = Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}_{n\in N_+}$$

has a further subsequence that converges weakly to some probability measure μ on $\mathbb{R}$, since the second statement follows immediately from this claim by symmetry.

Recall from the Lebesgue decomposition that, for any $n \in N_+$,

$$P_n = P_n^a + P_n^\perp,$$

where $P_n^a << Q_n$ and $P_n^\perp \perp Q_n$, and that $\frac{dP_n}{dQ_n} = \frac{\partial P_n^a}{\partial Q_n}$. Note that

$$\int_{E_n} \frac{dP_n}{dQ_n} dQ_n = \int_{E_n} \frac{\partial P_n^a}{\partial Q_n} dQ_n = P_n^a(E_n) \leq P_n(E_n) = 1$$

by the definition of the Radon-Nikodym derivative of P_n^a with respect to Q_n. For any

$k \in N_+$,

$$1 \geq \int_{E_n} \frac{dP_n}{dQ_n} dQ_n = \int_{\left\{\frac{dP_n}{dQ_n} > k\right\}} \frac{dP_n}{dQ_n} dQ_n + \int_{\left\{\frac{dP_n}{dQ_n} \leq k\right\}} \frac{dP_n}{dQ_n} dQ_n$$

$$\geq \int_{\left\{\frac{dP_n}{dQ_n} > k\right\}} \frac{dP_n}{dQ_n} dQ_n \geq k \cdot Q_n \left[\left\{\frac{dP_n}{dQ_n} > k\right\}\right] = k \cdot \mu_n((k, +\infty)),$$

or equivalently,

$$\mu_n((k, +\infty)) \leq \frac{1}{k}.$$

This holds for any $n, k \in N_+$, and since each μ_n is concentrated on $[0, +\infty)$ (since any likelihood ratio is non-negative valued),

$$\inf_{n \in N_+} \mu_n([0, k]) \geq 1 - \frac{1}{k}$$

for any $k \in N_+$. Thus, for any $\epsilon > 0$, letting $N \in N_+$ be chosen so that $\frac{1}{N} < \epsilon$,

$$\inf_{n \in N_+} \mu_n([0, N]) \geq 1 - \frac{1}{N} > 1 - \epsilon.$$

In other words, $[0, N]$ is a compact set to which every measure on $\{\mu_n\}_{n \in N_+}$ assigns measure at least $1 - \epsilon$. By definition, $\{\mu_n\}_{n \in N_+}$ is a sequence of tight probability measures, and Prohorov's theorem implies that it is relatively compact, that is, every subsequence of $\{\mu_n\}_{n \in N_+}$ contains a further subsequence that converges weakly to some probability measure μ on $\mathbb{R}$.

It remains to show that μ is concentrated on $[0, +\infty)$. This follows easily from Prohorov's theorem; letting $\{\mu_{n_k}\}_{k \in N_+}$ be the subsequence of $\{\mu_n\}_{n \in N_+}$ that converges weakly to μ, since $[0, +\infty)$ is a closed subset of $\mathbb{R}$ and $\mu_{n_k}([0, +\infty)) = 1$ for any $k \in N_+$,

$$1 = \limsup_{k \to \infty} \mu_{n_k}([0, +\infty)) \leq \mu([0, +\infty)),$$

so that $\mu([0, +\infty)) = 1$.

Q.E.D.

Let P, Q be two probability measures on some measurable space $(E, \mathcal{E})$. Then, we saw from the corollary to lemma 9.2. that the following are equivalent:

$$Q << P, \quad Q\left[\left\{\frac{dP}{dQ} = 0\right\}\right] = 0, \quad \int_E \frac{dQ}{dP} dP = 1.$$

Le Cam's first lemma shows that contiguity can be characterized using the asymptotic versions of the above conditions: specifically, letting μ be the weak limit of some subsequence of $\left\{Q_n \circ\right.$

$\left(\frac{dP_n}{dQ_n}\right)^{-1}\Big\}_{n\in N_+}$ and v the weak limit of some subsequence of $\left\{P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n\in N_+}$, we will show that

$$Q_n \lhd P_n \quad \text{if and only if} \quad \mu(\{0\}) = 0 \quad \text{if and only if} \quad \int_{\mathbb{R}} x\, dv(x) = 1.$$

We start with a preliminary result relating to weak convergence.

Lemma 9.4 Let (E,d) be a metric space and $\mathcal{E}$ the corresponding Borel σ-algebra. Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence of probability measures on $(E,\mathcal{E})$. If $\mu_n \to \mu$ weakly, then for any non-negative continuous function f on E,

$$\liminf_{n\to\infty} \int_E f d\mu_n \geq \int_E f d\mu.$$

Conversely, if

$$\liminf_{n\to\infty} \int_E f d\mu_n \geq \int_E f d\mu$$

for any non-negative continuous function f on E and every open subset of E is σ-compact, that is, any open subset of E can be expressed as the countable union of compact subsets of E, then $\mu_n \to \mu$ weakly.

Proof) **Necessity**

Construct the sequence $\{f_m\}_{m\in N_+}$ of non-negative continuous and bounded functions as

$$f_m(x) = \min(f(x), m)$$

for any $x \in E$ and $m \in N_+$. Then, $|f_m| = f_m \leq m$ on E for any $m \in N_+$ and $f_m \nearrow f$ pointwise. By the MCT, we have

$$\lim_{m\to\infty} \int_E f_m d\mu = \int_E f d\mu.$$

Choose any $M \in \mathbb{R}$ such that $\int_E f d\mu > M$. Then,

$$M < \int_E f d\mu = \sup_{m\in N_+} \int_E f_m d\mu,$$

so that there exists some $N \in N_+$ such that $M < \int_E f_N d\mu$. By the definition of weak convergence,

$$\int_E f_N d\mu = \lim_{n\to\infty} \int_E f_N d\mu_n,$$

and because

$$\int_E f_N d\mu_n \leq \int_E f d\mu_n$$

for any $n \in N_+$ by the fact that $f_N \leq f$ on E, we can see that

$$\lim_{n\to\infty} \int_E f_N d\mu_n \leq \liminf_{n\to\infty} \int_E f d\mu_n.$$

As such,

$$M < \liminf_{n\to\infty} \int_E f d\mu_n.$$

This holds for any $M < \int_E f d\mu$, so it must be the case that

$$\int_E f d\mu \leq \liminf_{n\to\infty} \int_E f d\mu_n.$$

Sufficiency

Suppose that every open subset of E is σ-compact, and that

$$\int_E f d\mu \leq \liminf_{n\to\infty} \int_E f d\mu_n$$

for any non-negative continuous function f on E.

Because every metric space is a locally compact Hausdorff space, and the measure of any Borel measurable subset of E has finite measure under the probability measure μ, it follows that μ is a regular Borel measure[1]. In particular, it is inner Borel regular, so that the measure of any open set under μ can be approximated from beneath by compact sets:

$$\mu(V) = \sup\{\mu(K) \mid K \subset V,\ K \text{ is compact}\}$$

for any open set V.

Choose any open set V and $\epsilon > 0$. Then, by inner regularity, there exists some compact subset K of E such that $K \subset V$ and

$$\mu(V) - \epsilon < \mu(K) \subset \mu(V).$$

Then, by Urysohn's lemma, there exists a continuous function $f : E \to [0, 1]$ such that

[1]This is the content of theorem 4.4 in the measure theory text.

$I_K \leq f \leq I_V$. Since f is a continuous non-negative function, by assumption

$$\liminf_{n\to\infty} \int_E f d\mu_n \geq \int_E f d\mu.$$

The monotonicity of integration tells us that $\int_E f d\mu_n \leq \mu_n(V)$ for any $n \in N_+$ and $\int_E f d\mu \geq \mu(K) > \mu(V) - \epsilon$, so that

$$\liminf_{n\to\infty} \mu_n(V) \geq \mu(V) - \epsilon.$$

This holds for any $\epsilon > 0$, so we have

$$\liminf_{n\to\infty} \mu_n(V) \geq \mu(V).$$

This in turn holds for any open subset V of E, so by the Portmanteau theorem, $\mu_n \to \mu$ weakly as $n \to \infty$.

Q.E.D.

Theorem 9.5 (Le Cam's First Lemma)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces and $\{P_n\}_{n\in N_+}$, $\{Q_n\}_{n\in N_+}$ sequences such that P_n, Q_n are probability measures on $(E_n, \mathcal{E}_n)$ for each $n \in N_+$. Then, the following are equivalent:

i) $Q_n \lhd P_n$.

ii) If there exists a subsequence of $\left\{Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}_{n\in N_+}$ that converges weakly to some probability measure μ on $\mathbb{R}$, then $\mu((0, +\infty)) = 1$.

iii) If there exists a subsequence of $\left\{P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n\in N_+}$ that converges weakly to some probability measure v on $\mathbb{R}$, then

$$\int_{\mathbb{R}} x \, dv(x) = 1.$$

iv) For any sequence $\{\mathcal{T}_n\}_{n\in N_+}$ such that each $\mathcal{T}_n : E_n \to \mathbb{R}^k$ is $\mathcal{E}_n$-measurable,

$$\lim_{n\to\infty} P_n[\{|\mathcal{T}_n| > \epsilon\}] = 0 \quad \text{for any } \epsilon > 0$$

implies

$$\lim_{n\to\infty} Q_n[\{|\mathcal{T}_n| > \epsilon\}] = 0 \quad \text{for any } \epsilon > 0.$$

Proof) **i) $\Leftrightarrow$ iv)**

The equivalence of i) and iv) is immediately evident. Suppose $Q_n \lhd P_n$, and that

$$\lim_{n\to\infty} P_n[\{|\mathcal{T}_n| > \epsilon\}] = 0 \quad \text{for any } \epsilon > 0.$$

Then, for any $\epsilon > 0$, the sequence $\{A_n = \{|\mathcal{T}_n| > \epsilon\}\}_{n\in N_+}$ is a sequence of sets such that $A_n \in \mathcal{E}_n$ for each $n \in N_+$ and $P_n(A_n) \to 0$, so that $Q_n(A_n) \to 0$ by contiguity, or equivalently,

$$\lim_{n\to\infty} Q_n[\{|\mathcal{T}_n| > \epsilon\}] = 0.$$

Conversely, suppose that iv) holds. Then, choose any sequence $\{A_n\}_{n\in N_+}$ of sets such that $A_n \in \mathcal{E}_n$ for any $n \in N_+$ and $P_n(A_n) \to 0$. Defining $\mathcal{T}_n = I_{A_n}$ for each $n \in N_+$, it follows that $\{\mathcal{T}_n\}_{n\in N_+}$ is a sequence such that each $\mathcal{T}_n : E_n \to \mathbb{R}^k$ is $\mathcal{E}_n$-measurable and, since

$$\{|\mathcal{T}_n| > \epsilon\} = \{I_{A_n} > \epsilon\} = A_n$$

for any $\epsilon > 0$ and $n \in N_+$,

$$\lim_{n\to\infty} P_n[\{|\mathcal{T}_n| > \epsilon\}] = 0 \quad \text{for any } \epsilon > 0.$$

It follows that

$$\lim_{n\to\infty} Q_n[\{|\mathcal{T}_n| > \epsilon\}] = 0 \quad \text{for any } \epsilon > 0,$$

or equivalently, $Q_n(A_n) \to 0$ as $n \to \infty$. This shows us that $Q_n \lhd P_n$.

Now we show that i) implies ii) implies iii) implies i), at which point the proof will be completed.

i) $\Rightarrow$ ii)

Suppose that i) holds, and that there exists a subsequence of $\left\{Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}_{n\in N_+}$ that converges weakly to some probability measure μ on $\mathbb{R}$. For notational simplicity,

let $\left\{Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}_{n\in N_+}$ itself be this weakly convergent subsequence. Denote $\mu_n = Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}$ for each n.

First, we show that $\mu([0,+\infty)) = 1$. This is simple, since $[0,+\infty)$ is a closed subset of $\mathbb{R}$ such that $\mu_n([0,+\infty) = 1$ for each n; thus, by the Portmanteau theorem,

$$1 = \limsup_{n\to\infty} \mu_n([0,+\infty)) \leq \mu([0,+\infty)),$$

which implies that $\mu([0,+\infty)) = 1$.

Now choose any $\epsilon > 0$. Since $(-\infty,\epsilon)$ is an open subset of $\mathbb{R}$, by the Portmanteau theorem

$$\liminf_{n\to\infty} \mu_n([0,\epsilon)) = \liminf_{n\to\infty} \mu_n((-\infty,\epsilon)) \geq \mu((-\infty,\epsilon)) = \mu([0,\epsilon)).$$

In particular, for any $k \in N_+$,

$$\liminf_{n\to\infty} \mu_n\Big[[0,\ 1/k)\Big] \geq \mu\Big[[0,\ 1/k)\Big].$$

This allows us to construct a subsequence $\{N_k\}_{k\in N_+}$ of N_+ such that

$$\mu_n\Big[[0,\ 1/k)\Big] - \mu\Big[[0,\ 1/k)\Big] \geq -\frac{1}{k}$$

for any $n \geq N_k$. Define the sequence $\{\epsilon_n\}_{n\in N_+}$ as

$$\epsilon_n = \frac{1}{k} \quad \text{if } N_k \leq n < N_{k+1}.$$

Then, $\epsilon_n \searrow 0$ as $n \to \infty$ and, for any $n \in N_+$, letting $N_k \leq n < N_{k+1}$ we have

$$\mu_n\Big[[0,\epsilon_n)\Big] - \mu\Big[[0,\epsilon_n)\Big] = \mu_n\Big[[0,\ 1/k)\Big] - \mu\Big[[0,\ 1/k)\Big] \geq -\frac{1}{k} = -\epsilon_n.$$

This holds for any $n \in N_+$, so

$$\liminf_{n\to\infty} \Big(\mu_n\Big[[0,\epsilon_n)\Big] - \mu\Big[[0,\epsilon_n)\Big]\Big) \geq 0.$$

By the Lebesgue decomposition lemma, for each $n \in N_+$ we can decompose P_n as

$$P_n = P_n^a + P_n^\perp,$$

where $P_n^a << Q_n$ and $\frac{dP_n}{dQ_n} = \frac{\partial P_n^a}{\partial Q_n}$, and $P_n^\perp \perp Q_n$, so that there exists a measurable

subset $E_n^Q \in \mathcal{E}$ such that $Q_n(E_n^Q) = 1$ for any $A \in \mathcal{E}_n$ and $P_n^{\perp}(E_n^Q) = 0$. Therefore,

$$\begin{aligned}\mu_n\Big[[0,\epsilon_n)\Big] &= Q_n\left[\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\}\right] \\ &= Q_n\left[\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\} \cap E_n^Q\right]\end{aligned}$$

and

$$\begin{aligned}P_n\left[\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\} \cap E_n^Q\right] &= P_n^a\left[\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\} \cap E_n^Q\right] \\ &= \int_{\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\} \cap E_n^Q} \frac{dP_n}{dQ_n} dQ_n \leq \epsilon_n.\end{aligned}$$

for any $n \in N_+$. Since $\epsilon_n \searrow 0$,

$$\lim_{n\to\infty} P_n\left[\Big\{x \in E_n \mid \frac{dP_n}{dQ_n} < \epsilon_n\Big\} \cap E_n^Q\right] = 0,$$

and by contiguity we have

$$\lim_{n\to\infty} \mu_n\Big[[0,\epsilon_n)\Big] = 0.$$

Sequential continuity tells us that

$$\lim_{n\to\infty} \mu\Big[[0,\epsilon_n)\Big] = \mu(\{0\}),$$

so

$$\begin{aligned}0 &\leq \liminf_{n\to\infty} \Big(\mu_n\Big[[0,\epsilon_n)\Big] - \mu\Big[[0,\epsilon_n)\Big]\Big) \\ &= \lim_{n\to\infty} \mu_n\Big[[0,\epsilon_n)\Big] - \lim_{n\to\infty} \mu\Big[[0,\epsilon_n)\Big] = -\mu(\{0\}),\end{aligned}$$

which implies that $\mu(\{0\}) = 0$. Therefore,

$$\mu((0,+\infty)) = \mu([0,+\infty)) - \mu(\{0\}) = 1.$$

ii) ⇒ iii)

Suppose that ii) holds, and that some subsequence of

$$\Big\{\mu_n = P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1}\Big\}_{n\in N_+}$$

converges weakly to some probability measure μ on $\mathbb{R}$; for notational simplicity, we let $\{\mu_n\}_{n\in N_+}$ itself be the weakly convergent subsequence. We want to show that

$$\int_{\mathbb{R}} x \, d\mu(x) = 1.$$

Define $R_n = \frac{1}{2}(P_n + Q_n)$ for each $n \in N_+$. Then, R_n is a probability measure on $(E_n, \mathcal{E}_n)$ such that $Q_n, P_n << R_n$, so that

$$P_n(A) = \int_A \frac{dP_n}{dR_n} dR_n, \quad Q_n(A) = \int_A \frac{dQ_n}{dR_n} dR_n$$

for any $A \in \mathcal{E}_n$. It follows that

$$2R_n(A) = P_n(A) + Q_n(A) = \int_A \left[\frac{dP_n}{dR_n} + \frac{dQ_n}{dR_n}\right] dR_n$$

and thus

$$\int_A \left[\frac{dP_n}{dR_n} + \frac{dQ_n}{dR_n} - 2\right] dR_n = 0$$

for any $A \in \mathcal{E}_n$; it follows that $\frac{dP_n}{dR_n} + \frac{dQ_n}{dR_n} = 2$ R_n-a.e. We choose versions $\frac{dP_n}{dR_n}$ and $\frac{dQ_n}{dR_n}$ of the densities so that this equation holds exactly. By implication, $\frac{dP_n}{dR_n}$ and $\frac{dQ_n}{dR_n}$ take values in the compact set $[0, 2]$, so that

$$\left\{v_n = R_n \circ \left(\frac{dP_n}{dR_n}\right)^{-1}\right\}$$

is a tight sequence of probability measures on $\mathbb{R}$. We showed in lemma 9.3. that

$$\left\{\lambda_n = Q_n \circ \left(\frac{dP_n}{dQ_n}\right)^{-1}\right\}$$

is also a tight sequence of probability measures on $\mathbb{R}$.

By Prohorov's theorem, there exist subsequences

$$\{\lambda_{n_k}\}_{k\in N_+} \quad \text{and} \quad \{v_{n_k}\}_{k\in N_+}$$

that converge weakly to probability measures λ and v on $\mathbb{R}$. Because $v_{n_k}([0, 2]) = 1$ for each $k \in N_+$, defining $M : \mathbb{R} \to \mathbb{R}$ as

$$M(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \in [0, 2] \\ 2 & \text{if } x > 2 \end{cases},$$

M is a bounded and continuous function on $\mathbb{R}$ such that

$$\int_{\mathbb{R}} x\, dv_{n_k}(x) = \int_{\mathbb{R}} M(x)\, dv_{n_k}(x) \to \int_{\mathbb{R}} M(x) dv(x) = \int_{\mathbb{R}} x\, dv(x)$$

as $k \to \infty$ by the definition of weak convergence. Note that

$$\int_{\mathbb{R}} x\, dv_n(x) = \int_{E_n} \frac{dP_n}{dR_n} dR_n = 1$$

for each $n \in N_+$ by the formula for integrals with respect to pushforward measures; it follows that

$$\int_{\mathbb{R}} x\, dv(x) = 1.$$

as well.

Now choose any $f \in C_b(\mathbb{R}, \mathbb{R})$, and define $g : [0,2] \to \mathbb{R}$ as

$$g(x) = \begin{cases} \frac{x}{2-x} & \text{if } x \in [0,2) \\ 0 & \text{otherwise} \end{cases}.$$

For any $n \in N_+$, from lemma 9.1 we can infer that, Q_n-a.e.,

$$\frac{dP_n}{dQ_n} = \frac{dP_n}{dR_n} \Big/ \frac{dQ_n}{dR_n} = \frac{\frac{dP_n}{dR_n}}{2 - \frac{dP_n}{dR_n}},$$

so that

$$\frac{dP_n}{dQ_n}(x) = g\left(\frac{dP_n}{dR_n}(x)\right)$$

for Q_n-a.e. $x \in E_n$. Therefore,

$$\begin{aligned} \int_{E_n} f\left(\frac{dP_n}{dQ_n}\right) dQ_n &= \int_{E_n} f\left(g\left(\frac{dP_n}{dR_n}\right)\right) dQ_n \\ &= \int_{E_n} f\left(g\left(\frac{dP_n}{dR_n}\right)\right) \frac{dQ_n}{dR_n} dR_n \\ &= \int_{E_n} f\left(g\left(\frac{dP_n}{dR_n}\right)\right) \left(2 - \frac{dP_n}{dR_n}\right) dR_n \\ &= \int_0^2 f(g(r))(2-r) dv_n(r) = \int_{\mathbb{R}} F(r) dv_n(r) \end{aligned}$$

for any $n \in N_+$, where we define $F : \mathbb{R} \to \mathbb{R}$ as

$$F(r) = \begin{cases} 2f(0) & \text{if } r < 0 \\ f(g(r))(2-r) & \text{if } r \in [0,2] \\ 0 & \text{if } r > 2 \end{cases}.$$

F is a bounded and continuous function on $\mathbb{R}$, so it follows from the definition of weak convergence that

$$\int_{E_{n_k}} f\left(\frac{dP_{n_k}}{dQ_{n_k}}\right) dQ_{n_k} = \int_{\mathbb{R}} F(r) dv_{n_k}(r) \to \int_{\mathbb{R}} F(r) dv(r) = \int_0^2 f(g(r))(2-r) dv(r)$$

as $k \to \infty$. Another application of the definition of weak convergence tells us that

$$\int_{E_{n_k}} f\left(\frac{dP_{n_k}}{dQ_{n_k}}\right) dQ_{n_k} = \int_{\mathbb{R}} f(x) d\lambda_{n_k} \to \int_{\mathbb{R}} f d\lambda$$

as $k \to \infty$, so that

$$\int_{\mathbb{R}} f d\lambda = \int_0^2 f(g(r))(2-r) dv(r).$$

Define $\{f_m\}_{m \in N_+}$ as

$$f_m(x) = \max(0, 1 - m|x|)$$

for any $x \in \mathbb{R}$ and $m \in N_+$. Then, $\{f_m\}_{m \in N_+}$ is a sequence of continuous and bounded functions bounded above by 1 that decreases to $I_{\{0\}}$. For any $m \in N_+$,

$$\int_{\mathbb{R}} f_m d\lambda = \int_0^2 f_m(g(r))(2-r) dv(r),$$

and by the DCT, taking $m \to \infty$ yields

$$\lambda(\{0\}) = \int_0^2 I_{\{0\}}(g(r))(2-r) dv(r) = \int_{[0,2)} I_{\{0\}}(g(r))(2-r) dv(r) = 2v(\{0\}),$$

since $g(r) = 0$ for $r \in [0,2)$ implies $r = 0$. Because we assumed that ii) holds, $\lambda(\{0\}) = 0$ and therefore $v(\{0\}) = 0$.

Choose any $f \in C_b(\mathbb{R}, \mathbb{R})$, and this time, define $\tilde{g} : [0,2] \to \mathbb{R}$ as

$$\tilde{g}(x) = \begin{cases} \frac{2-x}{x} & \text{if } x \in (0,2] \\ 0 & \text{otherwise} \end{cases}.$$

By lemma 9.1 again, we have, P_n-a.e.,

$$\frac{dQ_n}{dP_n} = \frac{dQ_n}{dR_n} / \frac{dP_n}{dR_n} = \frac{2 - \frac{dP_n}{dR_n}}{\frac{dP_n}{dR_n}},$$

so that

$$\frac{dQ_n}{dP_n}(x) = \tilde{g}\left(\frac{dP_n}{dR_n}(x)\right)$$

for P_n-a.e. $x \in E_n$. Therefore,

$$\begin{aligned}
\int_{E_n} f\left(\frac{dQ_n}{dP_n}\right) dP_n &= \int_{E_n} f\left(\tilde{g}\left(\frac{dP_n}{dR_n}\right)\right) dP_n \\
&= \int_{E_n} f\left(\tilde{g}\left(\frac{dP_n}{dR_n}\right)\right) \frac{dP_n}{dR_n} dR_n \\
&= \int_0^2 f(\tilde{g}(r)) r dv_n(r) = \int_{\mathbb{R}} \tilde{F}(r) dv_n(r)
\end{aligned}$$

for any $n \in N_+$, where $\tilde{F} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\tilde{F}(r) = \begin{cases} 0 & \text{if } r < 0 \\ f(\tilde{g}(r))r & \text{if } r \in [0,2] \\ 2f(0) & \text{if } r > 2 \end{cases}.$$

$\tilde{F}(\cdot)$ is a continuous and bounded function on $\mathbb{R}$, so that

$$\int_{E_{n_k}} f\left(\frac{dQ_{n_k}}{dP_{n_k}}\right) dP_{n_k} \to \int_{\mathbb{R}} \tilde{F}(r) dv(r) = \int_0^2 f(\tilde{g}(r)) r dv(r)$$

as $k \to \infty$ by the definition of weak convergence. Meanwhile,

$$\int_{E_n} f\left(\frac{dQ_n}{dP_n}\right) dP_n = \int_{\mathbb{R}} f(x) d\mu_n(x) \to \int_{\mathbb{R}} f(x) d\mu(x)$$

as $n \to \infty$ by our initial assumption, so

$$\int_{\mathbb{R}} f(x) d\mu(x) = \int_0^2 f(\tilde{g}(r)) r dv(r).$$

Letting $\{f_m\}_{m \in N_+}$ be a sequence of real valued functions defined as

$$f_m(x) = \begin{cases} -m & \text{if } x < -m \\ x & \text{if } x \in [-m, m] \\ m & \text{if } x > m \end{cases},$$

$\{f_m\}_{m \in N_+}$ is a sequence of bounded and continuous functions increasing to the identity

function on $\mathbb{R}$. Since $\{|f_m|\}_{m\in N_+}$ is a sequence of non-negative bounded and continuous functions increasing to the non-negative function $f:\mathbb{R}\to\mathbb{R}_+$ defined as $f(x)=|x|$, the MCT tells us that

$$\begin{aligned}\int_{\mathbb{R}}|x|d\mu(x) &= \lim_{m\to\infty}\int_{\mathbb{R}}|f_m|d\mu \\ &= \lim_{m\to\infty}\int_0^2|f_m(\tilde{g}(r))|rdv(r)=\int_0^2|\tilde{g}(r)|rdv(r) \\ &= \int_{(0,2]}(2-r)dv(r)=2\left(1-v(\{0\})\right)-1=1<+\infty,\end{aligned}$$

so that f is μ-integrable. It now follows from $|f_m|\leq|f|$ and the DCT that

$$\begin{aligned}\int_{\mathbb{R}}x\,d\mu(x) &= \lim_{m\to\infty}\int_{\mathbb{R}}f_m d\mu \\ &= \lim_{m\to\infty}\int_0^2 f_m(\tilde{g}(r))rdv(r)=\int_0^2\tilde{g}(r)rdv(r) \\ &= \int_{(0,2]}(2-r)dv(r)=2\left(1-v(\{0\})\right)-1=1.\end{aligned}$$

This completes the proof that iii) holds if ii) does.

iii) $\Rightarrow$ i)

It remains to show that iii) implies i). Suppose that iii) holds, and choose a sequence $\{A_n\}_{n\in N_+}$ of sets such that $A_n\in\mathcal{E}_n$ for each n and

$$\lim_{n\to\infty}P_n(A_n)=0.$$

Defining $B_n=A_n^c$ for each n, $P_n(B_n)\to 1$ as $n\to\infty$; defining $v_n=P_n\circ B_n^{-1}$ for each n, $\{v_n\}_{n\in N_+}$ is a sequence of probability measures on $(\{0,1\},2^{\{0,1\}})$ that is tight ($\{0,1\}$ is a compact set).

By lemma 9.3., we know that

$$\left\{\lambda_n=P_n\circ\left(\frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n\in N_+}$$

admits a subsequence that converges weakly to some probability measure λ on $\mathbb{R}$. For notational simplicity, let $\{\lambda_n\}_{n\in N_+}$ itself be this subsequence. Let $\{Z_n\}_{n\in N_+}$ be a sequence of random variables with distribution process $\{P_n\}_{n\in N_+}$, and define $\{X_n\}_{n\in N_+}$

as

$$X_n = \frac{dQ_n}{dP_n}(Z_n)$$

for any $n \in N_+$. Then, each X_n has distribution λ_n, and $X_n \xrightarrow{d} X$ for some real random variable X with distribution λ. Similarly, define

$$Y_n = I_{B_n}(Z_n)$$

for any $n \in N_+$. For any $\epsilon \in (0,1)$,

$$\mathbb{P}(|Y_n - 1| < \epsilon) = \mathbb{P}(I_{B_n}(Z_n) - 1) = P_n(B_n) - 1 \to 0$$

as $n \to \infty$, so that $Y_n \xrightarrow{p} 1$. By Slutsky's theorem, we can now see that

$$(X_n, Y_n) \xrightarrow{d} (X, 1).$$

Define the function $f : \mathbb{R}^2 \to \mathbb{R}$ as

$$f(x,y) = xy \cdot I_{[0,+\infty)^2}(x,y).$$

for any $(x,y) \in \mathbb{R}^2$. Then, f is a continuous non-negative function on $\mathbb{R}^2$, so that, by the necessity part of lemma 9.4,

$$\liminf_{n\to\infty} \mathbb{E}[f(X_n, Y_n)] \geq \mathbb{E}[f(X,1)].$$

For each $n \in N_+$, we can decompose

$$Q_n = Q_n^a + Q_n^\perp,$$

where $Q_n^a << P_n$ and $\frac{dQ_n}{dP_n} = \frac{\partial Q_n^a}{\partial P_n}$. Therefore,

$$\begin{aligned}
Q_n(B_n) =\geq Q_n^a(B_n) &= \int_{E_n} I_{B_n} \frac{dQ_n}{dP_n} dP_n \\
&= \int_\Omega X_n Y_n d\mathbb{P} && (P_n = \mathbb{P} \circ Z_n^{-1}) \\
&= \int_\Omega \left[X_n Y_n \cdot I_{[0,+\infty)^2}(X_n, Y_n) \right] d\mathbb{P} \\
& && (X_n, Y_n \text{ are non-negative valued}) \\
&= \mathbb{E}[f(X_n, Y_n)].
\end{aligned}$$

Meanwhile,

$$\mathbb{E}[f(X,1)] = \mathbb{E}[X] = \int_{\mathbb{R}} x \, d\lambda(x).$$

By iii), $\int_{\mathbb{R}} x \, d\lambda(x) = 1$, so that

$$\liminf_{n\to\infty} Q_n(B_n) \geq \liminf_{n\to\infty} \mathbb{E}[f(X_n, Y_n)] \geq \mathbb{E}[f(X,1)] = 1.$$

It follows that

$$\lim_{n\to\infty} Q_n(B_n) = 1,$$

or equivalently, $Q_n(A_n) \to 0$ as $n \to \infty$. We have shown that any subsequence of $\{Q_n(A_n)\}_{n\in N_+}$ has a further subsequence that converges to 0, so by the subsequence selection theorem, $\{Q_n(A_n)\}_{n\in N_+}$ itself converges to 0. This proves that $Q_n \lhd P_n$.

Q.E.D.

The following is one of the most popular applications of Le Cam's first lemma, which allows us to immediately establish continguity between sequences of measures whose likelihood ratio converges weakly to a lognormal distribution.

Corollary to Theorem 9.5 (Contiguity with Lognormal Limits)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces and $\{P_n\}_{n\in N_+}$, $\{Q_n\}_{n\in N_+}$ sequences such that P_n, Q_n are probability measures on $(E_n, \mathcal{E}_n)$ for each $n \in N_+$. Let v_{μ,σ^2} be the probability measure on $\mathbb{R}$ corresponding to a lognormal distribution with mean μ and variance $\sigma^2 > 0$. If

$$\left\{ Q_n \circ \left(\frac{dP_n}{dQ_n} \right)^{-1} \right\}_{n\in N_+}$$

converges weakly to v_{μ,σ^2}, then $Q_n \lhd P_n$. If in addition $\mu = -\frac{1}{2}\sigma^2$, then $P_n \lhd Q_n$ as well.

Proof) Since the support of the lognormal distribution is $(0, +\infty)$, it follows immediately that $v_{\mu,\sigma^2}((0,+\infty)) = 1$. Since every subsequence of $\left\{ Q_n \circ \left(\frac{dP_n}{dQ_n} \right)^{-1} \right\}_{n\in N_+}$ converges weakly to v_{μ,σ^2}, and $v_{\mu,\sigma^2}((0,+\infty)) = 1$, it follows from ii) in the theorem above that $Q_n \lhd P_n$. Suppose in addition that $\mu = -\frac{1}{2}\sigma^2$. Then, letting X be a real random variable with distribution v_{μ,σ^2}, and $Y \sim \mathcal{N}[\mu, \sigma^2]$,

$$\int_{\mathbb{R}} x \, d\mu(x) = \mathbb{E}[X] = \mathbb{E}[X] = \mathbb{E}[\exp(Y)]$$

$$= \exp\left(\mu + \frac{1}{2}\sigma^2\right) = 1$$

by the formula for the moment generating function of the normal distribution. It then follows from iii) of the theorem above that $P_n \lhd Q_n$.

Q.E.D.

9.1.3 Local Alternatives and Le Cam's Third Lemma

We now return to our original objective, that is, recovering the asymptotic distribution of some sequence of statistics $\{\mathcal{T}_n\}_{n\in N_+}$ under the sequence of probability measures $\{Q_n\}_{n\in N_+}$ when the asymptotic distribution of these statistics under the sequence $\{P_n\}_{n\in N_+}$ is known. It turns out that the exact same relationship holds under contiguity as under absolute continuity: that is, if we know the weak limit to which the probability measure

$$P_n \circ \left(\mathcal{T}_n,\ \frac{dQ_n}{dP_n}\right)^{-1}$$

converges, then we can recover the weak limit of

$$Q_n \circ \mathcal{T}_n^{-1}.$$

This result is called Le Cam's third lemma.

To give a heuristic motivation for the lemma, let (Θ, ϑ) be the parameter space and $\{\mu_n\}_{n\in N_+}$ a sequence of transition probability kernels from (Θ, ϑ) into $(E_n, \mathcal{E}_n)$; in light of the framework for asymptotic statistical problems, $\{\mu_n\}_{n\in N_+}$ represents a sequence of statistical models corresponding to each sample size. Suppose we are interested in testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$, and that we have obtained a sequence of real valued statistics $\{\mathcal{T}_n\}_{n\in N_+}$ designed to test H_0 against H_1. The distribution of $\mathcal{T}_n$ under the parameter θ is given as $\mu_n(\theta, \cdot) \circ \mathcal{T}_n^{-1}$.

Suppose that we know the limiting distribution of $\mathcal{T}_n$ under the null: that is, we have found some probability measure μ on $\mathbb{R}$ such that

$$\mu_n(\theta_0, \cdot) \circ \mathcal{T}_n^{-1} \to \mu$$

weakly. To evaluate the asymptotic performance of this test, we want to obtain its power under the local alternative $\{\theta_n\}_{n\in N_+}$, which is a sequence in $\Theta \setminus \{\theta_0\}$ converging to θ_0. In this case, what we want to find is the weak limit of the sequence of probability measures

$$\left\{\mu_n(\theta_n, \cdot) \circ \mathcal{T}_n^{-1}\right\}_{n\in N_+}.$$

Defining $P_n = \mu_n(\theta_0, \cdot)$ and $Q_n = \mu_n(\theta_n, \cdot)$, this is exactly the problem discussed in Le Cam's third lemma: given the knowledge of the asymptotic behavior of $\{P_n \circ \mathcal{T}_n^{-1}\}_n$, can we infer the asymptotic behavior of $\{Q_n \circ \mathcal{T}_n^{-1}\}$? The lemma says that we can, given that we also know the joint asymptotic behavior of $\mathcal{T}_n$ and the likelihood ratio $\frac{dP_n}{dQ_n}$ under P_n: that is, if we know the

weak limit of the sequence

$$\left\{\mu_n(\theta_0,\cdot)\circ\left(\mathcal{T}_n,\ \frac{d\mu_n(\theta_0,\cdot)}{d\mu_n(\theta_n,\cdot)}\right)^{-1}\right\}_{n\in N_+}.$$

Going forward, we use new notations that make this connection clear. Let $\{(E_n,\mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces, $\{G_n\}_{n\in N_+}$ a sequence of $\mathcal{E}_n$-measurable functions taking values in $\mathbb{R}^k$, and $\{P_n\}_{n\in N_+}$ a sequence of probability measures in $(E_n,\mathcal{E}_n)$. Then, for the result

$$P_n\circ G_n^{-1}\to\mu$$

weakly, where μ is some probability measure on $\mathbb{R}^k$, we write

$$G_n\overset{P_n}{\Rightarrow}G,$$

where G is some random variable taking values in $\mathbb{R}^k$ that has distribution μ. This makes clear that G_n converges in distribution to G in the sense that the distribution of G_n under P_n converges weakly to the distribution of G.

Theorem 9.6 (The General Version of Le Cam's Third Lemma)
Let $\{(E_n,\mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces and $\{P_n\}_{n\in N_+}$, $\{Q_n\}_{n\in N_+}$ sequences such that P_n,Q_n are probability measures on $(E_n,\mathcal{E}_n)$ for each $n\in N_+$. Let $\{\mathcal{T}_n\}_{n\in N_+}$ be a sequence of functions such that each $\mathcal{T}_n:E_n\to\mathbb{R}^k$ is measurable relative to $\mathcal{E}_n$ and $\mathcal{B}(\mathbb{R}^k)$.

Suppose that $Q_n\lhd P_n$ and that

$$\left(\mathcal{T}_n,\ \frac{dQ_n}{dP_n}\right)\overset{P_n}{\Rightarrow}(\mathcal{T}^P,V).$$

Then,

$$\mathcal{T}_n\overset{Q_n}{\Rightarrow}\mathcal{T}^Q,$$

and the function $\mu:\mathcal{B}(\mathbb{R}^k)\to[0,+\infty]$ defined as

$$\mu(A)=\mathbb{E}\left[I_A(\mathcal{T}^P)\cdot V\right]$$

for any $A\in\mathcal{B}(\mathbb{R}^k)$ is the distribution of L.

Proof) Define the sequence

$$\left\{v_n=P_n\circ\left(\mathcal{T}_n,\ \frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n\in N_+}$$

and let v be the joint distribution of $(\mathcal{T}^P,V)$. Then, we are assuming that $v_n\to v$

weakly. Similarly, define

$$\{\mu_n = Q_n \circ \mathcal{T}_n^{-1}\}_{n \in N_+};$$

our objective is to show that $\mu_n \to \mu$ weakly as $n \to \infty$.

μ is a probability measure

We first show that μ is a well-defined probability measure on $\mathbb{R}^k$. V is a non-negative valued random variable because each $\frac{dQ_n}{dP_n}(X_n^P)$ is non-negative valued, and from Le Cam's first lemma $\mathbb{E}[V] = 1$, since

$$\left\{P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1}\right\}_{n \in N_+}$$

converges weakly to the marginal distribution of V, and $Q_n \lhd P_n$. It follows that the mapping $\mu : \mathcal{B}(\mathbb{R}^k) \to [0, +\infty]$ is well-defined as the expectation of non-negative random variables, and that $\mu(\mathbb{R}^k) = \mathbb{E}[V] = 1$. It is clear that $\mu(\emptyset) = 0$, and for any disjoint sequence $\{A_n\}_{n \in N_+}$ of Borel measurable subsets of $\mathbb{R}^k$,

$$\mu\left(\bigcup_n A_n\right) = \mathbb{E}\left[\sum_{n=1}^{\infty} I_{A_n}(\mathcal{T}^P) \cdot V\right] = \sum_{n=1}^{\infty} \mathbb{E}\left[I_{A_n}(\mathcal{T}^P) \cdot V\right] = \sum_{n=1}^{\infty} \mu(A_n)$$

by the MCT for series. Therefore, $\mu(\cdot)$ is a well-defined probability measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$.

Integration with respect to μ

Let $f : \mathbb{R}^k \to \mathbb{R}$ be a non-negative Borel-measurable simple function with canonical form

$$f = \sum_{i=1}^{n} a_i \cdot I_{A_i}.$$

Then,

$$\int_{\mathbb{R}^k} f d\mu = \sum_{i=1}^{n} a_i \cdot \mu(A_i) = \mathbb{E}\left[\left[\left(\sum_{i=1}^{n} a_i \cdot I_{A_i}\right) \circ \mathcal{T}^P\right] \cdot V\right] = \mathbb{E}\left[f(\mathcal{T}^P)V\right].$$

Now let f be an arbitrary non-negative Borel-measurable function. Letting $\{f_n\}_{n \in N_+}$ be a sequence of non-negative measurable simple functions that increases to f, the MCT tells us that

$$\int_{\mathbb{R}^k} f d\mu = \lim_{n \to \infty} \int_{\mathbb{R}^k} f_n d\mu$$

$$= \lim_{n\to\infty} \mathbb{E}\left[f_n(\mathcal{T}^P)V\right] = \mathbb{E}\left[f(\mathcal{T}^P)V\right],$$

since $\{(f_n \circ \mathcal{T}^P)V\}_{n\in N_+}$ is a sequence of non-negative random variables that increases to $(f \circ \mathcal{T}^P)V$.

Proving Weak Convergence

For any $n \in N_+$ and non-negative continuous function f on $\mathbb{R}^k$, letting $Q_n = Q_n^a + Q_n^\perp$ be the Lebesgue decomposition of Q_n with respect to P_n, we have

$$\begin{aligned}\int_{\mathbb{R}^k} f d\mu_n = \mathbb{E}\left[f(\mathcal{T}_n(X_n^Q))\right] &= \int_{E_n} (f \circ \mathcal{T}_n) dQ_n \\ &\geq \int_{E_n} (f \circ \mathcal{T}_n) dQ_n^a = \int_{E_n} (f \circ \mathcal{T}_n) \frac{dQ_n}{dP_n} dP_n \\ &= \mathbb{E}\left[f(\mathcal{T}_n(X_n^P)) \cdot \frac{dQ_n}{dP_n}(X_n^P)\right].\end{aligned}$$

Defining the function $g : \mathbb{R}^{k+1} \to \mathbb{R}$ as

$$g(x,y) = f(x)y \cdot I_{[0,+\infty)}(y) \quad \text{for any } (x,y) \in \mathbb{R}^k \times \mathbb{R},$$

it follows that $g(\cdot)$ is a non-negative continuous function on $\mathbb{R}^{k+1}$ such that

$$\begin{aligned}\mathbb{E}\left[f(\mathcal{T}_n(X_n^P)) \cdot \frac{dQ_n}{dP_n}(X_n^P)\right] &= \int_{\mathbb{R}^{k+1}} f(x)y \; dv_n(x,y) \\ &= \int_{\mathbb{R}^k \times [0,+\infty)} f(x)y \; dv_n(x,y) \\ &\qquad (v_n \text{ is concentrated on } \mathbb{R}^k \times [0,+\infty)) \\ &= \int_{\mathbb{R}^{k+1}} g \; dv_n\end{aligned}$$

for any $n \in N_+$, and likewise,

$$\int_{\mathbb{R}^k} f d\mu = \mathbb{E}\left[f(\mathcal{T}^P)V\right] = \int_{\mathbb{R}^{k+1}} g \; dv.$$

By the necessity part of lemma 9.4, $v_n \to v$ weakly implies that

$$\liminf_{n\to\infty} \int_{\mathbb{R}^{k+1}} g \; dv_n \geq \int_{\mathbb{R}^{k+1}} g dv = \int_{\mathbb{R}^k} f d\mu.$$

Together with the fact that $\int_{\mathbb{R}^k} f d\mu_n \geq \int_{\mathbb{R}^{k+1}} g \; dv_n$ for any $n \in N_+$, we can now see that

$$\liminf_{n\to\infty} \int_{\mathbb{R}^k} f d\mu_n \geq \int_{\mathbb{R}^k} f d\mu.$$

This holds for any non-negative continuous function f on $\mathbb{R}^k$, and because any open subset of euclidean space is σ-compact, the sufficiency part of lemma 9.4 implies that $\mu_n \to \mu$ weakly.

Q.E.D.

Note that Le Cam's first lemma, and by extension the contiguity relationship $Q_n \lhd P_n$, is only used to ensure that the limiting measure μ has mass 1, or in other words to ensure that it is a probability measure. This is similar to Prohorov's theorem, in which we can prove that any collection of probability measures on a metric space admits a weakly convergent subsequence, but this weak limit is a probability measure only when the collection in question is tight.

Le Cam's third lemma can be used to obtain asymptotic results in the lognormal case discussed in the corollary to lemma 9.5. This result is what is usually referred to as the third lemma, and we state it separately in light of its incredible significance in the statistical literature.

First, we given meaning to the convergence in distribution of the logs of non-negative random variables. Given any sequence of non-negative valued real random variables $\{X_n\}_{n \in N_+}$ and some real-valued random variable X, we write

$$\log(X_n) \xrightarrow{d} X$$

for

$$X_n \xrightarrow{d} e^X.$$

If $X_n > 0$ with probability 1 for any $n \in N_+$, then the two are equivalent in light of the continuous mapping theorem. In the more general case, because $e^X > 0$ everywhere on the sample space, letting μ be the distribution of e^X, we have

$$\mu(\partial(0, +\infty)) = \mu(\{0\}) = \mathbb{P}\left(e^X = 0\right) = 0.$$

Therefore, by the Portmanteau theorem,

$$\lim_{n \to \infty} \mathbb{P}\left(X_n > 0\right) = \mu((0, +\infty)) = \mathbb{P}\left(e^X > 0\right) = 1,$$

which shows us that the random variables X_n are positive-valued with probability approaching 1. In other words, $\log(X_n)$ is well-defined (or real-valued if we let $\log(0) = -\infty$) with probability approaching 1 as $n \to \infty$.

Theorem 9.7 (Le Cam's Third Lemma for the Log-Normal Case)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces and $\{P_n\}_{n\in N_+}$, $\{Q_n\}_{n\in N_+}$ sequences such that P_n, Q_n are probability measures on $(E_n, \mathcal{E}_n)$ for each $n \in N_+$. Let $\{\mathcal{T}_n\}_{n\in N_+}$ be a sequence of functions such that each $\mathcal{T}_n : E_n \to \mathbb{R}^k$ is measurable relative to $\mathcal{E}_n$ and $\mathcal{B}(\mathbb{R}^k)$.

Suppose that $Q_n \lhd P_n$ and that

$$\left(\mathcal{T}_n,\, \log\frac{dQ_n}{dP_n}\right) \overset{P_n}{\Rightarrow} (\mathcal{T}^P, V) \sim \mathcal{N}\left[\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix},\, \begin{pmatrix} \Sigma & \tau' \\ \tau & \sigma^2 \end{pmatrix}\right]$$

for some $\mu \in \mathbb{R}^k$, $\sigma^2 > 0$, positive definite $\Sigma \in \mathbb{R}^{k\times k}$, and $\tau \in \mathbb{R}^k$. Then,

$$\mathcal{T}_n \overset{Q_n}{\Rightarrow} \mathcal{T}^Q \sim \mathcal{N}\left[\mu + \tau,\, \Sigma\right].$$

Proof) Focusing only on the marginal distribution of the likelihood ratio under P_n, we can see that

$$\log\frac{dQ_n}{dP_n} \overset{P_n}{\Rightarrow} V \sim \mathcal{N}\left[-\frac{1}{2}\sigma^2,\, \sigma^2\right],$$

which is equivalent to

$$\frac{dQ_n}{dP_n} \overset{P_n}{\Rightarrow} e^V,$$

which in turn means that

$$P_n \circ \left(\frac{dQ_n}{dP_n}\right)^{-1} \to v$$

weakly, where v is the distribution of e^V. By assumption, v is the probability measure on $\mathbb{R}$ corresponding to a lognormal distribution with mean $-\frac{1}{2}\sigma^2$ and variance σ^2, so by the corollary to theorem 9.5, $Q_n \lhd\rhd P_n$.

Now, the general version of Le Cam's third lemma tells us that, because

$$\left(\mathcal{T}_n,\, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\Rightarrow} (\mathcal{T}^P, e^V),$$

we have

$$\mathcal{T}_n \overset{Q_n}{\Rightarrow} \mathcal{T}^Q,$$

where the distribution μ of $\mathcal{T}^Q$ satisfies

$$\mu(A) = \mathbb{E}\left[I_A(\mathcal{T}^P)e^V\right]$$

for any $A \in \mathcal{B}(\mathbb{R}^k)$. We saw in the proof of the general version of the third lemma that

$$\int_{\mathbb{R}^k} f d\mu = \mathbb{E}\left[f(\mathcal{T}^P)e^V\right]$$

for any non-negative Borel measurable function f. The moment generating function of $\mathcal{T}^Q$ is therefore given as

$$\begin{aligned} m_Q(r) = \mathbb{E}\left[e^{r'\mathcal{T}_Q}\right] &= \int_{\mathbb{R}^k} e^{r't} d\mu(t) \\ &= \mathbb{E}\left[e^{r'\mathcal{T}^P + V}\right] = \exp\left(\begin{pmatrix} r' & 1 \end{pmatrix}\begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} r' & 1 \end{pmatrix}\begin{pmatrix} \Sigma & \tau' \\ \tau & \sigma^2 \end{pmatrix}\begin{pmatrix} r \\ 1 \end{pmatrix}\right) \\ &= \exp\left(r'(\mu+\tau) + \frac{1}{2}r'\Sigma r\right) \end{aligned}$$

for any $r \in \mathbb{R}^k$, where we used the formula for the moment generating function of a Gaussian random vector to justify the third equality. This is precisely the mgf of a normal random vector with mean $\mu + \tau$ and variance Σ, so

$$\mathcal{T}^Q \sim \mathcal{N}\left[\mu + \tau,\ \Sigma\right],$$

as we desired.

Q.E.D.

9.2 Local Asymptotic Normality

This section concerns an important generalization of Le Cam's third lemma. First, we introduce the statistical framework and approximation by normal limiting experiments; then, we introduce local asymptotic normality as a sufficient condition to ensure that such an approximation holds. Finally, we furnish some primitive conditions under which local asymptotic normality holds, with a focus on the differentiability in quadratic mean condition.

9.2.1 Motivation and the Gaussian Location Problem

We work with the following framework. Let $\{(E_n, \mathcal{E}_n)\}_{n \in N_+}$ be a sequence of measurable spaces and (Θ, ϑ) the k-dimensional parameter space. Let $\{\mu_n\}_{n \in N_+}$ denote a sequence of transition probability kernels from (Θ, ϑ) into $(E_n, \mathcal{E}_n)$, so that each μ_n represents a statistical model given a sample of size n. Suppose we are interested in testing the null $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$ using a sequence of k-dimensional estimators $\{\mathcal{T}_n\}_{n \in N_+}$ of the parameter, where each $\mathcal{T}_n$ is a $\mathcal{E}_n$-measurable function.

Suppose we know that $\mathcal{T}_n$ is asymptotically unbiased and normal under the null; that is, we know that

$$\mathcal{T}_n \overset{\mu_n(\theta_0, \cdot)}{\Rightarrow} \mathcal{T}^P \sim \mathcal{N}\left[O_{k \times 1},\, \Sigma^{-1}\right].$$

for $\bar{T} \in \mathbb{R}^k$ and positive definite $\Sigma \in \mathbb{R}^{k \times k}$. Consider a local alternative given as $\{\theta_0 + h/\sqrt{n}\}_{n \in N_+}$, where $h \in \mathbb{R}^k$. If we know that $\mathcal{T}_n$ and the log likelihood ratio $\log \frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}$ converge jointly to a normal random variable under the null:

$$\left(\mathcal{T}_n,\, \log \frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}\right) \overset{\mu_n(\theta_0, \cdot)}{\Rightarrow} \mathcal{N}\left[\begin{pmatrix} O_{k \times 1} \\ -\frac{1}{2} h' \Sigma h \end{pmatrix}, \begin{pmatrix} \Sigma^{-1} & h' \\ h & h' \Sigma h \end{pmatrix}\right],$$

then by Le Cam's third lemma, we know that

$$\mathcal{T}_n \overset{\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{\Rightarrow} \mathcal{N}\left[h,\, \Sigma^{-1}\right].$$

In other words, $\mathcal{T}_n$ can be approximated in the limit for any $h \in \mathbb{R}^k$ by a random vector that follows the normal distribution $\mathcal{N}\left[h,\, \Sigma^{-1}\right]$.

The Gaussian Location Problem

Another way to put this is as follows. Let μ be a Markov kernel on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ such that

$$\mu(h, \cdot) \sim \mathcal{N}\left[h,\, \Sigma^{-1}\right]$$

for any $h \in \mathbb{R}^k$. Then,

$$\mu_n(\theta_0 + h/\sqrt{n}, \cdot) \circ \mathcal{T}_n^{-1} \to \mu(h, \cdot)$$

weakly as $n \to \infty$ for any $h \in \mathbb{R}^k$. In other words, the limiting distribution of the statistic $\mathcal{T}_n$ under the local alternative $\{\theta_0 + h/\sqrt{n}\}_{n \in N_+}$ can be approximated by the distribution $\mu(h, \cdot)$ for any choice of $h \in \mathbb{R}^k$.

Focusing on this kernel μ, we can see that this is the model corresponding to a statistical decision problem in which we want to make inferences about the mean of a normal distribution using a single observation drawn from the true normal distribution. Problems involving inferences about the mean of a normal distribution are called Gaussian location problems; as such, μ is the model corresponding to a Gaussian location problem utilizing a single observation. The asymptotic approximation above thus shows us that the statistical problem of interest that uses the statistic $\mathcal{T}_n$ to make inferrences about the parameter θ is equivalent, in the limit, to a Gaussian location problem making use of a single observation. In other words, the asymptotic properties of the original statistical problem can be obtained by simply studying the simpler Gaussian location problem. We will repeatedly take advantage of this simplification below when formalizing the concepts of asymptotic parametric and semi-parametric efficiency.

The asymptotic equivalence between the original statistical problem and the Gaussian location problem in the limit can be phrased in another way. Let X be a random variable be drawn from a normal distribution with variance Σ^{-1} and unknown mean. Under the local alternative $\theta_0 + h/\sqrt{n}$, the statistic $\mathcal{T}_n$ behaves, asymptotically, in the same manner as the random variable X if h were its true mean. To make this interpretation clear, we write

$$\mathcal{T}_n \overset{h}{\Rightarrow} X.$$

Here, the superscript h serves two roles; it represents how we are studying the asymptotic behavior of $\mathcal{T}_n$ under the measure $\mu_n(\theta_0 + h/\sqrt{n}, \cdot)$; and it also represents that the true mean of X is h, that is, that we are treating X as a draw from the distribution $\mathcal{N}\left[h, \Sigma^{-1}\right]$.

Sufficient Conditions for Approximation by a Gaussian Location Problem

The natural next step is to generalize this result. Specifically, we are interested in furnishing sufficient conditions under which a sequence of d-dimensional statistics $\{\mathcal{T}_n\}_{n \in N_+}$ can be approximated in the limit by a statistic in a normal limiting experiment, that is, if there exists some Gaussian location problem utilizing a single observation X and a statistic $G : \mathbb{R}^k \to \mathbb{R}^d$ from this problem such that

$$\mathcal{T}_n \overset{h}{\Rightarrow} G(X)$$

for any $h \in \mathbb{R}^k$.

It turns out that there are two main sufficient conditions that ensure a statistic can be

approximated in the limit by a statistic in a Gaussian location problem. The first is for $\mathcal{T}_n$ to converge in distribution under the null; this is a standard result that is established in the process of constructing a test statistic. The second sufficient condition is weak convergence of the log likelihood ratio under the null:

$$\log \frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} \overset{\mu_n(\theta_0,\cdot)}{\Rightarrow} \mathcal{N}\left[-\frac{1}{2}h'\Sigma h,\ h'\Sigma h\right],$$

which can equivalently be written as

$$\frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} \overset{\mu_n(\theta_0,\cdot)}{\Rightarrow} \exp\left(-\frac{1}{2}h'\Sigma h + h'\Delta\right)$$

for some $\Delta \sim \mathcal{N}[O_{k\times 1}, \Sigma]$. This condition is called local asymptotic normality, since it requires the log likelihood ratio of local alternatives to converge to a normal distribution in the limit. Below we show that these two conditions are sufficient for the distribution of the statistics of interest to be approximated by a Gaussian location problem, and furnish sufficient conditions under which local asymptotic normality holds.

9.2.2 The Main Approximation Result

We continue to work with the framework introduced above: $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ is a sequence of measurable spaces, (Θ, ϑ) the k-dimensional parameter space, and $\{\mu_n\}_{n\in N_+}$ a sequence of transition probability kernels from (Θ, ϑ) into $(E_n, \mathcal{E}_n)$. The sequence $\{\mu_n\}_{n\in N_+}$ is said to satisfy the local asymptotic normality condition at $\theta_0 \in \Theta^o$ if

> There exists a sequence $\{r_n\}_{n\in N_+}$ of nonsingular $k \times k$ matrices, a sequence $\{\Delta_n\}_{n\in N_+}$ such that each $\Delta_n : E_n \to \mathbb{R}^k$ is a measurable function, a positive definite matrix $I_0 \in \mathbb{R}^{k\times k}$ and a random vector $\Delta_0 \sim \mathcal{N}[O_{k\times 1},\ I_0]$ such that $\Delta_n \overset{\mu_n(\theta_0,\cdot)}{\Rightarrow} \Delta_0$ and
>
> $$\frac{d\mu_n(\theta_0 + r_n^{-1}h_n, \cdot)}{d\mu_n(\theta_0, \cdot)} - \exp\left(h'\Delta_n - \frac{1}{2}h'I_0h\right) \overset{\mu_n(\theta_0,\cdot)}{\Rightarrow} 0$$
>
> for any sequence $\{h_n\}_{n\in N_+}$ in $\mathbb{R}^k$ converging to some $h \in \mathbb{R}^k$, where $\theta_0 + r_n^{-1}h_n \in \Theta$ for sufficiently large n.

Generalizing the notation introduced above, we denote weak convergence under the distribution $\mu_n(\theta_0 + r_n^{-1}h_n, \cdot)$, where $h_n \to h$, by $\overset{h}{\Rightarrow}$[2]; local asymptotic normality then implies that

$$\log \frac{d\mu_n(\theta_0 + r_n^{-1}h_n, \cdot)}{d\mu_n(\theta_0, \cdot)} \overset{0}{\Rightarrow} h'\Delta_0 - \frac{1}{2}h'I_0h.$$

Before proving the result of interest, we introduce an additional concept. Consider a Gaussian location problem with variance I_0^{-1} and a single observation X. We call any Borel measurable

[2]This may be misleading because the convergence depends not only on the limit h but also on the choice of the sequence h_n; but in most cases we make explicit which sequence h_n we are considering before utilizing this notation, so that there is no confusion.

function $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ a randomized statistic from the location problem, since in this case the function $G(\cdot)$ depends not only on the observation X, but also on an independent uniformly distributed random variable U, which adds an additional element of randomness to the value of the statistic.

What we will show below is not that the statistics of interest can be approximated in the limit by a statistic in a Gaussian location problem, but rather by a randomized statistic in this problem; specifically, that

$$\mathcal{T}_n \overset{h}{\Rightarrow} G(X, U)$$

for any $h \in \mathbb{R}^k$, where $U \sim \text{Uniform}[0,1]$ and $U \perp X$. Formally, the above expression is shorthand for the result

$$\mu_n(\theta_0 + r_n^{-1} h_n, \cdot) \circ \mathcal{T}_n^{-1} \to (\mu(h, \cdot) \times v) \circ G^{-1}$$

weakly, where $\mu(h, \cdot) \sim \mathcal{N}\left[h, I_0^{-1}\right]$ and v is the uniform distribution on $[0,1]$.

First, we present a lemma of interest.

Lemma 9.8 Let $\mathcal{T}$ be a real valued random variable and Δ a real k-dimensional random vector. Then, there exists a measurable function $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}$ such that, for any random variable U that is independent of Δ and uniformly distributed on $[0,1]$, we have

$$(G(\Delta, U), \Delta) \sim (\mathcal{T}, \Delta).$$

Proof) Because $\mathcal{T}$ is a random variable taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, there exists a transition probability kernel μ from $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\mu(\cdot, A) = \mathbb{P}(\mathcal{T} \in A \mid \Delta)$$

for any $A \in \mathcal{B}(\mathbb{R})$. $\mu(\delta, \cdot)$ now represents the conditional distribution of $\mathcal{T}$ given $\Delta = \delta$ for any $\delta \in \mathbb{R}^k$.

Let the conditional distribution function $F : \mathbb{R} \times \mathbb{R}^k \to [0,1]$ be defined as

$$F(t \mid \delta) = \mu(\delta, \ (-\infty, t])$$

for any $t \in \mathbb{R}$; $F(\cdot \mid \delta)$ is the distribution function of $\mathcal{T}$ given $\Delta = \delta$. The Borel measurability of the section $F(t \mid \cdot)$ and the right continuity of the section $F(\cdot \mid \delta)$ implies that $F(\cdot)$ is jointly Borel measurable on $\mathbb{R}^{k+1}$.

We can now define the conditional quantile function $Q : (0,1) \times \mathbb{R}^k \to [-\infty, +\infty]$ as

$$Q(u \mid \delta) = \inf\{t \in \mathbb{R} \mid F(t \mid \delta) \geq u\}$$

for any $(u,\delta) \in [0,1] \times \mathbb{R}^k$, where we define $Q(0 \mid \delta) = -\infty$ and $Q(1 \mid \delta) = +\infty$. We know that

$$Q(u \mid \delta) \leq t \quad \text{if and only if} \quad u \leq F(t \mid \delta)$$

for any $u \in (0,1)$ and $t \in \mathbb{R}$.

Note that $Q(\cdot)$ is jointly Borel measurable on $(0,1) \times \mathbb{R}^k$: for any $\alpha \in \mathbb{R}$, $\inf\{t \in \mathbb{R} \mid F(t \mid \delta) \geq u\} \geq \alpha$ implies that $t \geq \alpha$ for any $(t,\delta) \in \mathbb{R}^{k+1}$ such that $F(t \mid \delta) \geq u$, so that

$$\begin{aligned}\{(u,\delta) \in [0,1] \times \mathbb{R}^k \mid Q(u \mid \delta) \geq \alpha\} &= \{(u,\delta) \in [0,1] \times \mathbb{R}^k \mid \inf\{t \in \mathbb{R} \mid F(t \mid \delta) \geq u\} \geq \alpha\} \\ &= \left(\bigcap_{q<\alpha,\ q \in \mathbb{Q}} \{(u,\delta) \in [0,1] \times \mathbb{R}^k \mid F(q \mid \delta) < u\} \right) \cup \Big(\{1\} \times \mathbb{R}^k \Big),\end{aligned}$$

where the rightmost term is a measurable rectangle on $[0,1] \times \mathcal{B}(\mathbb{R}^k)$.

Define $G : [0,1] \times \mathbb{R}^k \to \mathbb{R}$ as

$$G(u,\delta) = Q(u \mid \delta) \cdot I_{(0,1)}(u)$$

for any $(u,\delta) \in [0,1] \times \mathbb{R}^k$. Then, $G(\cdot)$ is clearly jointly Borel measurable. Let U be a uniformly distributed random variable that is independent of Δ. Then,

$$\mathbb{P}\left(Q(U \mid \Delta) \neq G(U,\Delta) \mid \Delta\right) = \mathbb{P}\left(U \in \{0,1\} \mid \Delta\right) = \mathbb{P}\left(U \in \{0,1\}\right) = 0,$$

so that $Q(U \mid \Delta) = G(U,\Delta)$ almost surely by the law of iterated expectations. Furthermore, for any $\delta \in \mathbb{R}^k$,

$$\mathbb{P}\left(G(U,\Delta) \leq t \mid \Delta\right) = \mathbb{P}\left(Q(U \mid \Delta) \leq t \mid \Delta\right) = \mathbb{P}\left(U \leq F(t \mid \Delta) \mid \Delta\right) = F(t \mid \Delta).$$

As such, for any $t \in \mathbb{R}$ and $B \in \mathcal{B}(\mathbb{R}^k)$, we have

$$\begin{aligned}\mathbb{P}\left(G(U,\Delta) \leq t,\ \Delta \in B\right) &= \mathbb{P}\left(\mathbb{P}\left(G(U,\Delta) \leq t \mid \Delta\right)\left(I_B \circ \Delta\right)\right) \\ &= \mathbb{P}\left(F(t \mid \Delta) \cdot \left(I_B \circ \Delta\right)\right) = \mathbb{P}\left(\mu(\Delta, (-\infty, t])(I_B \circ \Delta)\right) \\ &= \mathbb{P}\left(\mathcal{T} \leq t,\ \Delta \in B\right).\end{aligned}$$

Since $\{(-\infty,t] \times B \mid t \in \mathbb{R}, B \in \mathcal{B}(\mathbb{R}^k)\}$ is a π-system that generates $\mathcal{B}(\mathbb{R}^{k+1})$, it follows that

$$(G(U,\Delta),\Delta) \sim (\mathcal{T},\Delta).$$

Q.E.D.

Theorem 9.9 (Approximation via a Gaussian Location Problem)
Let $\{(E_n, \mathcal{E}_n)\}_{n \in N_+}$ be a sequence of measurable spaces, (Θ, ϑ) the k-dimensional parameter space and $\{\mu_n\}_{n \in N_+}$ the corresponding statistical models. Let $\{\mathcal{T}_n\}_{n \in N_+}$ be a sequence such that each $\mathcal{T}_n : E_n \to \mathbb{R}$ is $\mathcal{E}_n$-measurable. Suppose that the following hold true:

i) **(Local Asymptotic Normality)**
$\{\mu_n\}_{n \in N_+}$ is locally asymptotically normal at $\theta_0 \in \Theta$, so that there exists a sequence $\{r_n\}_{n \in N_+}$ of non-singular $k \times k$ matrices, a sequence $\{\Delta_n\}_{n \in N_+}$ such that each $\Delta_n : E_n \to \mathbb{R}^k$ is a measurable function, a positive definite $k \times k$ matrix I_0 and a k-dimensional random vector $\Delta_0 \sim \mathcal{N}[O_{k \times 1},\ I_0]$ such that $\Delta_n \overset{0}{\Rightarrow} \Delta_0$ and

$$\frac{d\mu_n(\theta_0 + r_n^{-1}h, \cdot)}{d\mu_n(\theta_0, \cdot)} - \exp\left(h'\Delta_n - \frac{1}{2}h'I_0h\right) \overset{0}{\Rightarrow} 0$$

for any $h \in \mathbb{R}^k$ such that $\theta_0 + r_n^{-1}h_n \in \Theta$ for sufficiently large n.

ii) **(Convergence in Distribution under Local Alternatives)**
For each $h \in \mathbb{R}^k$ such that $\theta_0 + r_n^{-1}h \in \Theta$ for sufficiently large n, there exists some real random variable $\mathcal{T}^h$ such that

$$\mathcal{T}_n \overset{h}{\Rightarrow} \mathcal{T}^h.$$

Then, there exists a measurable function $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}$ such that

$$\mathcal{T}^h \overset{h}{\sim} G(X, U),$$

where X is an observation from the Gaussian location problem with variance I_0^{-1} and U is a uniformly distributed random variable independent of X, for any $h \in \mathbb{R}^k$ such that $\theta_0 + r_n^{-1}h \in \Theta$ for sufficiently large n.

Proof) Denote

$$P_n = \mu_n(\theta_0, \cdot)$$

for any $n \in N_+$. By assumption,

$$\mathcal{T}_n \overset{P_n}{\Rightarrow} \mathcal{T}^0, \quad \Delta_n \overset{P_n}{\Rightarrow} \Delta_0.$$

It follows from the necessity part of Prohorov's theorem that the sequences $\{P_n \circ \mathcal{T}_n^{-1}\}_{n \in N_+}$ and $\{P_n \circ \Delta_n^{-1}\}_{n \in N_+}$ are tight. By definition, for any $\epsilon > 0$, there exist

compact subsets K_1 and K_2 of $\mathbb{R}$ and $\mathbb{R}^k$ such that, for any $\epsilon > 0$,

$$(P_n \circ \mathcal{T}_n^{-1})(K_1) > 1 - \frac{\epsilon}{2}, \quad (P_n \circ \Delta_n^{-1})(K_2) > 1 - \frac{\epsilon}{2}$$

for any n. Letting $K = K_1 \times K_2$, K is a compact subset of $\mathbb{R}^{k+1}$ by Tychonoff's theorem that satisfies

$$\begin{aligned}(P_n \circ (\mathcal{T}_n, \Delta_n)^{-1})(K) &= P_n\left[\{\mathcal{T}_n \in K_1\} \cap \{\Delta_n \in K_2\}\right] \\ &\geq P_n\left[\{\mathcal{T}_n \in K_1\}\right] + P_n\left[\{\Delta_n \in K_2\}\right] - 1 > 1 - \epsilon\end{aligned}$$

for any n. We can find such a compact set for each $\epsilon > 0$, so by definition $\{P_n \circ (\mathcal{T}_n, \Delta_n)^{-1}\}_{n \in N_+}$ is a tight sequence of probability measures on $\mathbb{R}^{k+1}$. By the sufficiency part of Prohorov's theorem, there exists a weakly convergent subsequence $\{P_{n_m} \circ (\mathcal{T}_{n_m}, \Delta_{n_m})^{-1}\}_{m \in N_+}$ of $\{P_n \circ (\mathcal{T}_n, \Delta_n)^{-1}\}_{n \in N_+}$.

Denote the weak limit of $\{P_{n_m} \circ (\mathcal{T}_{n_m}, \Delta_{n_m})^{-1}\}_{m \in N_+}$ be the distribution of the $k+1$-dimensional random vector $(\mathcal{T}', \Delta')$. Let $\pi_{-1} : \mathbb{R}^{k+1} \to \mathbb{R}^k$ be the projection function defined as $\pi_{-1}(x_0, \cdots, x_k) = (x_1, \cdots, x_k)$ for any $(x_0, \cdots, x_k) \in \mathbb{R}^{k+1}$. For any bounded and continuous function $f : \mathbb{R}^k \to \mathbb{R}$, the function $f \circ \pi_{-1}$ is a bounded and continuous function on $\mathbb{R}^{k+1}$, so that

$$\begin{aligned}\mathbb{E}\left[f(\Delta_0)\right] &= \lim_{m \to \infty} \int_{E_{n_m}} f(\Delta_{n_m}) dP_{n_m} \\ &= \lim_{m \to \infty} \int_{E_{n_m}} (f \circ \pi_{-1})(\mathcal{T}_{n_m}, \Delta_{n_m}) dP_{n_m} \\ &= \mathbb{E}\left[(f \circ \pi_{-1})(\mathcal{T}', \Delta')\right] = \mathbb{E}\left[f(\Delta')\right]\end{aligned}$$

by the definition of weak convergence. As such, $\Delta' \sim \Delta_0 \sim \mathcal{N}\left[O_{k \times 1},\ I_0\right]$.

Now choose any $h \in \mathbb{R}^k$ such that $\theta_0 + r_n^{-1} h \in \Theta$ for large enough n, and denote

$$Q_n = \mu_n(\theta_0 + r_n^{-1} h, \cdot)$$

for any $n \in N_+$ such that Q_n is well-defined; we assume without loss of generality that Q_n is well-defined for any $n \in N_+$. Let $\{X_n\}_{n \in N_+}$ be a sequence of random variables such that each X_n has distribution P_n. Then, by assumption,

$$\frac{dQ_n}{dP_n}(X_n) - \exp\left(h' \Delta_n(X_n) - \frac{1}{2} h' I_0 h\right) \xrightarrow{d} 0,$$

which implies that

$$\frac{dQ_n}{dP_n}(X_n) - \exp\left(h'\Delta_n(X_n) - \frac{1}{2}h'I_0h\right) = o_p(1).$$

Likewise,

$$\left(\mathcal{T}_{n_m}(X_{n_m}), \quad \exp\left(h'\Delta_{n_m}(X_{n_m}) - \frac{1}{2}h'I_0h\right)\right) \xrightarrow{d} \left(\mathcal{T}', \quad \exp\left(h'\Delta' - \frac{1}{2}h'I_0h\right)\right).$$

It follows from Slutsky's theorem that

$$\left(\mathcal{T}_{n_m}(X_{n_m}), \quad \frac{dQ_{n_m}}{dP_{n_m}}(X_{n_m})\right) \xrightarrow{d} \left(\mathcal{T}', \quad \exp\left(h'\Delta' - \frac{1}{2}h'I_0h\right)\right),$$

or equivalently,

$$\left(\mathcal{T}_{n_m}, \quad \frac{dQ_{n_m}}{dP_{n_m}}\right) \overset{P_{n_m}}{\Rightarrow} \left(\mathcal{T}', \quad \exp\left(h'\Delta' - \frac{1}{2}h'I_0h\right)\right).$$

By implication,

$$\log \frac{dQ_{n_m}}{dP_{n_m}} \overset{P_{n_m}}{\Rightarrow} h'\Delta' - \frac{1}{2}h'I_0h \sim \mathcal{N}\left[-\frac{1}{2}h'I_0h,\ h'I_0h\right],$$

so by Le Cam's first lemma, Q_{n_m} and P_{n_m} are mutually contiguous. It now follows from the general version of Le Cam's third lemma that

$$\mathcal{T}_{n_m} \overset{Q_{n_m}}{\Rightarrow} \mathcal{T}^{h\prime},$$

where $\mathcal{T}^{h\prime}$ is a real random variable with distribution v given as

$$v(A) = \mathbb{E}\left[I_A(\mathcal{T}') \cdot e^{h'\Delta' - \frac{1}{2}h'I_0h}\right]$$

for any $A \in \mathcal{B}(\mathbb{R})$. Since we assumed that the $\mathcal{T}_n$ converges in distribution to $\mathcal{T}^h$ under Q_n, any subsequence of $\mathcal{T}_n$ converges in distribution to $\mathcal{T}^h$ under Q_n, and it follows from the uniqueness of weak limits that v is the distribution of $\mathcal{T}^h$. It remains to show that there exists a randomized statistic in the normal limiting experiment with variance I_0^{-1} that has distribution v.

By the preceding lemma, there exists some measurable function $H : \mathbb{R}^k \times [0,1] \to \mathbb{R}$ such that

$$(H(\Delta', U), \Delta') \sim (\mathcal{T}', \Delta')$$

for any uniformly distributed U that is independent of Δ'. We define $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}$ as

$$G(x,u) = H(I_0x, u)$$

for any $(x,u) \in \mathbb{R}^k \times [0,1]$.

Now choose some $X \sim \mathcal{N}\left[h, I_0^{-1}\right]$, and let U be a uniformly distributed random variable that is independent of X and Δ'. By independence, the joint density $f_{X,U} : \mathbb{R}^k \times [0,1] \to \mathbb{R}_+$ of $(X,\, U)$ is the product of the individual densities:

$$f_{X,U}(x,u) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |I_0|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x-h)' I_0 (x-h)\right)$$

for any $(x,u) \in \mathbb{R} \times [0,1]$. Similarly, the joint density $f_{\Delta',U}\mathbb{R}^k \times [0,1] \to \mathbb{R}_+$ is given as

$$f_{\Delta',U}(\delta,u) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |I_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\delta' I_0^{-1} \delta\right).$$

for any $(x,u) \in \mathbb{R} \times [0,1]$. Therefore,

$$\begin{aligned}
\mathbb{P}(G(X,\, U) \in A) &= \mathbb{P}(H(I_0 X,\, U) \in A) \\
&= \int_{\mathbb{R}\times[0,1]} I_{H^{-1}(A)}(I_0 x, u) \cdot f_{X,U}(x,u) d(x,u) \\
&= \int_0^1 \int_{\mathbb{R}} I_{H^{-1}(A)}(I_0 x, u) \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |I_0|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(x-h)' I_0 (x-h)\right) dx du \\
&= \int_0^1 \int_{\mathbb{R}} \left[I_{H^{-1}(A)}(I_0 x, u) \cdot \exp\left(h' I_0 x - \frac{1}{2} h' I_0 h\right)\right] \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |I_0|^{\frac{1}{2}} \exp\left(-\frac{1}{2} x' I_0 x\right) dx du \\
&= \int_0^1 \int_{\mathbb{R}} \left[I_{H^{-1}(A)}(\delta, u) \cdot \exp\left(h'\delta - \frac{1}{2} h' I_0 h\right)\right] \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |I_0|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \delta' I_0^{-1} \delta\right) d\delta du \\
&= \int_{\mathbb{R}\times[0,1]} \left[I_{H^{-1}(A)}(\delta, u) \cdot \exp\left(h'\delta - \frac{1}{2} h' I_0 h\right)\right] f_{\Delta',U}(\delta,u) d(\delta,u) \\
&= \mathbb{E}\left[I_B(H(\Delta', U)) \cdot e^{h'\Delta' - \frac{1}{2} h' I_0 h}\right],
\end{aligned}$$

for any $A \in \mathcal{B}(\mathbb{R})$, where we used the change of variables $x = I_0^{-1}\delta$ to justify the fourth equality. Since

$$(H(\Delta', U), \Delta') \sim (\mathcal{T}', \Delta'),$$

it follows that

$$\mathbb{P}(G(X,\, U) \in A) = \mathbb{E}\left[I_B(\mathcal{T}') \cdot e^{h'\Delta' - \frac{1}{2} h' I_0 h}\right] = v(A)$$

for any $A \in \mathcal{B}(\mathbb{R})$: we finally have

$$\mathcal{T}_n \overset{Q_n}{\Rightarrow} G(X,U).$$

Letting $\mu(h,\cdot)$ be the distribution corresponding to $\mathcal{N}\left[h, I_0^{-1}\right]$ and v the distribution of U, the distribution of $G(X,U)$ is $(\mu(h,\cdot)\times v)\circ G^{-1}$ by the independence of X and U, meaning we have just shown that

$$\mu_n(\theta_0 + r_n^{-1}h,\cdot)\circ\mathcal{T}_n^{-1} \to (\mu(h,\cdot)\times v)\circ G^{-1}.$$

Since this holds for any $h\in\mathbb{R}^k$ such that $\theta_0 + r_n^{-1}h \in \Theta$ for large enough n, out proof is complete.

Q.E.D.

While the theorem above has been formulated only for real-valued statistics, it can easily be generated to higher dimensional statistics. This involves a version of lemma 9.8 that holds for higher dimensional statistics, which can be proved via a straightforward generalization of the quantile transformation approach used in lemma 9.8.

It is important to clarify what this theorem does and does not say. The theorem says that, if the log likelihood ratio satisfies the local asymptotic normality property and the statistic of interest has some limiting distribution under any local alternative for which the LAN property holds, then under the set of local alternatives for which the LAN property holds, the limiting distribution of the statistic under each local alternative is determined by a single observation from a Gaussian location problem and an independent random component. It does not, however, tell us that this limiting distribution is Gaussian; it may even be a distribution whose name is unknown. In this sense, the term "local asymptotic normality" refers only to the asymptotic distribution of the log likelihood ratio, and not the limiting distributions of the statistics that we are interested in.

9.2.3 Differentiability in Quadratic Mean

We have established above the importance of the local asymptotic normality condition as one of two sufficient conditions required for the approximation of the distribution of a statistic by a normal limiting experiment. This section is devoted to sufficient conditions that ensure that local asymptotic normality holds. In particular, we focus on how differentiability in quadratic mean (DQM) implies local asymptotic normality when the data is iid.

We consider the following statistical framework with iid data. Let $(E,\mathcal{E})$ be a measurable space, (Θ,ϑ) the k-dimensional parameter space, and μ a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$. For each $\theta\in\Theta$, if θ is the true parameter, then the observations are drawn independently from the distribution $\mu(\theta,\cdot)$. It follows that the sequence $\{\mu_n\}_{n\in N_+}$ of statistical models is such that each μ_n is a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})^n$ and

$$\mu_n(\theta,\cdot) = \underbrace{\mu(\theta,\cdot)\times\cdots\times\mu(\theta,\cdot)}_{n}$$

for any $\theta \in \Theta$. We say that μ is differentiable in quadratic mean at $\theta_0 \in \Theta^o$ if

> There exists a measurable function $g : E \to \mathbb{R}^k$ such that $\int_E |g|^2 d\mu(\theta_0, \cdot) < +\infty$ and a σ-finite measure λ such that $\mu(\theta, \cdot) << \lambda$ for any $\theta \in \Theta$ and
>
> $$\lim_{|h| \to 0} \frac{1}{|h|^2} \int_E \left(\sqrt{d\mu(\theta_0 + h, \cdot)} - \sqrt{d\mu(\theta_0, \cdot)} - \frac{1}{2} h' g \cdot \sqrt{d\mu(\theta_0, \cdot)} \right)^2 d\lambda = 0,$$
>
> where $d\mu(\theta_0 + h, \cdot)$, $d\mu(\theta_0, \cdot)$ are the Radon-Nikodym derivatives of $\mu(\theta_0 + h, \cdot)$ and $\mu(\theta_0, \cdot)$ with respect to λ.

Since θ_0 is contained in the interior of Θ, we can see that $\theta_0 + h \in \Theta$ for small enough h, which means that DQM is well-defined. The function $g(\cdot)$ is called the score function. Indeed, it turns out that $g(\cdot)$ is exactly the score function that we know under the usual regularity conditions:

Lemma 9.10 (DQM under Regularity Conditions)
Let $(E, \mathcal{E})$ be a measurable space, (Θ, ϑ) the k-dimensional parameter space, and μ a transition probability kernel from (Θ, ϑ) into $(E, \mathcal{E})$ that determines the sequence of statistical models $\{\mu_n\}_{n \in N_+}$ under iid data. Suppose that:

i) There exists some σ-finite measure λ on $(E, \mathcal{E})$ and a jointly measurable function $p : \Theta \times E \to \mathbb{R}_+$ such that

$$\int_A p(\theta, x) d\lambda(x) = \mu(\theta, A)$$

for any $\theta \in \Theta$ and $A \in \mathcal{E}$.

ii) For any $x \in E$, $p(\cdot, x)$ is differentiable on the interior of Θ, and there exists some λ-integrable non-negative function $b : E \to \mathbb{R}_+$ such that

$$\left| \frac{\partial p(\theta, x)}{\partial \theta} \right| \leq b(x)$$

for any $(x, \theta) \in E \times \Theta^o$.

iii) For any $\theta_0 \in \Theta^o$, let $s_{\theta_0} : E \to \mathbb{R}^k$ be defined as

$$s_{\theta_0}(x) = \begin{cases} \frac{1}{p(\theta_0, x)} \cdot \frac{\partial p(\theta_0, x)}{\partial \theta} & \text{if } p(\theta_0, x) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Assume that s_{θ_0} is $\mu(\theta_0, \cdot)$-integrable, and that there exists some positive definite $k \times k$

matrix I_{θ_0} such that

$$\int_E s_{\theta_0} s'_{\theta_0} d\mu(\theta_0, \cdot) = I_{\theta_0}.$$

Then, μ is differentiable in quadratic mean at any $\theta_0 \in \Theta^o$ with score function s_{θ_0}:

$$\lim_{|h|\to 0} \frac{1}{|h|^2} \int_E \left(\sqrt{p(\theta_0 + h, \cdot)} - \sqrt{p(\theta_0, \cdot)} - \frac{1}{2} h' s_{\theta_0} \cdot \sqrt{p(\theta_0, \cdot)} \right)^2 d\lambda = 0.$$

Proof) Choose any $\theta_0 \in \Theta^o$. For any $n \in N_+$, For any $n \in N_+$,

$$\begin{aligned} \frac{1}{|h_n|^2} \left[\sqrt{p(\theta_0 + h_n, \cdot)} - \left(1 - \frac{1}{2} h'_n s_{\theta_0} \right) \sqrt{p(\theta_0, \cdot)} \right]^2 &\leq 2 \left[\frac{p(\theta_0 + h_n, \cdot) - p(\theta_0, \cdot)}{|h_n|} + \frac{1}{4} |s_{\theta_0}|^2 p(\theta_0, \cdot) \right] \\ &\leq 2 \left[b + \frac{1}{4} |s_{\theta_0}|^2 p(\theta_0, \cdot) \right], \end{aligned}$$

where

$$\int_E 2 \left[b + \frac{1}{4} |s_{\theta_0}|^2 p(\theta_0, \cdot) \right] d\lambda = 2 \left[\int_E b d\lambda + \frac{1}{4} \operatorname{tr}(I_{\theta_0}) \right] < +\infty.$$

Finally, since

$$\begin{aligned} \frac{\partial \sqrt{p(\theta_0, x)}}{\partial \theta} &= \frac{1}{2\sqrt{p(\theta_0, x)}} \times \frac{\partial p(\theta_0, x)}{\partial \theta} \\ &= \frac{1}{2} \frac{\partial \log p(\theta_0, x)}{\partial \theta} \cdot \sqrt{p(\theta_0, x)} = \frac{1}{2} s_{\theta_0}(x) \cdot \sqrt{p(\theta_0, x)} \end{aligned}$$

for any $x \in E_0$, it follows from the DCT that

$$\begin{aligned} &\lim_{n\to\infty} \int_{E_0} \frac{1}{|h_n|^2} \left(\sqrt{p(\theta_0 + h_n, \cdot)} - \sqrt{p(\theta_0, \cdot)} - \frac{1}{2} h'_n s_{\theta_0} \cdot \sqrt{p(\theta_0, \cdot)} \right)^2 d\lambda \\ &= \int_E \lim_{n\to\infty} \left(\frac{\sqrt{p(\theta_0 + h_n, \cdot)} - \sqrt{p(\theta_0, \cdot)} - \frac{1}{2} h'_n s_{\theta_0} \cdot \sqrt{p(\theta_0, \cdot)}}{|h_n|} \right)^2 d\lambda = 0. \end{aligned}$$

This holds for any sequence $\{h_n\}_{n \in N_+}$ such that $|h_n| \to 0$, so

$$\lim_{|h|\to 0} \frac{1}{|h|^2} \int_E \left(\sqrt{p(\theta_0 + h, \cdot)} - \sqrt{p(\theta_0, \cdot)} - \frac{1}{2} h' s_{\theta_0} \cdot \sqrt{p(\theta_0, \cdot)} \right)^2 d\lambda = 0.$$

Q.E.D.

The next result shows us that the DQM condition implies local asymptotic normality:

Theorem 9.11 (DQM Implies LAN)
Let $(E,\mathcal{E})$ be a measurable space, (Θ,ϑ) the k-dimensional parameter space, and μ a transition probability kernel from (Θ,ϑ) into $(E,\mathcal{E})$ that determines the sequence of statistical models $\{\mu_n\}_{n\in N_+}$ under iid data. Suppose that the DQM condition holds at $\theta_0\in\Theta$ with score function $g_0 : E\to\mathbb{R}^k$ and dominating measure λ:

$$\lim_{|h|\to 0}\frac{1}{|h|^2}\int_E\left(\sqrt{d\mu(\theta_0+h,\cdot)}-\sqrt{d\mu(\theta_0,\cdot)}-\frac{1}{2}h'g_0\cdot\sqrt{d\mu(\theta_0,\cdot)}\right)^2 d\lambda=0.$$

Then, the following hold true:

i) The mean of g_0 under $\mu(\theta_0,\cdot)$ is 0:

$$\int_E g_0 d\mu(\theta_0,\cdot)=O_{k\times 1}.$$

ii) Letting $\{X_i\}_{i\in N_+}$ be a sequence of random variables drawn iid from the sequence $\mu(\theta_0,\cdot)$, the log likelihood ratio admits the asymptotic expansion

$$\frac{d\mu_n(\theta_0+h_n/\sqrt{n},\cdot)}{d\mu_n(\theta_0,\cdot)}(X_1,\cdots,X_n)=\exp\left(h'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g_0(X_i)\right)-\frac{1}{2}h'I_0h\right)+o_p(1).$$

for any $h_n\to h$. By implication,

$$\log\frac{d\mu_n(\theta_0+h_n/\sqrt{n},\cdot)}{d\mu_n(\theta_0,\cdot)}\overset{\mu_n(\theta_0,\cdot)}{\Rightarrow}h'\Delta_0-\frac{1}{2}h'I_0h$$

for any $h_n\to h$, where $I_0=\int_E g_0g_0'd\mu(\theta_0,\cdot)$ and $\Delta_0\sim\mathcal{N}[O_{k\times 1},I_0]$.

Proof) Choose any $h_n\to h$, and denote

$$Q_n=\mu(\theta_0+h_n/\sqrt{n},\cdot)$$

$$P=\mu(\theta_0,\cdot)$$

$$q_n=\frac{\partial\mu(\theta_0+h_n/\sqrt{n},\cdot)}{\partial\lambda}$$

$$p=\frac{\partial\mu(\theta_0,\cdot)}{\partial\lambda}$$

for any $n\in N_+$. We now proceed in steps.

Step 1: Convergence Results for Densities

In this step, we establish that $\sqrt{n}\,(\sqrt{q_n}-\sqrt{p})$ converges in mean square to $\frac{1}{2}h'g_0$, and use this fact to show that $\int_E h'g_0 dP = 0$.

By the DQM property,

$$\lim_{n\to\infty} \frac{1}{|h_n/\sqrt{n}|^2} \int_E \left(\sqrt{q_n}-\sqrt{p}-\frac{1}{2}\frac{h_n'}{\sqrt{n}}g_0\cdot\sqrt{p}\right)^2 d\lambda = 0;$$

since $h_n \to h$, this implies

$$\lim_{n\to\infty} \int_E \left[\sqrt{n}\,(\sqrt{q_n}-\sqrt{p})-\frac{1}{2}h_n'g_0\cdot\sqrt{p}\right]^2 d\lambda = 0.$$

Furthermore,

$$\int_E \left[\sqrt{n}\,(\sqrt{q_n}-\sqrt{p})-\frac{1}{2}h'g_0\cdot\sqrt{p}\right]^2 d\lambda$$

$$\leq 2\int_E \left[\sqrt{n}\,(\sqrt{q_n}-\sqrt{p})-\frac{1}{2}h_n'g_0\cdot\sqrt{p}\right]^2 d\lambda + \frac{1}{2}|h_n-h|^2\int_E |g_0|^2 dP \to 0$$

as $n\to\infty$, which tells us that

$$\sqrt{n}\,(\sqrt{q_n}-\sqrt{p}) \overset{L^2(\mathcal{E},\lambda)}{\to} \frac{1}{2}h'g_0\cdot\sqrt{p}$$

as $n\to\infty$. Multiplying both sides by $\frac{1}{\sqrt{n}}$ also shows us that

$$\sqrt{q_n}-\sqrt{p} \overset{L^2(\mathcal{E},\lambda)}{\to} 0$$

as $n\to\infty$, or equivalently,

$$\lim_{n\to\infty}\int_E (\sqrt{q_n}-\sqrt{p})\,d\lambda = 0.$$

Letting $\left\langle\cdot,\cdot\right\rangle_\lambda$ be the usual inner product on the L^2 space $L^2(\mathcal{E},p)$, it follows, from the continuity of the inner product in both arguments with respect to the metric induced by the inner product, that

$$\int_E h'g_0\,dP = \int_E \left(\frac{1}{2}h'g_0\cdot\sqrt{p}\right)(2\sqrt{p})\,d\lambda = \left\langle \frac{1}{2}h'g_0\cdot\sqrt{p}, 2\sqrt{p}\right\rangle_\lambda$$

$$= \lim_{n\to\infty}\left\langle \sqrt{n}\,(\sqrt{q_n}-\sqrt{p}),\ \sqrt{q_n}+\sqrt{p}\right\rangle_\lambda$$

$$= \lim_{n\to\infty}\sqrt{n}\int_E (\sqrt{q_n}-\sqrt{p})\,(\sqrt{q_n}-\sqrt{p})\,d\lambda = \lim_{n\to\infty}\sqrt{n}\int_E (q_n-p)\,d\lambda = 0$$

because densities integrate to 1.

Random Variables Constructed using the Likelihood Ratio

In this step, we construct a triangular array $\{W_{ni}\}_{1\leq i\leq n, n\in N_+}$ of random variables using the likelihood ratio $\frac{q_n}{p}$ which will play a crucial role in the final asymptotic step.

Let $\{X_i\}_{i\in N_+}$ be a sequence of iid random variables taking values in $(E,\mathcal{E})$ with distribution $P=\mu(\theta_0,\cdot)$, and define

$$E_0=\{x\in E\mid p(x)>0\}$$

and

$$W_{ni}=2\left(\sqrt{\frac{q_n}{p}(X_i)}-1\right)$$

for any $n\in N_+$ and $1\leq i\leq n$. As is customary, we define

$$\frac{q_n}{p}(x)=\begin{cases}\frac{q_n(x)}{p(x)} & \text{if } x\in E_0\\ 0 & \text{otherwise}\end{cases},$$

where $P(E_0)=1$. We want to obtain asymptotic expansions of the following terms:

$$\sum_{i=1}^n W_{ni},\quad \sum_{i=1}^n W_{ni}^2,\quad \max_{1\leq i\leq n}|W_{ni}|.$$

Step 2-1: Asymptotic Expansion of the Sum

For any $n\in N_+$,

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^n W_{ni}-\frac{1}{\sqrt{n}}\sum_{i=1}^n h'g_0(X_i)\right]&=\mathrm{Var}\left[\sum_{i=1}^n\left(W_{ni}-\frac{1}{\sqrt{n}}h'g_0(X_i)\right)\right]\\&=\sum_{i=1}^n\mathrm{Var}\left[W_{ni}-\frac{1}{\sqrt{n}}h'g_0(X_i)\right]\\&\qquad\qquad\text{(Independence of } (X_1,\cdots,X_n))\\&\leq\sum_{i=1}^n\mathbb{E}\left|W_{ni}-\frac{1}{\sqrt{n}}h'g_0(X_i)\right|^2\\&=4\int_E\left(\sqrt{n}\left(\sqrt{\frac{q_n}{p}}-1\right)-\frac{1}{2}h'g_0\right)^2dP\\&=4\int_{E_0}\left(\sqrt{n}\left(\sqrt{\frac{q_n}{p}}-1\right)-\frac{1}{2}h'g_0\right)^2dP\\&\qquad\qquad\text{(Because } P(E_0)=1)\end{aligned}$$

$$= 4\int_{E_0}\left(\sqrt{n}(\sqrt{q_n}-\sqrt{p})-\frac{1}{2}h'g_0\cdot\sqrt{p}\right)^2 d\lambda$$

(Because $\frac{q_n}{p}\cdot p = q_n$ on E_0)

$$\leq 4\int_{E}\left(\sqrt{n}(\sqrt{q_n}-\sqrt{p})-\frac{1}{2}h'g_0\cdot\sqrt{p}\right)^2 d\lambda$$

(Monotonicity of Integration and $E_0 \subset E$)

Taking $n\to\infty$ on both sides, we can see that

$$\lim_{n\to\infty}\operatorname{Var}\left[\sum_{i=1}^{n}W_{ni}-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h'g_0(X_i)\right]=0.$$

Similarly,

$$\mathbb{E}\left[\sum_{i=1}^{n}W_{ni}\right] = 2n\left(\int_E\sqrt{\frac{q_n}{p}}dP-1\right)$$

$$= 2n\left(\int_{E_0}\sqrt{\frac{q_n}{p}}dP-1\right)$$

$$= 2n\left(\int_{E_0}(\sqrt{q_n}\sqrt{p})\,d\lambda-1\right)$$

$$= 2n\left(\int_{E}(\sqrt{q_n}\sqrt{p})\,d\lambda-1\right),$$

where the last equality is justified because $p=0$ on E_0^c, so that

$$\int_{E_0^c}(\sqrt{q_n}\sqrt{p})\,d\lambda=0.$$

It follows that

$$\mathbb{E}\left[\sum_{i=1}^{n}W_{ni}\right] = -n\int_E(\sqrt{q_n}-\sqrt{p})^2\,d\lambda = -\int_E\left[\sqrt{n}(\sqrt{q_n}-\sqrt{p})\right]^2 d\lambda$$

where we used the fact that the densities q_n, p integrate to 1 under λ. Since

$$\sqrt{n}(\sqrt{q_n}-\sqrt{p}) \overset{L^2(\mathcal{E},\lambda)}{\to} \frac{1}{2}h'g_0\cdot\sqrt{p},$$

it follows that

$$\lim_{n\to\infty}\mathbb{E}\left[\sum_{i=1}^{n}W_{ni}\right] = -\frac{1}{4}\int_E(h'g_0)^2 p\,d\lambda = -\frac{1}{4}\int_E(h'g_0)^2 dP.$$

Therefore,

$$\sum_{i=1}^{n}W_{ni} = h'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_0(X_i)\right)-\frac{1}{4}\int_E(h'g_0)^2dP+r_n,$$

where $\mathbb{E}[r_n] = 0$ and $\text{Var}[r_n] \to 0$ as $n \to \infty$, and as such $r_n \xrightarrow{L^2} 0$ as $n \to \infty$. Mean square convergence implies convergence in probability, and $\int_E (h'g_0)^2 dP = h' \left(\int_E g_0 g_0' dP \right) h = h' I_0 h$, so

$$\sum_{i=1}^n W_{ni} = h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n g_0(X_i) \right) - \frac{1}{4} h' I_0 h + o_p(1).$$

Step 2-2: Asymptotic Expansion of the Sum of Squares

We now shift our focus to the sum of squares $\sum_{i=1}^n W_{ni}^2$. We already saw above that

$$\mathbb{E}\left| \sqrt{n} W_{ni} - h' g_0(X_i) \right|^2 = 4 \int_E \left(\sqrt{n} \left(\sqrt{\frac{q_n}{p}} - 1 \right) - \frac{1}{2} h' g_0 \right)^2 dP \to 0$$

as $n \to \infty$. Defining

$$A_{ni} = \sqrt{n} W_{ni} - h' g_0(X_i),$$

we can express the sum of squares of W_{ni} as

$$\begin{aligned} \sum_{i=1}^n W_{ni}^2 &= \frac{1}{n} \sum_{i=1}^n \left(A_{ni} + h' g_0(X_i) \right)^2 \\ &= h' \left(\frac{1}{n} \sum_{i=1}^n g_0(X_i) g_0(X_i)' \right) h + 2h' \left(\frac{1}{n} \sum_{i=1}^n g_0(X_i) A_{ni} \right) + \frac{1}{n} \sum_{i=1}^n A_{ni}^2. \end{aligned}$$

By Khinchin's LLN for iid sequences of random variables with finite first moments,

$$h' \left(\frac{1}{n} \sum_{i=1}^n g_0(X_i) g_0(X_i)' \right) h \xrightarrow{p} h' I_0 h.$$

Meanwhile,

$$\mathbb{E}\left| \frac{1}{n} \sum_{i=1}^n A_{ni}^2 \right| \leq \mathbb{E}|A_{n1}|^2 \to 0$$

as $n \to \infty$, so that

$$\frac{1}{n} \sum_{i=1}^n A_{ni}^2 \xrightarrow{L^1} 0.$$

Finally, by Hölder's inequality,

$$\mathbb{E}\left| \frac{1}{n} \sum_{i=1}^n g_0(X_i) A_{ni} \right| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[|g_0(X_i)| \cdot |A_{ni}| \right]$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}|g_0(X_i)|^2\right)^{\frac{1}{2}}\left(\mathbb{E}|A_{ni}|^2\right)^{\frac{1}{2}}$$

$$= \sqrt{\mathrm{tr}(I_0)}\cdot\left(\mathbb{E}|A_{n1}|^2\right)^{\frac{1}{2}} \to 0$$

as $n \to \infty$, so that

$$\frac{1}{n}\sum_{i=1}^{n} g_0(X_i)A_{ni} \xrightarrow{L^1} O_{k\times 1}.$$

Therefore,

$$\sum_{i=1}^{n} W_{ni}^2 = h'\left(\frac{1}{n}\sum_{i=1}^{n} g_0(X_i)g_0(X_i)'\right)h + o_p(1) \xrightarrow{p} h'I_0h.$$

Step 2-3: Asymptotic Expansion of the Maximum

It remains to study the properties of the sequence $\{\max_{1\leq i\leq n}|W_{ni}|\}_{n\in N_+}$ of random variables. For any $\epsilon > 0$,

$$\mathbb{P}\left(\max_{1\leq i\leq n}|W_{ni}| > \epsilon\right) \leq \sum_{i=1}^{n}\mathbb{P}\left(|W_{ni}| > \epsilon\right)$$

$$\leq \sum_{i=1}^{n}\mathbb{P}\left(|A_{ni}| > \frac{\epsilon\sqrt{n}}{2}\right) + \sum_{i=1}^{n}\mathbb{P}\left(\left|h'g_0(X_i)\right| > \frac{\epsilon\sqrt{n}}{2}\right)$$
$$\left(W_{ni} = \tfrac{1}{\sqrt{n}}\left(A_{ni} + h'g_0(X_i)\right) \text{ for } 1 \leq i \leq n\right)$$

$$\leq \frac{4}{\epsilon^2}\times\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|A_{ni}|^2 + n\mathbb{P}\left(\left|h'g_0(X_1)\right| > \frac{\epsilon\sqrt{n}}{2}\right)$$

$$= \frac{4}{\epsilon^2}\mathbb{E}|A_{n1}|^2 + n\cdot\mathbb{P}\left(\left|h'g_0(X_1)\right|^2 > \frac{\epsilon^2 n}{4}\right).$$

Letting Y be a non-negative random variable and $b > 0$, since

$$b\cdot I_{\{Y>b\}} \leq Y\cdot I_{\{Y>b\}},$$

it follows that

$$b\cdot\mathbb{P}\left(Y > b\right) \leq \mathbb{E}\left[Y\cdot I_{\{Y>b\}}\right].$$

As such,

$$\frac{\epsilon^2 n}{4} \cdot \mathbb{P}\left(\left|h' g_0(X_1)\right|^2 > \frac{\epsilon^2 n}{4} \right) \leq \mathbb{E}\left[\left|h' g_0(X_1)\right|^2 \cdot I_{\{|h' g_0(X_1)|^2 > \frac{\epsilon^2 n}{4}\}} \right]$$

$$= \int_E \left[\left|h' g_0\right|^2 \cdot I_{\{|h' g_0|^2 > (\epsilon^2 n)/4\}} \right] dP,$$

or equivalently,

$$n \cdot \mathbb{P}\left(\left|h' g_0(X_1)\right|^2 > \frac{\epsilon^2 n}{4} \right) \leq \frac{4}{\epsilon^2} \int_E \left[\left|h' g_0\right|^2 \cdot I_{\{|h' g_0|^2 > (\epsilon^2 n)/4\}} \right] dP.$$

Since $|h' g_0|^2$ is P-integrable and $\frac{\epsilon^2 n}{4} \to +\infty$ as $n \to \infty$, it follows from the characterization of integrable functions such that

$$\lim_{n \to \infty} \int_E \left[\left|h' g_0\right|^2 \cdot I_{\{|h' g_0|^2 > (\epsilon^2 n)/4\}} \right] dP = 0,$$

which implies that $n \cdot \mathbb{P}\left(\left|h' g_0(X_1)\right|^2 > \frac{\epsilon^2 n}{4} \right) \to 0$ as $n \to \infty$.

Meanwhile, we already established above that $\mathbb{E}|A_{n1}|^2 \to 0$ as $n \to \infty$, so that

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{1 \leq i \leq n} |W_{ni}| > \epsilon \right) = 0.$$

This holds for any $\epsilon > 0$, so

$$\max_{1 \leq i \leq n} |W_{ni}| \xrightarrow{p} 0$$

as $n \to \infty$.

Step 3: Taylor Expansion of the Exponential Function

In this final step, we use a second-order Taylor expansion of an exponential function to establish the desired LAN result.

Defining $X^n = (X_1, \cdots, X_n)$ for each $n \in N_+$, it follows that $\{X^n\}_{n \in N_+}$ is a sequence of random variables such that each X^n has distribution $\mu_n(\theta_0, \cdot)$. We want to show that the likelihood ratio can be asymptotically expanded as

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) = \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) + o_p(1).$$

For any $n \in N_+$, the mapping

$$(x_1, \cdots, x_n) \mapsto \prod_{i=1}^{n} q_n(x_i)$$

is the Radon-Nikodym derivative of $\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)$ with respect to λ; similarly, the mapping

$$(x_1, \cdots, x_n) \mapsto \prod_{i=1}^{n} p(x_i)$$

is the density of $\mu_n(\theta_0, \cdot)$ with respect to λ. By lemma 9.1, it follows that

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(x_1, \cdots, x_n) = \prod_{i=1}^{n} \frac{q_n}{p}(x_i)$$

for any $(x_1, \cdots, x_n) \in E_0^n$, where

$$\mu_n(\theta_0, E_0^n) = \prod_{i=1}^{n} P(E_0) = 1.$$

$\mu_n(\theta_0, \cdot)$ is the distribution of X^n, so

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) = \prod_{i=1}^{n} \frac{q_n}{p}(X_i)$$

almost surely for any $n \in N_+$. Therefore, our problem reduces to showing that

$$\prod_{i=1}^{n} \frac{q_n}{p}(X_i) = \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) + o_p(1),$$

since

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) - \prod_{i=1}^{n} \frac{q_n}{p}(X_i) = o_p(1)$$

as $n \to \infty$.

To this end, we utilize the following Taylor expansion of the mapping

$$x \mapsto e^{x-\frac{1}{2}x^2}$$

around 0. This function has first and second derivatives equal to

$$x \mapsto (1-x)e^{x-\frac{1}{2}x^2} \quad \text{and} \quad x \mapsto (1-x)^2 e^{x-\frac{1}{2}x^2} - e^{x-\frac{1}{2}x^2},$$

which, evaluated at 0, are equal to 1 and 0, respectively. Therefore, the second order Taylor expansion of the above mapping around 0 is given as

$$e^{x-\frac{1}{2}x^2} = 1 + x + R(x),$$

where

$$\lim_{x\to 0} \frac{|R(x)|}{x^2} = 0.$$

Letting

For any $n \in N_+$ and $1 \le i \le n$,

$$\sqrt{\frac{q_n}{p}(X_i)} = \left(\frac{1}{2}W_{ni} + 1\right) = \exp\left(\frac{1}{2}W_{ni} - \frac{1}{8}W_{ni}^2\right) - R\left(\frac{1}{2}W_{ni}\right),$$

so that

$$\begin{aligned}\left(\prod_{i=1}^n \frac{q_n}{p}(X_i)\right)^{\frac{1}{2}} &= \prod_{i=1}^n \left[\exp\left(\frac{1}{2}W_{ni} - \frac{1}{4}W_{ni}^2\right) - R\left(\frac{1}{2}W_{ni}\right)\right] \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^n W_{ni} - \frac{1}{8}\sum_{i=1}^n W_{ni}^2\right) + R_n,\end{aligned}$$

where

$$R_n = \prod_{i=1}^n \left[\exp\left(\frac{1}{2}W_{ni} - \frac{1}{8}W_{ni}^2\right) - R\left(\frac{1}{2}W_{ni}\right)\right] - \prod_{i=1}^n \exp\left(\frac{1}{2}W_{ni} - \frac{1}{8}W_{ni}^2\right).$$

Step 3-1: Asymptotic Negligibility of the Remainder Term

We deal first with the remainder term. Note that the mapping $x \mapsto e^{x-\frac{1}{2}x^2}$ attains its maximum at $x = 1$, at which its value is $\sqrt{e}$. Therefore,

$$\frac{1}{2\sqrt{e}} e^{x-\frac{1}{2}x^2} \le \frac{1}{2}$$

for any $x \in \mathbb{R}$. Since $R(x) \to 0$ as $x \to 0$, there exists some $\eta > 0$ such that $|R(x)| < \sqrt{e}$ for any $x \in \mathbb{R}$ such that $|x| \leq \eta$. Defining the set

$$H_n = \{\max_{1 \leq i \leq n} |W_{ni}| \leq \eta\}.$$

for any $n \in N_+$, it follows that, on H_n, $\left|R\left(\frac{1}{2}W_{ni}\right)\right| < \sqrt{e}$ for any $1 \leq i \leq n$, and by what we proved above,

$$\lim_{n \to \infty} \mathbb{P}(H_n) = 1.$$

For any $\epsilon > 0$ and $n \in N_+$,

$$\begin{aligned}
\mathbb{P}(|R_n| > \epsilon) &= \mathbb{P}\left(\frac{1}{2\sqrt{e}}|R_n| > \frac{\epsilon}{2\sqrt{e}}\right) \\
&\leq \mathbb{P}\left(\frac{1}{2\sqrt{e}}|R_n| \cdot I_{H_n} > \frac{\epsilon}{4\sqrt{e}}\right) + \mathbb{P}\left(\frac{1}{2\sqrt{e}}|R_n| \cdot I_{H_n^c} > \frac{\epsilon}{4\sqrt{e}}\right) \\
&\leq \mathbb{P}\left(\frac{1}{2\sqrt{e}}|R_n| \cdot I_{H_n} > \frac{\epsilon}{4\sqrt{e}}\right) + \mathbb{P}(H_n^c),
\end{aligned}$$

where the final inequality follows because $\frac{1}{2\sqrt{e}}|R_n| \cdot I_{H_n^c} > \frac{\epsilon}{4\sqrt{e}} > 0$ implies that $I_{H_n^c} = 1$. Focusing on the first term, choose any $\omega \in H_n$. Then, since

$$\frac{1}{2\sqrt{e}}\left|R\left(\frac{1}{2}W_{ni}(\omega)\right)\right| < \frac{1}{2}$$

for any $1 \leq i \leq n$, it follows that

$$\begin{aligned}
\frac{1}{2\sqrt{e}}|R_n(\omega)| &= \left|\prod_{i=1}^{n}\left[\frac{1}{2\sqrt{e}}\exp\left(\frac{1}{2}W_{ni}(\omega) - \frac{1}{8}W_{ni}(\omega)^2\right) - \frac{1}{2\sqrt{e}}R\left(\frac{1}{2}W_{ni}(\omega)\right)\right] - \prod_{i=1}^{n}\frac{1}{2\sqrt{e}}\exp\left(\frac{1}{2}W_{ni}(\omega) - \frac{1}{8}W_{ni}(\omega)^2\right)\right| \\
&\leq \frac{1}{2\sqrt{e}}\sum_{i=1}^{n}\left|R\left(\frac{1}{2}W_{ni}(\omega)\right)\right|,
\end{aligned}$$

where we used the fact that

$$\left[\frac{1}{2\sqrt{e}}\exp\left(\frac{1}{2}W_{ni}(\omega) - \frac{1}{8}W_{ni}(\omega)^2\right) - \frac{1}{2\sqrt{e}}R\left(\frac{1}{2}W_{ni}(\omega)\right)\right], \quad \frac{1}{2\sqrt{e}}R\left(\frac{1}{2}W_{ni}(\omega)\right)$$

are contained in the unit circle for each $1 \leq i \leq n$ and that

$$\left|\prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i\right| \leq \sum_{i=1}^{n}|z_i - w_i|$$

for any $z_1, \cdots, z_n, w_1, \cdots, w_n \in \mathbb{C}$ contained in the unit circle. Therefore,

$$\frac{1}{2\sqrt{e}}|R_n| \cdot I_{H_n} \leq \frac{1}{2\sqrt{e}}\sum_{i=1}^{n}\left|R\left(\frac{1}{2}W_{ni}\right)\right|.$$

Defining the function $r : \mathbb{R} \to \mathbb{R}$ as

$$r(x) = \begin{cases} \frac{R(x)}{x^2} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

it follows that $r(x) \to 0$ as $x \to 0$ and $R(x) = x^2 r(x)$ for any $x \in \mathbb{R}$ because $R(0) = 0$. As such,

$$\begin{aligned} \frac{1}{2\sqrt{e}} |R_n| \cdot I_{H_n} &\leq \frac{1}{8\sqrt{e}} \sum_{i=1}^{n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| \cdot W_{ni}^2 \\ &\leq \frac{1}{8\sqrt{e}} \max_{1 \leq i \leq n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| \left(\sum_{i=1}^{n} W_{ni}^2 \right). \end{aligned}$$

For any $\zeta > 0$, there exists a $\xi > 0$ such that

$$|r(x)| \leq \zeta \quad \text{for any } x \in \mathbb{R} \text{ s.t. } |x| \leq \xi.$$

It follows that, if $\max_{1 \leq i \leq n} |W_{ni}| \leq \xi$, then $\max_{1 \leq i \leq n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| \leq \zeta$, so that

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| > \zeta \right) \leq \mathbb{P}\left(\max_{1 \leq i \leq n} |W_{ni}| > \xi \right).$$

Since $\max_{1 \leq i \leq n} |W_{ni}| \xrightarrow{p} 0$ as $n \to \infty$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{1 \leq i \leq n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| > \zeta \right) = 0,$$

and because this holds for any $\zeta > 0$,

$$\max_{1 \leq i \leq n} \left| r\left(\frac{1}{2} W_{ni}\right) \right| \xrightarrow{p} 0.$$

We know that $\sum_{i=1}^{n} W_{ni}^2 = O_p(1)$ as $n \to \infty$, so

$$\frac{1}{2\sqrt{e}} |R_n| \cdot I_{H_n} \leq o_p(1) O_p(1) = o_p(1),$$

which implies that $\frac{1}{2\sqrt{e}} |R_n| \cdot I_{H_n} = o_p(1)$ as $n \to \infty$ as well. By implication,

$$\limsup_{n \to \infty} \mathbb{P}\left(|R_n| > \epsilon \right) \leq \lim_{n \to \infty} \mathbb{P}\left(\frac{1}{2\sqrt{e}} |R_n| \cdot I_{H_n} > \frac{\epsilon}{4\sqrt{e}} \right) + \lim_{n \to \infty} \mathbb{P}\left(H_n^c \right) = 0,$$

which implies that

$$\lim_{n \to \infty} \mathbb{P}\left(|R_n| > \epsilon \right) = 0.$$

This in turn holds for any $\epsilon > 0$, so $R_n = o_p(1)$ as $n \to \infty$.

Step 3-2: Asymptotic Expansion of the Likelihood Ratio

Recall how we showed that

$$\sum_{i=1}^{n} W_{ni} = h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{4} h' I_0 h + o_p(1),$$

$$\sum_{i=1}^{n} W_{ni}^2 \xrightarrow{p} h' I_0 h$$

as $n \to \infty$. By the Lindeberg-Levy CLT, it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \xrightarrow{d} \Delta_0 \sim \mathcal{N}\left[O_{k\times 1},\ I_0\right],$$

so that

$$\sum_{i=1}^{n} W_{ni} \xrightarrow{d} h' \Delta_0 - \frac{1}{4} h' I_0 h$$

by Slutsky's theorem. By implication,

$$\exp\left(\frac{1}{2} \sum_{i=1}^{n} W_{ni} - \frac{1}{8} \sum_{i=1}^{n} W_{ni}^2 \right) \xrightarrow{d} \exp\left(\frac{1}{2} h' \Delta_0 - \frac{1}{4} h' I_0 h \right) = O_p(1)$$

as $n \to \infty$.

From the preceding expansions, we can show that the likelihood ratio can be expanded as

$$\begin{aligned} \prod_{i=1}^{n} \frac{q_n}{p}(X_i) &= \left(\exp\left(\frac{1}{2} \sum_{i=1}^{n} W_{ni} - \frac{1}{8} \sum_{i=1}^{n} W_{ni}^2 \right) + R_n \right)^2 \\ &= \exp\left(\frac{1}{2} \sum_{i=1}^{n} W_{ni} - \frac{1}{8} \sum_{i=1}^{n} W_{ni}^2 \right)^2 + o_p(1) \\ &= \exp\left(\sum_{i=1}^{n} W_{ni} - \frac{1}{4} \sum_{i=1}^{n} W_{ni}^2 \right) + o_p(1), \end{aligned}$$

so that

$$\prod_{i=1}^{n} \frac{q_n}{p}(X_i) = \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h + \epsilon_n \right) + o_p(1),$$

where $\epsilon_n = o_p(1)$ as $n \to \infty$. Since

$$\exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) \xrightarrow{d} \exp\left(h' \Delta_0 - \frac{1}{2} h' I_0 h \right),$$

and $\exp(\epsilon_n) - 1 = o_p(1)$ as $n \to \infty$ by the continuous mapping theorem,

$$\prod_{i=1}^{n} \frac{q_n}{p}(X_i) - \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right)$$

$$= \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) [\exp(\epsilon_n) - 1] + o_p(1)$$

$$= O_p(1) o_p(1) + o_p(1) = o_p(1)$$

as $n \to \infty$. This proves that

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) = \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) + o_p(1),$$

and by Slutsky's theorem

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) \xrightarrow{d} \exp\left(h' \Delta_0 - \frac{1}{2} h' I_0 h \right),$$

or equivalently,

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} \overset{\mu_n(\theta_0, \cdot)}{\Rightarrow} \exp\left(h' \Delta_0 - \frac{1}{2} h' I_0 h \right).$$

This holds for any $h_n \to h$, so the proof is complete.

Q.E.D.

9.3 Asymptotic Efficiency for Parametric Models

In this section we use the machinery developed above, in particular Le Cam's lemmas and approximation via normal limiting experiments, to formalize the notion of efficient estimation in an asymptotic context for models with finite-dimensional parameters. We briefly touched upon asymptotic efficiency in the section on maximum likelihood methods, where we claimed that the MLE, under the appropriate regularity conditions, is asymptotically efficient because its asymptotic variance is equal to the asymptotic version of the CRLB. Efficiency was brought up again in the context of GMM estimation, where we derived the efficient GMM estimator as the estimator with the smallest asymptotic variance among GMM estimators using the same moment conditions.

While these examples may lead us to conflate minimizing the asymptotic variance with efficiency, in reality the concept of asymptotic efficiency demands much more subtlety. For one, while we claimed the MLE is asymptotically efficient, there is no reason to believe a prior that a limiting normal distribution is the "best" limiting distribution. Furthermore, we have not even established that the CRLB is the smallest asymptotic variance that can be achieved by some class of estimators with normal limiting distributions.

9.3.1 A Criterion for Efficiency

We define asymptotic efficiency in terms of how concentrated the limiting distribution is around 0. The mathematical framework is as follows. Suppose that (Θ, ϑ) is the k-dimensional parameter space, and that $\psi : \Theta \to \mathbb{R}^d$ is some measurable function. We are interested in the value $\psi(\theta)$ given the parameter θ. Suppose that $\{(E_n, \mathcal{E}_n)\}_{n \in N_+}$ is a sequence of measurable spaces and $\{\mu_n\}_{n \in N_+}$ the corresponding sequence of transition probability kernels representing statistical models for each sample size.

We focus on statistics $\mathcal{T}_n$ such that, for some $\theta_0 \in \Theta^o$ and any $h \in \mathbb{R}^k$, there exists a random vector $\mathcal{T}_{\theta_0,h}$ such that

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} \mathcal{T}_{\theta_0,h}.$$

Letting $L_{\theta_0,h}$ be the distribution of $\mathcal{T}_{\theta_0,h}$, this is equivalent to requiring that

$$\mu_n(\theta_0 + h/\sqrt{n}, \cdot) \circ \left(\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right]\right)^{-1} \to L_{\theta_0,h}$$

weakly as $n \to \infty$ for any $h \in \mathbb{R}^k$. We call such statistics smooth at θ_0. If the limiting distribution $L_{\theta_0,h}$ does not depend on h, then we say that $\mathcal{T}_n$ is regular at θ_0. Regular statistics also play an important role later on.

We want the limiting distributions $L_{\theta_0,h}$ to be as concentrated around 0 as possible across different values of h, so that the asymptotic bias and variance are minimized locally and uniformly around θ_0. Formally, the degree of concentration is measured via the use of loss functions,

as in decision theory for finite sample problems. Let $l : \mathbb{R}^d \to \mathbb{R}_+$ be some measurable function that is minimized at 0 and increase the farther we move away from 0. Then, under the loss function $l(\cdot)$, the asymptotic risk associated with the limiting distribution $L_{\theta_0,h}$ is defined as

$$\int_{\mathbb{R}^d} l(x) dL_{\theta_0,h}(x).$$

Our goal is to search for the $L_{\theta_0,h}$ under which this risk is suitably minimized across different values of h. As in statistical decision theory for fixed finite sample sizes, we can think of different criteria that allow us to optimally choose the collection $\{L_{\theta_0,h}\}_{h\in\mathbb{R}^k}$ of limiting distributions:

1. **Bayes Criterion**
   Let π be a distribution on $\mathbb{R}^k$. We may search for the collection $\{L_{\theta_0,h}\}_{h\in\mathbb{R}^k}$ of limiting distributions that minimizes the Bayes criterion

   $$\int_{\mathbb{R}^k} \int_{\mathbb{R}^d} l \; dL_{\theta_0,h}(x) d\pi(h),$$

   granted that the mapping

   $$h \mapsto \int_{\mathbb{R}^d} l \; dL_{\theta_0,h}(x)$$

   is Borel measurable.

2. **Minimax Criterion**
   We may search for the collection $\{L_{\theta_0,h}\}_{h\in\mathbb{R}^k}$ of limiting distributions that minimizes the maximum asymptotic risk

   $$\sup_{h\in\mathbb{R}^k} \int_{\mathbb{R}^d} l \; dL_{\theta_0,h}(x).$$

To make the problem more tractable, we must impose more structure on the statistical problem at hand. Once again, local asymptotic normality proves useful to this end, as the following theorem shows. Specifically, smooth statistics in a statistical model that satisfies the LAN property have local limiting distributions that are location shifts of randomized statistics from a Gaussian location problem.

Theorem 9.12 (Smooth Statistics under Local Asymptotic Normality)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces, (Θ, ϑ) the k-dimensional parameter space and $\{\mu_n\}_{n\in N_+}$ a sequence of statistical models. Let $\psi : \Theta \to \mathbb{R}^d$ be a function that is differentiable at $\theta_0 \in \Theta^o$, and denote by $\dot{\psi}(\theta_0)$ the derivative of ψ at θ_0. Let $\{\mathcal{T}_n\}_{n\in N_+}$ be the sequence of d-dimensional statistics of interest. Suppose that the following hold true:

i) **(Smooth Statistics)**
$\mathcal{T}_n$ is smooth at θ_0, so that

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} \mathcal{T}_{\theta_0,h} \sim L_{\theta_0,h}$$

for each $h \in \mathbb{R}^k$.

ii) **(Local Asymptotic Normality)**
$\{\mu_n\}_{n\in N_+}$ is locally asymptotically normal at θ_0, that is, there exists some positive definite $k \times k$ matrix I_0, a sequence $\{\Delta_n\}_{n\in N_+}$ where each $\Delta_n : E_n \to \mathbb{R}^k$ is a measurable function, and $\Delta_0 \sim \mathcal{N}[O_{k\times 1},\ I_0]$ such that $\Delta_n \overset{0}{\Rightarrow} \Delta_0$ and

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} - \exp\left(h'\Delta_n - \frac{1}{2}h'I_0h\right) \overset{0}{\Rightarrow} 0$$

for any sequence $\{h_n\}_{n\in N_+}$ in $\mathbb{R}^k$ converging to $h \in \mathbb{R}^k$.

Then, there exists some measurable function $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ such that, for any observation X drawn from the Gaussian location model with variance I_0^{-1} and $U \sim \text{Uniform}[0,1]$ independent of X, we have

$$\mathcal{T}_{\theta_0,h} \overset{h}{\sim} G(X,U) - \dot{\psi}(\theta_0)h$$

for any $h \in \mathbb{R}^k$.

Proof) Choose any non-zero $h \in \mathbb{R}^k$. By the definition of differentiability,

$$\lim_{n\to\infty} \frac{\left|\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0) - \dot{\psi}(\theta_0)\frac{h}{\sqrt{n}}\right|}{\frac{1}{\sqrt{n}}} = 0,$$

which implies that

$$\lim_{n\to\infty} \sqrt{n}\left[\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right] = \dot{\psi}(\theta_0)h.$$

It follows from Slutsky's theorem that

$$\sqrt{n}\,[\mathcal{T}_n - \psi(\theta_0)] = \sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] + \sqrt{n}\left[\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right]$$

$$\stackrel{h}{\Rightarrow} \mathcal{T}_{\theta_0,h} + \dot{\psi}(\theta_0)h.$$

Meanwhile,

$$\sqrt{n}\,[\mathcal{T}_n - \psi(\theta_0)] \stackrel{0}{\Rightarrow} \mathcal{T}_{\theta_0,0}.$$

Since $\{\mu_n\}_{n\in N_+}$ satisfies the LAN property and

$$\sqrt{n}\,[\mathcal{T}_n - \psi(\theta_0)]$$

converges in distribution under any $h \in \mathbb{R}^k$, by the main approximation theorem, theorem 9.9, it follows that there exists some measurable function $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ such that, for any $h \in \mathbb{R}^k$,

$$\mathcal{T}_{\theta_0,h} + \dot{\psi}(\theta_0)h \stackrel{h}{\sim} G(X,U),$$

where X is a draw from the Gaussian location model with variance I_0^{-1} and U is a uniformly distributed random variable that is independent of X. This is equivalent to

$$\mathcal{T}_{\theta_0,h} \stackrel{h}{\sim} G(X,U) - \dot{\psi}(\theta_0)h$$

for any $h \in \mathbb{R}^k$, which is the result we desired.

Q.E.D.

The local asymptotic normality of the model at θ_0 and the weak convergence of $\mathcal{T}_n$ under the null not only ensure that $\mathcal{T}_n$ is smooth at θ_0, but also shows us that the limiting distributions $L_{\theta_0,h}$ must be given as the convolution of a randomized statistic $G(X,U)$ from a Gaussian location model and a non-random location shift $\dot{\psi}(\theta_0)h$. This means that the asymptotic efficiency of $\mathcal{T}_n$ around the point θ_0 depends entirely on the behavior of the convolution

$$G(X,U) - \dot{\psi}(\theta_0)h$$

for each value of $h \in \mathbb{R}^k$, that is, on the distribution

$$L_{\theta_0,h}(G) = \left[(\mu(h,\cdot) \times v) \circ G^{-1}\right] * \delta_{-\dot{\psi}(\theta_0)h},$$

where μ is the single-observation Gaussian location model with variance I_0^{-1} and v is the uniform distribution on $[0,1]$.

As such, when we restrict our focus to smooth statistics in models that satisfy the LAN property, our objective can be simplified to searching for the statistic $G(\cdot)$ that makes is the "best" in the sense of minimizing the asymptotic risk

$$\mathcal{R}(h,G) = \int_{\mathbb{R}^d} l(x)\, d[L_{\theta_0,h}(G)](x) = \int_{\mathbb{R}^k} \int_0^1 l\left[G(x,u) - \dot{\psi}(\theta_0)h\right] dv(u)\mu(h,dx)$$

across different values of h, where $l : \mathbb{R}^d \to \mathbb{R}_+$ is our loss function of choice. Since this is intimately tied to the statistical properties of the Gaussian location problem with a single observation, this is what we focus on in the subsequent section.

9.3.2 Efficiency in the Gaussian Location Problem

We take a brief detour to focus on the statistical properties of the Gaussian location model μ, where

$$\mu(h,\cdot) \sim \mathcal{N}[h,\Sigma]$$

for each $h \in \mathbb{R}^k$, where Σ is a $k \times k$ positive definite matrix. Let $\Psi \in \mathbb{R}^{d\times k}$. Given a loss function $l : \mathbb{R}^d \to \mathbb{R}_+$, We are interested in searching for a randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ from the Gaussian location problem of interest such that the maximum risk

$$\sup_{h\in\mathbb{R}^k} \int_{\mathbb{R}^k} \int_0^1 l\left[G(x,u) - \Psi h\right] dv(u)\mu(h,dx)$$

is minimized, where v is the uniform distribution on $[0,1]$.

To make any sort of progress, we need to impose some structure on the loss function $l(\cdot)$. Specifically, we will be interested in bowl-shaped loss functions, that is, loss functions such that the sublevel set

$$\{x \in \mathbb{R}^d \mid l(x) \leq c\}$$

is convex and symmetric around the origin (a subset A of $\mathbb{R}^d$ is said to be symmetric around the origin if $A = -A$) for any $c \in \mathbb{R}_+$. If, in addition, the sublevel sets of $l(\cdot)$ are closed, we say that $l(\cdot)$ is subconvex. Many types of loss functions are subconvex, including the absolute loss function $l(x) = |x|$, the quadratic loss function $l(x) = |x|^2$ and the indicator $l(x) = I_{|x|\geq a}$ for any $a > 0$.

A core result in statistics, called Anderson's lemma, shows us that the risk associated with a normal distribution centered at the origin is always lower than the risk associated with any convolution of this normal distribution. First, we must introduce an additional concept. We say that a measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is log-concave if

$$\mu(\lambda A + (1-\lambda)B) \geq \mu(A)^\lambda \mu(B)^{1-\lambda}$$

for any $A, B \in \mathcal{B}(\mathbb{R}^d)$ and $\lambda \in [0,1]$, where $\lambda A + (1-\lambda B$ is defined as the Minkowski sum

$$\lambda A + (1-\lambda)B = \{\lambda x + (1-\lambda)y \mid x \in A, y \in B\}.$$

The name log-concave is made clear if we apply logs on both sides of this inequality (granted that it is well-defined). Similarly, a non-negative function $f : \mathbb{R}^d \to \mathbb{R}_+$ is said to be log-concave if

$$f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$$

for any $x, y \in \mathbb{R}^d$ and $\lambda \in [0,1]$. A well-known result by Prekopa (which we do not prove here) tells us that:

> **(Prekopa's Theorem)**
> A measure μ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is log-concave if and only if it has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ that is itself log-concave.

This theorem shows us that the Lebesgue measure is log-concave (since it has density 1 with respect to itself) and also that any normal distribution is log-concave (since the normal density is a log-concave function).

We can now prove Anderson's lemma for log-concave distributions that are symmetric around the origin:

Lemma 9.13 (Generalized Version of Anderson's Lemma)
Let μ be a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ that is log-concave and symmetric around the origin, that is, $\mu(A) = \mu(-A)$ for any $A \in \mathcal{B}(\mathbb{R}^d)$. Then, for any bowl-shaped loss function $l : \mathbb{R}^d \to \mathbb{R}_+$ and probability measure v on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$,

$$\int_{\mathbb{R}^d} l(x) d\mu(x) \leq \int_{\mathbb{R}^d} l(x) d(\mu * v)(x).$$

Proof) Let X be a random vector with distribution μ. Because μ is symmetric around the origin, $-X$ also has distribution μ. We first prove that

$$\mathbb{P}(X + y \in K) \leq \mathbb{P}(X \in K)$$

for any $y \in \mathbb{R}^k$ and Borel measurable convex set K that is symmetric around the origin. To this end, note that

$$K = \frac{1}{2}(K + y) + \frac{1}{2}(K - y);$$

clearly, $\frac{1}{2}(K+y) + \frac{1}{2}(K-y) \subset K$ by convexity, and for any $x \in K$, $x + y \in K + y$,

$x-y \in K-y$ and $\frac{1}{2}(x+y)+\frac{1}{2}(x-y)=x$, so that $x \in \frac{1}{2}(K+y)+\frac{1}{2}(K-y)$. Therefore,

$$\begin{aligned}
\mathbb{P}(X \in K) &= \mu\left[\frac{1}{2}(K+y)+\frac{1}{2}(K-y)\right] \\
&\geq \mu(K+y)^{\frac{1}{2}}\mu(K-y)^{\frac{1}{2}} && \text{(Log-Concavity of } \mu\text{)} \\
&= \sqrt{\mathbb{P}(-X-y \in K)\,\mathbb{P}(X+y \in K)} && (-X \text{ has distribution } \mu) \\
&= \sqrt{\mathbb{P}(X+y \in -K)\,\mathbb{P}(X+y \in K)} \\
&= \mathbb{P}(X+y \in K). && (K \text{ is symmetric around the origin})
\end{aligned}$$

Now let v be some probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, and Y a random vector with distribution v that is independent of X. By definition, $X+Y$ has distribution $\mu * v$, and the joint distribution of (X,Y) is the product measure $\mu \times v$. Let $l : \mathbb{R}^d \to \mathbb{R}_+$ be some bowl-shaped loss function, so that the sublevel set

$$K_c = \{x \in \mathbb{R}^d \mid l(x) \leq c\}$$

is convex and symmetric around the origin for any $c \in \mathbb{R}_+$; it is also Borel measurable becuase $l(\cdot)$ is a Borel measurable function. By the integral representation of any non-negative random variable, we can see that

$$\begin{aligned}
\mathbb{E}[l(X+Y)] &= \int_0^\infty \mathbb{P}(l(X+Y) > c)\,dc \\
&= \int_0^\infty [1-\mathbb{P}(X+Y \in K_c)]\,dc \\
&= \int_0^\infty \left[1-\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} I_{K_c}(x+y)d\mu(x)dv(y)\right]dc \\
&= \int_0^\infty \left[1-\int_{\mathbb{R}^d} \mathbb{P}(X+y \in K_c)\,dv(y)\right]dc \\
&\geq \int_0^\infty \left[1-\int_{\mathbb{R}^d} \mathbb{P}(X \in K_c)\,dv(y)\right]dc \\
&= \int_0^\infty [1-\mathbb{P}(X \in K_c)]\,dc \\
&= \int_0^\infty \mathbb{P}(l(X) > c)\,dc = \mathbb{E}[l(X)],
\end{aligned}$$

where we used the fact that $\mathbb{P}(X+y \in K_c) \leq \mathbb{P}(X \in K_c)$ for each $y \in \mathbb{R}^k$ because K_c is convex, symmetric around the origin and Borel measurable to justify the inequality.

Since

$$\mathbb{E}[l(X+Y)] = \int_{\mathbb{R}^d} l(x)d(\mu * v)(x), \quad \mathbb{E}[l(X)] = \int_{\mathbb{R}^d} l(x)d\mu(x),$$

the proof is complete.

Q.E.D.

Corollary to Lemma 9.13 (Anderson's Lemma)
Let $\mathcal{N}[O_{d\times 1}, \Sigma]$ be the distribution corresponding to the normal distribution centered at the origin with positive definite variance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Then, for any bowl-shaped loss function $l : \mathbb{R}^d \to \mathbb{R}_+$ and probability measure v on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$,

$$\int_{\mathbb{R}^d} l \; d\mathcal{N}[O_{d\times 1}, \Sigma] \leq \int_{\mathbb{R}^d} l \; d(\mathcal{N}[O_{d\times 1}, \Sigma] * v).$$

Proof) This follows immediately from the previous lemma because the normal distribution is log-concave by Prekopa's theorem, and a normal distribution centered at the origin is also symmetric around the origin.

Q.E.D.

We say that a randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ is equivariant if the distribution

$$\left[(\mu(h,\cdot) \times v) \circ G^{-1}\right] * \delta_{-\Psi h}$$

does not depend on h. Note that this implies that $(\mu(O_{k\times 1}, \cdot) \times v) \circ G^{-1}$ is the distribution of the random vector

$$G(X,U) - \Psi h$$

for any possible value $h \in \mathbb{R}^k$ of the true mean of X, where X is a single observation from the Gaussian location model μ and U is a uniformly distributed random variable independent of X. We call this the distribution of G.

Anderson's lemma plays a crucial role in the following result, which proves two facts: the function $G(x,u) = \Psi x$ minimizes maximum risk under subconvex loss functions, and the distribution of $G(X,U) - \Psi h$ for any equivariant randomized statistic $G(\cdot)$ can be represented as a convolution involving a normal distribution centered at the origin.

Lemma 9.14 Let μ be the Gaussian location model for a single observation with positive definite variance $\Sigma \in \mathbb{R}^{k\times k}$, v the uniform distribution on $[0,1]$, and $\Psi \in \mathbb{R}^{d\times k}$ a non-random

matrix with rank d. Then, the following hold true:

i) For any subconvex loss function $l : \mathbb{R}^d \to \mathbb{R}_+$ and randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ for the Gaussian location model,

$$\sup_{h \in \mathbb{R}^k} \int_{\mathbb{R}^k} \int_0^1 l\left[G(x,u) - \Psi h\right] dv(u) \mu(h, dx) \geq \int_{\mathbb{R}^d} l \, d\mathcal{N}\left[O_{d\times 1},\ \Psi \Sigma \Psi'\right].$$

ii) For any equivariant randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ for the Gaussian location model with distribution L, there exists a probability measure λ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ such that

$$L = \mathcal{N}\left[O_{d\times 1},\ \Psi \Sigma \Psi'\right] * \lambda.$$

Proof) Let $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ be a randomized statistic for the Gaussian location model. The proof of the theorem makes use of a clever Bayesian construction based on the observation that the Gaussian location model can be viewed as a sort of asymptotic posterior of a model with likelihood μ and an uninformative, or flat, prior imposed on the "parameter" H. Intuitively, this tells us that the properties of μ can be recovered by considering a sequence of Bayesian models with likelihood μ and prior distributions that become increasing flat.

To formalize the above intuition, first let λ_n be the distribution corresponding to the normal distribution $\mathcal{N}[O_{k\times 1},\ \Lambda_n]$, where $\Lambda_n = n \cdot I_k$. $\{\lambda_n\}_{n \in N_+}$ is a sequence of probability measures representing the prior distributions that become flatter as n increases to infinity. Now let (X, H_n) be a pair of random vectors with joint distribution $\lambda_n \times \mu$, and choose U as a uniformly distributed random variable that is independent of X and each H_n[3]. By design, the conditional distribution of X given H_n is μ, that is,

$$\mu(\cdot, A) \circ H_n = \mathbb{P}\left(X \in A \mid H_n\right)$$

for any $A \in \mathcal{B}(\mathbb{R}^k)$ and $n \in N_+$, so that

$$X \mid H_n \sim \mathcal{N}\left[H_n,\ \Sigma\right].$$

Since the prior on H_n is given by

$$H_n \sim \mathcal{N}\left[O_{k\times 1},\ \Lambda_n\right],$$

the posterior distribution of H_n given X can be obtained via the usual normal-normal

[3]We can construct the desired random variables $X, U, \{H_n\}_n$ using Kolmogorov's extension theorem.

update as

$$H_n \mid X \sim \mathcal{N}\left[\left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1} \Sigma^{-1} X,\ \left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1}\right].$$

Now define the sequence of random vectors $\{V_n\}_{n \in N_+}$ and $\{W_n\}_{n \in N_+}$ as

$$W_n = G(X,U) - \Psi\left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1} \Sigma^{-1} X$$

$$V_n = -\Psi\left[H_n - \left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1} \Sigma^{-1} X\right].$$

Since

$$V_n \mid X \sim \mathcal{N}\left[O_{d \times 1},\ \Psi\left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1} \Psi'\right],$$

it follows that $V_n \perp X$. Furthermore, because U was chosen so that $U \perp \{(H_n, X)\}_{n \in N_+}$, it follows that $V_n \perp U$ as well, indicating that $V_n \perp W_n$. Furthermore,

$$V_n + W_n = G(X,U) - \Psi H_n$$

for any $n \in N_+$.

Let $l : \mathbb{R}^d \to \mathbb{R}_+$ be a subconvex loss function. Then, defining $g : \mathbb{R}^k \to [0, +\infty]$ as

$$g(h) = \int_{\mathbb{R}^k} \int_0^1 l\left[G(x,u) - \Psi h\right] dv(u) \mu(h, dx)$$

for any $h \in \mathbb{R}^k$, it follows that

$$g(H_n) = \mathbb{E}\left[l\left(G(X,U) - \Psi H_n\right) \mid H_n\right]$$

for any $n \in N_+$. As such, because $\sup_{h \in \mathbb{R}^k} g(h) \geq g(H_n)$, it follows from the monotonicity of integration that

$$\sup_{h \in \mathbb{R}^k} g(h) \geq \mathbb{E}\left[g(H_n)\right] = \mathbb{E}\Big[l\left[G(X,U) - \Psi H_n\right]\Big] = \mathbb{E}\left[l(V_n + W_n)\right].$$

Since

$$V_n \sim \mathcal{N}\left[O_{d \times 1},\ \Psi\left(\Sigma^{-1} + \frac{1}{n} I_k\right)^{-1} \Psi'\right],$$

and subconvex loss functions are by definition bowl-shaped, it follows from Anderson's

lemma that

$$\sup_{h\in\mathbb{R}^k} g(h) \geq \mathbb{E}\left[l(V_n)\right].$$

Finally, since

$$\mathbb{E}\left[e^{ir'V_n}\right] = e^{-\frac{1}{2}r'\Psi\left(\Sigma^{-1}+\frac{1}{n}I_k\right)^{-1}\Psi' r} \to e^{-\frac{1}{2}r'\Psi\Sigma\Psi' r}$$

as $n\to\infty$ for any $r\in\mathbb{R}^d$, the continuity theorem implies that

$$V_n \xrightarrow{d} V \sim \mathcal{N}\left[O_{d\times 1},\ \Psi\Sigma\Psi'\right]$$

as $n\to\infty$. By the integral representation of the expected value of non-negative functions,

$$\begin{aligned}\mathbb{E}\left[l(V_n)\right] &= \int_0^\infty \mathbb{P}\left(l(V_n) > c\right) dc \\ &= \int_0^\infty \left[1 - \mathbb{P}\left(V_n \in K_c\right)\right] dc,\end{aligned}$$

where

$$K_c = \left\{x\in\mathbb{R}^d \mid l(x)\leq c\right\}$$

is the sublevel set of $l(\cdot)$ associated with $c\geq 0$. By the subconvexity of $l(\cdot)$, each K_c is closed, so that, by the Portmanteau theorem,

$$\limsup_{n\to\infty} \mathbb{P}\left(V_n\in K_c\right) \leq \mathbb{P}\left(V\in K_c\right)$$

for each $c\geq 0$. As such, Fatou's lemma implies that

$$\begin{aligned}\liminf_{n\to\infty}\mathbb{E}\left[l(V_n)\right] &= \liminf_{n\to\infty}\int_0^\infty \left[1-\mathbb{P}\left(V_n\in K_c\right)\right] dc \\ &\geq \int_0^\infty \left[1 - \limsup_{n\to\infty}\mathbb{P}\left(V_n\in K_c\right)\right] dc \\ &\geq \int_0^\infty \left[1-\mathbb{P}\left(V\in K_c\right)\right] dc \\ &= \int_0^\infty \mathbb{P}\left(l(V)>c\right) dc = \mathbb{E}\left[l(V)\right].\end{aligned}$$

In other words,

$$\sup_{h\in\mathbb{R}^k} g(h) \geq \liminf_{n\to\infty}\mathbb{E}\left[l(V_n)\right] \geq \mathbb{E}\left[l(V)\right] = \int_{\mathbb{R}^d} l\ d\mathcal{N}\left[O_{d\times 1},\ \Psi\Sigma\Psi'\right].$$

This proves the first claim.

To prove the second claim, let $G(\cdot)$ be an equivariant randomized statistic, so that there exists some probability measure L such that

$$L = (\mu(h,\cdot) \times v) \circ G^{-1} * \delta_{-\Psi h}$$

for any $h \in \mathbb{R}^k$. It follows that

$$G(X,U) - \Psi H_n \mid H_n \sim L,$$

so that

$$\varphi(r) := \int_{\mathbb{R}^d} e^{ir'x} dL(x) = \mathbb{E}\left[e^{ir'(G(X,U)-\Psi H_n)} \mid H_n\right]$$

and therefore

$$\varphi(r) = \mathbb{E}\left[e^{ir'(G(X,U)-\Psi H_n)}\right]$$

for any $r \in \mathbb{R}^d$ by the law of iterated expectations. By implication, the distribution of $G(X,U) - \Psi H_n$ is equal to L for each n.

Since $G(X,U) - \Psi H_n = V_n + W_n$, and $V_n \perp W_n$, we can see that

$$\varphi(r) = \mathbb{E}\left[e^{ir'V_n + ir'W_n}\right] = \mathbb{E}\left[e^{ir'V_n}\right]\mathbb{E}\left[e^{ir'W_n}\right]$$

for any $n \in N_+$. We saw above that

$$\mathbb{E}\left[e^{ir'V_n}\right] \to e^{-\frac{1}{2}r'\Psi\Sigma\Psi'r}$$

as $n \to \infty$, so

$$\lim_{n\to\infty} \mathbb{E}\left[e^{ir'W_n}\right] = \int_{\mathbb{R}^d} e^{ir'x + \frac{1}{2}r'\Psi\Sigma\Psi'r} dL(x) = e^{\frac{1}{2}r'\Psi\Sigma\Psi'r}\varphi(r).$$

Defining $\phi : \mathbb{R}^k \to \mathbb{C}$ as

$$\phi(r) = e^{\frac{1}{2}r'\Psi\Sigma\Psi'r}\varphi(r)$$

for any $r \in \mathbb{R}^d$, we have shown that the characteristic function of W_n converges pointwise to the continuous function $\phi(\cdot)$ (since $\varphi(\cdot)$, being a characteristic function, is uniformly continuous on $\mathbb{R}^d$). Therefore, by the continuity theorem, it follows that there exists some random vector W such that $W_n \xrightarrow{d} W$. The independence of W_n and V_n now allows

us to conclude that

$$\begin{pmatrix} V_n \\ W_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} V \\ W \end{pmatrix}$$

jointly, so that, by the continuous mapping theorem,

$$V_n + W_n \xrightarrow{d} V + W.$$

Since $V_n + W_n = G(X,U) - \Psi H_n$ has distribution L, the uniqueness of weak limits tells us that $V + W \sim L$. We already saw that $V \sim \mathcal{N}[O_{d\times 1},\ \Psi\Sigma\Psi']$, so the proof is complete.

Q.E.D.

9.3.3 The Convolution Theorem

Based on the results derived above about the Gaussian location problem with a single observation, we can see that the randomized statistic $G^* : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ defined as $G^*(x,u) = \Psi x$ for any $(x,u) \in \mathbb{R}^k \times [0,1]$ is the best estimator of Ψh in the minimax sense and, among equivariant statistics, in any sense, minimax or otherwise. As such, we can conclude that a sequence of statistics in a model satisfying the LAN property whose local limiting distribution around θ_0 is given as $G^*(X,U) - \dot{\psi}(\theta_0)h$ for any $h \in \mathbb{R}^k$ has the optimal limiting distribution among the class of smooth statistics in terms of the minimax criterion.

Theorem 9.15 (Minimax Optimality of the Normal Distribution)
Let $\{(E_n, \mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces, (Θ, ϑ) the k-dimensional parameter space and $\{\mu_n\}_{n\in N_+}$ a sequence of statistical models. Let $\psi : \Theta \to \mathbb{R}^d$ be a function that is differentiable at $\theta_0 \in \Theta^o$ of the parameter space, and denote the derivative by $\dot{\psi}(\theta_0)$. Suppose that the following hold true:

i) **(Local Asymptotic Normality)**
$\{\mu_n\}_{n\in N_+}$ is locally asymptotically normal at θ_0, that is, there exists some positive definite $k \times k$ matrix I_0, a sequence $\{\Delta_n\}_{n\in N_+}$ where each $\Delta_n : E_n \to \mathbb{R}^k$ is a measurable function, and $\Delta_0 \sim \mathcal{N}[O_{k\times 1},\ I_0]$ such that $\Delta_n \xRightarrow{0} \Delta_0$ and

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} - \exp\left(h'\Delta_n - \frac{1}{2}h'I_0h\right) \xRightarrow{0} 0$$

for any sequence $\{h_n\}_{n\in N_+}$ in $\mathbb{R}^k$ converging to $h \in \mathbb{R}^k$.

ii) **(Nonsingularity of Gradient)**
The $d \times k$ matrix $\dot{\psi}(\theta_0)$ has full rank d.

Then, defining the randomized statistic $G^* : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ from the Gaussian location model with variance I_0^{-1} as $G^*(x,u) = \dot{\psi}(\theta_0)x$ for any $(x,u) \in \mathbb{R}^k \times [0,1]$, $G^*(\cdot)$ is an equivariant statistic with distribution

$$L \sim \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right].$$

Furthermore, the sequence of smooth statistics $\{\mathcal{T}_n^*\}_{n\in N_+}$ such that

$$\sqrt{n}\left[\mathcal{T}_n^* - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} G^*(X,U) - \dot{\psi}(\theta_0)h \sim \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

for any $h \in \mathbb{R}^k$ is optimal in the minimax sense among smooth statistics at θ_0. That is, for any sequence $\{\mathcal{T}_n\}_{n\in N_+}$ of statistics that are smooth at θ_0 with limiting distribution

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} \mathcal{T}_{\theta_0,h} \sim L_{\theta_0,h}$$

for any $h \in \mathbb{R}^k$, and subconvex loss function $l : \mathbb{R}^d \to \mathbb{R}_+$,

$$\sup_{h\in\mathbb{R}^k} \int_{\mathbb{R}^d} l(x)\ dL_{\theta_0,h}(x) \geq \int_{\mathbb{R}^d} l(x)\ dL(x).$$

Proof) Let μ be the Gaussian location model with variance I_0^{-1}, and v the uniform distribution on $[0,1]$. Define the randomized statistic $G^*(\cdot)$ as above. Then,

$$G^*(X,U) - \dot{\psi}(\theta_0)h = \dot{\psi}(\theta_0)(X-h) \sim \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

when $h \in \mathbb{R}^k$ is the true mean, that is,

$$\left[\mu(h,\cdot) \times v) \circ (G^*)^{-1}\right] * \delta_{-\dot{\psi}(\theta_0)h} \sim \mathcal{N}\left[O_{k\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

for any $h \in \mathbb{R}^k$. In other words, $G^*(\cdot)$ is equivariant with distribution $\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$.

Now let the sequence $\{\mathcal{T}_n^*\}_{n\in N_+}$ be the sequence of smooth statistics with limiting distribution involving $G^*(\cdot)$ defined in the statement of the theorem, and choose any sequence $\{\mathcal{T}_n\}_{n\in N_+}$ of statistics that are smooth at θ_0. From theorem 9.12, the local asymptotic normality of $\{\mu_n\}_{n\in N_+}$ at θ_0 and the differentiability of $\psi(\cdot)$ on the interior of the parameter space imply that there exists some randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ from the Gaussian location model μ such that

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} G(X,U) - \dot{\psi}(\theta_0)h \sim L_{\theta_0,h}$$

for any $h \in \mathbb{R}^k$. For any subconvex loss function $l : \mathbb{R}^d \to \mathbb{R}_+$, claim i) from lemma 9.14

tells us that

$$\sup_{h\in\mathbb{R}^k}\int_{\mathbb{R}^d} l(x)dL_{\theta_0,h}(x) = \sup_{h\in\mathbb{R}^k}\int_{\mathbb{R}^k}\int_0^1 l\left[G(x,u)-\dot{\psi}(\theta_0)h\right]dv(u)\mu(h,dx) \geq \int_{\mathbb{R}^d} l(x)dL(x).$$

This completes the proof.

Q.E.D.

Since the distribution of the equivariant statistic $G^*(\cdot)$ has distribution $\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$, the preceding theorem can be used as a heuristic justification for why this distribution is the best limiting distribution. Note also that the variance $\dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'$ of this limiting distribution is precisely the Cramer-Rao lower bound (since I_0 is the limiting Fisher information matrix for generally dependent data and the one-observation information matrix for iid data), suggesting that the CRLB is not only a lower bound for the variance of an estimator in finite samples but that it also serves the same role for the variance of smooth statistics in the limit.

The last claim concerning the CRLB does not hold generally for smooth statistics, since we showed only that the maximum asymptotic risk associated with a smooth statistic is bounded below by the asymptotic risk incurred under the optimal statistic $G^*(\cdot)$. As such, it may be the case that there exist values of $h\in\mathbb{R}^k$ such that

$$\int_{\mathbb{R}^d} l(x)dL_{\theta_0,h}(x) < \int_{\mathbb{R}^d} l(x)dL(x),$$

so that, in the special case that $l(x)=|x|^2$, there may exist a local alternative around θ_0 that achieves an asymptotic variance that is smaller than the CRLB.

The following theorem shows us that, if we restrict our attention to regular statistics, then the CRLB serves also as the lower bound for the asymptotic variance. Called the convolution theorem, it actually goes further and reveals that the asymptotic risk from the normal distribution with asymptotic variance equal to the CRLB is never exceeds the asymptotic risk of any regular statistic for any values of $h\in\mathbb{R}^k$.

Theorem 9.16 (The Convolution Theorem)
Let $\{(E_n,\mathcal{E}_n)\}_{n\in N_+}$ be a sequence of measurable spaces, (Θ,ϑ) the k-dimensional parameter space and $\{\mu_n\}_{n\in N_+}$ a sequence of statistical models. Let $\psi:\Theta\to\mathbb{R}^d$ be a function that is differentiable at $\theta_0\in\Theta^o$, and denote the derivative by $\dot{\psi}(\theta_0)$. Suppose that the following hold true:

i) **(Local Asymptotic Normality)**
$\{\mu_n\}_{n\in N_+}$ is locally asymptotically normal at θ_0, that is, there exists some positive definite $k\times k$ matrix I_0, a sequence $\{\Delta_n\}_{n\in N_+}$ where each $\Delta_n: E_n\to\mathbb{R}^k$ is a measurable function,

and $\Delta_0 \sim \mathcal{N}[O_{k\times 1},\, I_0]$ such that $\Delta_n \overset{0}{\Rightarrow} \Delta_0$ and

$$\frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} - \exp\left(h'\Delta_n - \frac{1}{2}h'I_0h\right) \overset{0}{\Rightarrow} 0$$

for any sequence $\{h_n\}_{n\in N_+}$ in $\mathbb{R}^k$ converging to $h \in \mathbb{R}^k$.

ii) **(Nonsingularity of Gradient)**
The $d \times k$ matrix $\dot{\psi}(\theta_0)$ has full rank d.

Let $G^* : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ be the randomized statistic from the Gaussian location model with variance I_0^{-1} and $\mathcal{T}_n^*$ be the regular at θ_0 statistic with limiting distribution given as

$$L^* \sim \mathcal{N}\left[O_{d\times 1},\, \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

defined in the previous theorem. Then, for any sequence $\{\mathcal{T}_n\}_{n\in N_+}$ of statistics that are regular at θ_0 with limiting distribution

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} \mathcal{T} \sim L$$

for any $h \in \mathbb{R}^k$, the following hold true:

i) **(Convolution Representation)**
There exists a distribution λ on $\mathbb{R}^d$ such that

$$L = L^* * \lambda = \mathcal{N}\left[O_{d\times 1},\, \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right] * \lambda.$$

ii) **(Uniform Dominance of Asymptotic Risk)**
For any bowl-shaped loss function $l : \mathbb{R}^d \to \mathbb{R}_+$,

$$\int_{\mathbb{R}^d} l(x)\, dL(x) \geq \int_{\mathbb{R}^d} l(x)\, dL^*(x)$$

for any $h \in \mathbb{R}^k$.

iii) **(CRLB as Lower Bound on Asymptotic Variance)**
If $\Sigma = \int_{\mathbb{R}^d} xx' dL(x)$ exists in $\mathbb{R}^{d\times d}$,

$$\Sigma - \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'$$

is positive semidefinite.

Proof) We showed in theorem 9.15 that $G^*(\cdot)$ is an equivariant statistic with distribution

$$L^* \sim \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)'\right],$$

so that the statistic $\mathcal{T}_n^*$ such that

$$\sqrt{n}\left[\mathcal{T}_n^* - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} G^*(X,U) - \dot{\psi}(\theta_0)h \sim L^*$$

for any $h \in \mathbb{R}^k$, where X is a single observation from the Gaussian location model μ with variance I_0^{-1} and U an independent uniformly distributed random variable, is regular at θ_0.

Let $\mathcal{T}_n$ be a statistic from the statistical model $\{\mu_n\}_{n\in N_+}$ that is regular at θ_0, so that there exists some d-dimensional random vector $\mathcal{T}$ with distribution L such that

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} \mathcal{T}$$

for any $h \in \mathbb{R}^k$. By theorem 9.12, the local asymptotic normality of $\{\mu_n\}_{n\in N_+}$ and the continuous differentiability of $\psi(\cdot)$ on the interior of the parameter space imply that there exists some randomized statistic $G : \mathbb{R}^k \times [0,1] \to \mathbb{R}^d$ from the Gaussian location model μ such that

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] \overset{h}{\Rightarrow} G(X,U) - \dot{\psi}(\theta_0)h$$

for any $h \in \mathbb{R}^k$. Because the distribution of $G(X,U) - \dot{\psi}(\theta_0)h$ is the same at L for any value of $h \in \mathbb{R}^k$, it follows that $G(\cdot)$ is an equivariant statistic for the Gaussian location model μ, and by the second claim of lemma 9.14, there exists some probability measure λ on $\mathbb{R}^d$ such that

$$L = \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)'\right] * \lambda = L^* * \lambda.$$

Now let $l : \mathbb{R}^d \to \mathbb{R}_+$ be any bowl-shaped loss function. By Anderson's lemma, the convolution representation $L = L^* * \lambda$ and the fact that L^* is a normal distribution centered at the origin with nonsingular variance, we can now conclude that

$$\int_{\mathbb{R}^d} l(x) dL^*(x) \leq \int_{\mathbb{R}^d} l(x) dL(x).$$

Finally, suppose that the asymptotic variance

$$\Sigma = \int_{\mathbb{R}^d} xx' dL(x)$$

of the statistic $\mathcal{T}_n$ exists in $\mathbb{R}^{d\times d}$. Then, for any $\alpha \in \mathbb{R}^d$, define $l(x) = \alpha' x x' \alpha$ for any $x \in \mathbb{R}^d$. $l(\cdot)$ is a non-negative valued function that is continuous, and since $l(\cdot)$ is twice continuously differentiable with Hessian $2\alpha\alpha'$, a positive semidefinite matrix, $l(\cdot)$ is a convex function. It follows that, for any $c \in \mathbb{R}_+$, the sublevel set

$$\{x \in \mathbb{R}^d \mid l(x) \leq c\}$$

is convex. Finally, the sublevel set is symmetric because $l(-x) = l(x)$ for any $x \in \mathbb{R}^d$. It follows that $l(\cdot)$ is a subconvex and thus bowl-shaped loss function, which indicates that

$$\alpha'\Sigma\alpha = \int_{\mathbb{R}^d} l(x) dL(x) \geq \int_{\mathbb{R}^d} l(x) dL^*(x) = \alpha' \left[\dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)'\right] \alpha.$$

This holds for any $\alpha \in \mathbb{R}^d$, so it follows by definition that

$$\Sigma - \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)'$$

is positive semidefinite.

Q.E.D.

9.3.4 DQM and Asymptotically Efficient Estimators

So far, we have considered asymptotic efficiency for general models $\{\mu_n\}_{n\in N_+}$ that satisfy the LAN property. In light of the DQM property for models with iid data that imply the LAN property, for models with iid data it is possible to use DQM as a primitive condition under which the results above hold. In fact, under additional parametric assumptions we can obtain an expression for asymptotically efficient estimators, that is, regular estimators that achieve the distribution $\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)'\right]$ at the limit, in terms of the efficient influence function, which is defined using the score function and the Fisher information matrix.

We begin by formulating the convolution theorem under DQM:

Theorem 9.17 (Convolution Theorem under DQM)
Let $(E, \mathcal{E})$ be a measurable space, (Θ, ϑ) the k-dimensional parameter space, and μ a transition probability kernel from (Θ, ϑ) into $(E, \mathcal{E})$ that determines the sequence of statistical models $\{\mu_n\}_{n\in N_+}$ under iid. Let $\psi : \Theta \to \mathbb{R}^d$ be a function that is differentiable at $\theta_0 \in \Theta^o$, and denote the derivative by $\dot{\psi}(\theta_0)$. Suppose that the DQM condition holds at θ_0 with score function $g_0 : E \to \mathbb{R}^k$

and dominating measure λ:

$$\lim_{|h|\to 0}\frac{1}{|h|^2}\int_E\left(\sqrt{d\mu(\theta_0+h,\cdot)}-\sqrt{d\mu(\theta_0,\cdot)}-\frac{1}{2}h'g_0\cdot\sqrt{d\mu(\theta_0,\cdot)}\right)^2 d\lambda=0.$$

Assume in addition that $\int_E g_0 g_0' d\mu = I_0$ is positive definite, and that the $d\times k$ matrix $\dot{\psi}(\theta_0)$ has full rank d. Then, for any sequence $\{\mathcal{T}_n\}_{n\in N_+}$ of statistics that are regular at θ_0 with limiting distribution

$$\sqrt{n}\left[\mathcal{T}_n-\psi\left(\theta_0+\frac{h}{\sqrt{n}}\right)\right]\overset{h}{\Rightarrow}\mathcal{T}\sim L$$

for any $h\in\mathbb{R}^k$, the following hold true:

i) **(Convolution Representation)**
There exists a distribution M on $\mathbb{R}^d$ such that

$$L=\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]*M.$$

ii) **(Uniform Dominance of Asymptotic Risk)**
For any bowl-shaped loss function $l:\mathbb{R}^d\to\mathbb{R}_+$,

$$\int_{\mathbb{R}^d} l\ dL\geq\int_{\mathbb{R}^d} l\ d\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

for any $h\in\mathbb{R}^k$.

iii) **(CRLB as Lower Bound on Asymptotic Variance)**
If $\Sigma=\int_{\mathbb{R}^d}xx'dL(x)$ exists in $\mathbb{R}^{d\times d}$,

$$\Sigma-\dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'$$

is positive semidefinite.

Furthermore, suppose the sequence $\{\mathcal{T}_n^*\}_{n\in N_+}$ of statistics satisfies

$$\sqrt{n}[\mathcal{T}_n^*(X_1,\cdots,X_n)-\psi(\theta_0)]=\frac{1}{\sqrt{n}}\sum_{i=1}^n\dot{\psi}(\theta_0)I_0^{-1}g_0(X_i)+o_p(1),$$

where $\{X_i\}_{i\in N_+}$ is an iid sequence of random variables drawn from the distribution $\mu(\theta_0,\cdot)$. Then, $\mathcal{T}_n^*$ is a regular statistic at θ_0 such that

$$\sqrt{n}\left[\mathcal{T}_n^*-\psi\left(\theta_0+\frac{h}{\sqrt{n}}\right)\right]\overset{h}{\Rightarrow}\mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right]$$

for any $h \in \mathbb{R}^k$.

Proof) Recall from theorem 9.11 that, if the DQM condition holds at θ_0 for a score function g_0, then the LAN property holds:

$$\log \frac{d\mu_n(\theta_0 + h_n/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} \overset{0}{\Rightarrow} h'\Delta_0 - \frac{1}{2}h'I_0h$$

for any $h_n \to h$, where $\Delta_0 \sim \mathcal{N}[O_{k\times 1},\, I_0]$. As such, statements i) to iii) follow immediately from the general convolution theorem.

It remains to show that statement about the asymptotic efficiency of the asymptotically linear statistic $\mathcal{T}_n^*$ holds. Letting $\{X_i\}_{i\in N_+}$ be a sequence drawn iid from the distribution $\mu(\theta_0, \cdot)$, and defining $X^n = (X_1, \cdots, X_n)$ for any $n \in N_+$, by assumption

$$\sqrt{n}\,[\mathcal{T}_n^*(X^n) - \psi(\theta_0)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\psi}(\theta_0) I_0^{-1} g_0(X_i) + o_p(1)$$

for any $n \in N_+$, where

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \overset{d}{\to} \Delta_0$$

by the Lindeberg-Levy CLT. We also know from theorem 9.11 that

$$\frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) = \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) + o_p(1)$$

for any $n \in N_+$ and $h \in \mathbb{R}^k$. It follows from Slutsky's theorem that

$$\begin{aligned}
&\left(\sqrt{n}\,[\mathcal{T}_n^*(X^n) - \psi(\theta_0)], \quad \frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)}(X^n) \right) \\
&\quad = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) + o_p(1), \quad \exp\left(h' \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_0(X_i) \right) - \frac{1}{2} h' I_0 h \right) + o_p(1) \right) \\
&\qquad \overset{d}{\to} \left(\dot{\psi}(\theta_0) I_0^{-1} \Delta_0, \quad e^{h'\Delta_0 - \frac{1}{2}h'I_0h} \right),
\end{aligned}$$

so that, by the continuous mapping theorem, we have

$$\left(\sqrt{n}\,[\mathcal{T}_n^* - \psi(\theta_0)], \quad \log \frac{d\mu_n(\theta_0 + h/\sqrt{n}, \cdot)}{d\mu_n(\theta_0, \cdot)} \right) \overset{0}{\Rightarrow} \mathcal{N}\left[\begin{pmatrix} O_{d\times 1} \\ -\frac{1}{2}h'I_0h \end{pmatrix}, \begin{pmatrix} \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)' & \dot{\psi}(\theta_0)h \\ h'\dot{\psi}(\theta_0)' & h'I_0h \end{pmatrix} \right],$$

for any $h \in \mathbb{R}^k$. By Le Cam's third lemma,

$$\sqrt{n}\,[\mathcal{T}_n^* - \psi(\theta_0)] \overset{h}{\Rightarrow} \mathcal{N}\left[\dot{\psi}(\theta_0)h,\; \dot{\psi}(\theta_0) I_0^{-1} \dot{\psi}(\theta_0)' \right]$$

for any $h \in \mathbb{R}^k$. We noted in the proof of theorem 9.12 that

$$\lim_{n\to\infty} \sqrt{n}\left[\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right] = \dot{\psi}(\theta_0)h$$

for any $h \in \mathbb{R}^k$, so that, by Slutsky's theorem,

$$\sqrt{n}\left[\mathcal{T}_n^* - \psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right] = \sqrt{n}\left[\mathcal{T}_n^* - \psi(\theta_0)\right] - \sqrt{n}\left[\psi\left(\theta_0 + \frac{h}{\sqrt{n}}\right) - \psi(\theta_0)\right]$$

$$\stackrel{h}{\Rightarrow} \mathcal{N}\left[\dot{\psi}(\theta_0)h,\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right] * \delta_{-\dot{\psi}(\theta_0)h} \sim \mathcal{N}\left[O_{d\times 1},\ \dot{\psi}(\theta_0)I_0^{-1}\dot{\psi}(\theta_0)'\right].$$

for any $h \in \mathbb{R}^k$. This proves that $\mathcal{T}_n^*$ is regular at θ_0 with the optimal limiting distribution.

Q.E.D.

The above theorem shows us that, in a setting with iid data, if $\sqrt{n}\left[\mathcal{T}_n^* - \psi(\theta_0)\right]$ admits an asymptotic linear expansion under the null that θ_0 is the true parameter, then $\mathcal{T}_n^*$ is a regular statistic with the optimal limiting distribution, that is, the regular statistic that attains the CRLB. The function

$$(x_1, \cdots, x_n) \mapsto \dot{\psi}(\theta_0)I_0^{-1}g_0(x_i)$$

used to construct the linear part of $\mathcal{T}_n^*$ is called the efficient influence function.

In particular, we showed that, under the appropriate regularity conditions, the MLE of θ is asymptotically linear under the null with efficient influence function $I_0^{-1}g_0(\cdot)$, where the score $g_0(\cdot)$ is here given as the gradient of the log likelihood for a single observation; in this sense, the MLE of θ is a regular statistic that attains the optimal limiting distribution $\mathcal{N}\left[O_{k\times 1}, I_0^{-1}\right]$.

9.4 Asymptotic Efficiency in a Non-parametric Setting

So far, we defined asymptotic efficiency for regular statistics in a model with a finite-dimensional parameter space. These types of models are called parametric models. It is also possible to extend the results above to a model with an infinite-dimensional parameter space, as suggested in the beginning of the chapter. However, the extension is not so straightforward. For beginners, all the results we have obtained so far have been local in the sense that we study a shrinking neighborhood of a chosen parameter θ_0 in the parameter space. If the parameter space is infinite-dimensional, so that, for example, it is a subset of the space of probability measures on some euclidean space, how exactly would we define a shrinking neighborhood of a chosen parameter? Furthermore, the results for asymptotic efficiency obtained above were predicated on the assumption of the differentiability of the function $\psi(\cdot)$. How do we define the derivative of a function $\psi(\cdot)$ that maps probability measures to finite dimensional spaces?

The solution to both these complications is to consider finite-dimensional paths on the parameter space that pass through the chosen parameter. By considering these finite-dimensional paths, we reduce an infinite-dimensional problem into a collection of finite-dimensional problems, for which the notions of locality and differentiability are well-defined. The lower bound for the asymptotic variance of regular statistics, in this context, can just be taken to be the supremum of the CRLB of these finite-dimensional paths. In this sense, we can reduce the problem of defining asymptotic efficiency for a non-parametric problem into one of finding the finite-dimensional subproblem with the largest, or worst-case, asymptotic variance.

Below, we motivate non-parametric asymptotic efficiency via approximation by one-dimensional paths, and show that we can obtain an analogue to the CRLB. This quantity is then shown to admit an interpretation as the lower bound for the asymptotic variance of regular statistics via an analogue of the convolution theorem proved in the previous section. Finally, we apply the machinery developed here to semi-parametric models and obtain an asymptotically linear expansion for semi-parametrically efficient estimators. We work with iid data throughout this chapter, and touch on more general forms of dependency and heterogeneity at the end.

9.4.1 Differentiability and the Efficient Influence Function

Consider the following theoretical framework. Let $(E,\mathcal{E})$ be a measurable space and P a probability measure on $(E,\mathcal{E})$. Suppose that P belongs to the class $\mathcal{P} \subset \mathcal{M}_p(\mathcal{E})$ of probability measures on $(E,\mathcal{E})$, where $\mathcal{M}_p(\mathcal{E})$ is the set of all probability measures on $(E,\mathcal{E})$. $\mathcal{P}$ here is our (infinite-dimensional) parameter space. Recall that, in a parametric setting, a statistical model represented by a transition probability kernel collects probability measures on $(E,\mathcal{E})$ corresponding to each parameter value. Analogously, in a general non-parametric setting we let the class $\mathcal{P}$ itself be the statistical model, since it collects all probability measures under consideration that we may or may not be able to index using a finite-dimensional parameter.

Given a function $\psi : \mathcal{P} \to \mathbb{R}^d$, we want to estimate the value $\psi(P)$. Recall that our discussion of asymptotic efficiency in the parametric case held under two assumptions: local asymptotic

normality and the continuous differentiability of the mapping $\psi(\cdot)$. The LAN property, in turn, held if the DQM property was satisfied, from which we can obtain the Fisher information matrix associated with the parametric model. Thus, we require a generalization of DQM and differentiability for the case where the parameter space $\mathcal{P}$ is infinite-dimensional.

Differentiability in Quadratic Mean for One-Dimensional Submodels

We first discuss DQM and the generalization of Fisher information. Let $\mathcal{P}_0$ be a parametric submodel of $\mathcal{P}$, that is, $\mathcal{P}_0$ is a collection of probability measures on $(E, \mathcal{E})$ such that $\mathcal{P}_0 = \{\mu_0(\theta, \cdot) \mid \theta \in \Theta\}$ for some finite-dimensional parameter space Θ and transition probability kernel μ from Θ into E, that satisfies $P \in \mathcal{P}_0 \subset \mathcal{P}$. We might describe the probability kernel μ, or equivalently the collection $\mathcal{P}_0$, as a parametric model that passes through the true distribution P. Because of the parametric nature of $\mathcal{P}_0$, if it satisfies the DQM property then its Fisher information matrix for estimating $\psi(P)$ is well-defined.

Since the parametric submodel μ passing through P contains more information about $\psi(P)$ than the general non-parametric model, it follows that the information about $\psi(P)$ in the general model cannot exceed that of the submodel μ. Therefore, a natural way to define the information contained in the general model about $\psi(P)$ is as the infimum of the information about $\psi(P)$ contained in each parametric submodel; in terms of the information matrix, the information matrix on $\psi(P)$ should be defined as the supremum over the information matrices of each parametric submodel.

It turns out that, in many cases including any parametric model or completely non-parametric model, it suffices to consider one-dimensional submodels of $\mathcal{P}$. Formally, we consider one-dimensional parametric submodels μ of $\mathcal{P}$, where μ is a transition probability kernel from $[0, +\infty)$ into E such that $\mu(0, \cdot) = P$. We say that the one-dimensional submodel μ satisfies the DQM property if:

> There exists a measurable function $g : E \to \mathbb{R}$ and measure λ on $(E, \mathcal{E})$ such that $\int_E |g|^2 dP < +\infty$, $\mu(t, \cdot) << \lambda$ for each $t \geq 0$, and
>
> $$\lim_{t \to 0} \int_E \left(\frac{\sqrt{d\mu(t, \cdot)} - \sqrt{dP}}{t} - \frac{1}{2} g \cdot \sqrt{dP} \right)^2 d\lambda = 0.$$

Note that this is equivalent to requiring that the DQM property holds at the point $0 \in [0, +\infty)$. The collection of all score functions $g(\cdot)$ corresponding to one-dimensional submodels of $\mathcal{P}$ satisfying the DQM property is denoted by $\dot{\mathcal{P}}_P$, and it is called the tangent set of the model $\mathcal{P}$ at P. The tangent set $\dot{\mathcal{P}}_P$ is a subset of the L^2-space $L^2(\mathcal{E}, P)$ by definition (each score function $g(\cdot)$ is square integrable with respect to P). The following lemma shows us that the tangent set is a cone, and that any submodel associated with a score function in the tangent set satisfies the LAN property:

Lemma 9.18 (Properties of the Tangent Set)
Let $(E,\mathcal{E})$ be a measurable space, and P a probability measure in some statistical model $\mathcal{P} \subset \mathcal{M}_+(\mathcal{E})$. Let $\dot{\mathcal{P}}_P \subset L^2(\mathcal{E},P)$ be the tangent set of the model $\mathcal{P}$ at P. Then, the following hold true:

i) **(Tangent Set is a Cone)**
$\dot{\mathcal{P}}_P$ is a cone, that is, it is closed under non-negative scalar multiplication.

ii) **(LAN Property of Submodels)**
For any one-dimensional submodel μ of $\mathcal{P}$ that satisfies the DQM property with score function $g \in \dot{\mathcal{P}}_P$, we have $\int_E g dP = 0$ and

$$\prod_{i=1}^{n} \frac{d\mu(1/\sqrt{n},\cdot)}{dP}(X_i) = \exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i) - \frac{1}{2}I(g)\right) + o_p(1),$$

where $\{X_i\}_{i\in N_+}$ is an iid sequence of random variables with distribution P and $I(g) = \int_E g^2 dP$. By implication,

$$\log \frac{d\mu_n(1/\sqrt{n},\cdot)}{dP^n} \overset{P^n}{\Rightarrow} \Delta(g) - \frac{1}{2}I(g),$$

where $\Delta(g) \sim \mathcal{N}[0, I(g)]$ and μ_n is the transition probability kernel from $[0,+\infty)$ into E^n such that $\mu_n(t,\cdot) = \underbrace{\mu(t,\cdot) \times \cdots \times \mu(t,\cdot)}_{n}$ for any $t \geq 0$.

Proof) Let $g \in \dot{\mathcal{P}}_P$ and $a > 0$. Then, there exists some one-dimensional submodel μ of $\mathcal{P}$ such that $\mu(0,\cdot) = P$ and

$$\lim_{t\to 0} \int_E \left(\frac{\sqrt{d\mu(t,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} g \cdot \sqrt{dP}\right)^2 d\lambda = 0.$$

for some measure λ on $(E,\mathcal{E})$ such that $\mu(t,\cdot) << \lambda$ for any $t \geq 0$. Defining $\tilde{\mu}$ as

$$\tilde{\mu}(t,\cdot) = \mu(at,\cdot)$$

for any $t \geq 0$, it follows that $\tilde{\mu}$ is also a transition probability kernel from $[0,+\infty)$ into E such that $\tilde{\mu}(0,\cdot) = \mu(0,\cdot) = P$, so that it constitutes a well-defined one-dimensional submodel of $\mathcal{P}$. Furthermore, $\tilde{\mu}(t,\cdot) << \lambda$ for any $t \geq 0$ and

$$\int_E \left(\frac{\sqrt{d\tilde{\mu}(t,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} ag \cdot \sqrt{dP}\right)^2 d\lambda = a^2 \int_E \left(\frac{\sqrt{d\mu(at,\cdot)} - \sqrt{dP}}{at} - \frac{1}{2} g \cdot \sqrt{dP}\right)^2 d\lambda \to 0$$

as $t \to 0$, so that $ag \in \dot{\mathcal{P}}_P$ as well. Finally, the zero function is contained in $\dot{\mathcal{P}}_P$ because it corresponds to the one-dimensional submodel μ of $\mathcal{P}$ defined as $\mu(t,\cdot) = P$ for any $t \geq 0$ with dominating measure P.

Meanwhile, if μ is a one-dimensional submodel of $\mathcal{P}$ satisfying the DQM property with score function $g(\cdot)$, then putting $\theta_0 = 0$ and $h_n = h = 1$ for each $n \in N_+$ in the proof of theorem 9.11 shows us that $\int_E g dP = 0$ and

$$\prod_{i=1}^{n} \frac{d\mu(1/\sqrt{n}, \cdot)}{dP}(X_i) = \exp\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \frac{1}{2} I(g) \right) + o_p(1)$$

for any sequence $\{X_i\}_{i \in N_+}$ of iid random variables drawn from the distribution P and Fisher information $I(g) = \int_E g^2 dP$. The LAN property follows immediately.

Q.E.D.

The collection $\dot{\mathcal{P}}_P$ of score functions is sometimes called the maximal tangent set, since it collects every score function associated with a one-dimensional submodel satisfying the DQM property. Often we will be interested in subsets of the maximal tangent set, or collections of score functions that satisfy additional properties other than being the score function associated with a one-dimensional submodel. For instance, in many cases we want to focus on the collection of score functions $g(\cdot)$ for which the function $\psi(\cdot)$ is differentiable in the direction $g(\cdot)$ (which will be defined formally below), or score functions that satisfy some sort of orthogonality condition (which arises when we deal with models with moment restrictions). As such, without loss of generality below we let tangent set $\dot{\mathcal{P}}_P$ denote a subset of the maximal tangent set; when $\dot{\mathcal{P}}_P$ is maximal, we will state as such.

Differentiability of Functionals

For estimating the quantity $\psi(P)$, we want to restrict our attention to submodels μ such that the mapping $t \mapsto \psi(\mu(t, \cdot))$ is differentiable at $t = 0$.

Let $\dot{\mathcal{P}}_P$ be the tangent set at P, and let $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ be the closure of the linear span of $\dot{\mathcal{P}}_P$, where the closure is taken relative to the topology induced by the standard L^2 metric on $L^2(\mathcal{E}, P)$. We say that $\psi(\cdot)$ is differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear transformation $\dot{\psi}_P : L^2(\mathcal{E}, P) \to \mathbb{R}^d$ such that, for any $g \in \dot{\mathcal{P}}_P$ there exists some one-dimensional submodel μ with score function $g(\cdot)$ such that

$$\lim_{t \to 0} \frac{\psi(\mu(t, \cdot)) - \psi(P)}{t} = \dot{\psi}_P(g).$$

Note the similarities of this definition to the directional derivative of a multivariate function. Since the measure $\mu(t, \cdot)$ is an approximation to P in the "direction" $g(\cdot)$ for small t, it follows that the limit of $\frac{\psi(\mu(t,\cdot)) - \psi(P)}{t}$ as t goes to 0 can be interpreted as the directional derivative of $\psi(\cdot)$ at P in the direction of $g(\cdot)$. Analogously to the case of directional derivatives for functions defined on euclidean spaces, the directional derivative of P in the direction $g(\cdot)$ is given as a linear function of $g(\cdot)$, where the function depends on P and can be interpreted as the total

derivative of $\psi(\cdot)$ at P.

We emphasize here that the above equation need only hold for at least one one-dimensional submodel of $\mathcal{P}$ of g for $\psi(\cdot)$ to be differentiable at P relative to $\dot{\mathcal{P}}_P$. Theoretically this is because the properties of the optimal limiting distribution is determined by the score functions in $\dot{\mathcal{P}}_P$, so that the specific submodel for which $\psi(\cdot)$ is differentiable becomes irrelevant. The irrelevance of the choice of μ is also advantageous from a technical standpoint, since in many cases we are only interested in certain submodels with a given score function. For instance, in semiparametric models, we are only interested in submodels that are separable in the parameter of interest and the nuisance parameter.

For each $1 \leq i \leq d$, define $\dot{\psi}_{P,i} : L^2(\mathcal{E}, P) \to \mathbb{R}$ as the ith coordinate function of $\dot{\psi}_P(\cdot)$. Since $\dot{\psi}_{P,i}$ is a continuous linear functional on the Hilbert space $(\overline{\text{sp}(\dot{\mathcal{P}}_P)}, \left\langle \cdot, \cdot \right\rangle_2)$ defined over the real field, where $\left\langle \cdot, \cdot \right\rangle_2$ is the inner product on $L^2(\mathcal{E}, P)$, it follows from the Riesz-Frechet representation theorem that there exists a unique $L_i \in L^2(\mathcal{E}, P)$ such that

$$\dot{\psi}_{P,i}(g) = \left\langle L_i,\, g \right\rangle_2 = \int_E L_i g \, dP$$

for any $g \in L^2(\mathcal{E}, P)$. Since we are interested in the value of $\dot{\psi}_{P,i}(\cdot)$ on the subset $\dot{\mathcal{P}}_P$, we would like to choose a square integrable function L_i that takes values in this set. While this is not always possible, we can take the next best option and choose a function L_i that lies in the closure of the linear span of $\dot{\mathcal{P}}_P$. $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ is a closed linear subspace of $L^2(\mathcal{E}, P)$, so by the Hilbert projection theorem, there exists a unique $\tilde{\psi}_{P,i} \in \overline{\text{sp}(\dot{\mathcal{P}}_P)}$ such that

$$\left\langle L_i - \tilde{\psi}_{P,i},\, g \right\rangle_2 = 0$$

for any $g \in \overline{\text{sp}(\dot{\mathcal{P}}_P)}$. In other words, $\tilde{\psi}_{P,i}$ is the orthogonal projection of L_i onto $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$. It follows that, for any $g \in \dot{\mathcal{P}}_P$, since $g \in \overline{\text{sp}(\dot{\mathcal{P}}_P)}$ as well,

$$\dot{\psi}_{P,i}(g) = \left\langle L_i,\, g \right\rangle_2 = \left\langle \tilde{\psi}_{P,i},\, g \right\rangle_2.$$

This holds for each $1 \leq i \leq d$, so defining $\tilde{\psi}_P : E \to \mathbb{R}^d$ as

$$\tilde{\psi}_P(x) = \begin{pmatrix} \tilde{\psi}_{P,1}(x) \\ \vdots \\ \tilde{\psi}_{P,d}(x) \end{pmatrix}$$

for any $x \in E$, each component of $\tilde{\psi}_P(\cdot)$ lies in the closed linear subspace $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$, and

$$\dot{\psi}_P(g) = \int_E \tilde{\psi}_P g \, dP$$

for any $g \in \overline{\text{sp}(\dot{\mathcal{P}}_P)}$. We call $\tilde{\psi}_P(\cdot)$ the efficient influence function.

9.4.2 Revisiting the Convolution Theorem

Having generalized the notions of differentiability in quadratic mean and differentiability of the mapping $\psi(\cdot)$, we can now define the generalized CRLB for the model $\mathcal{P}$. Let $\dot{\mathcal{P}}_P$ be the tangent set, $\psi : \mathcal{P} \to \mathbb{R}^d$ a function that is differentiable at P relative to $\dot{\mathcal{P}}_P$, and $\tilde{\psi}_P(\cdot)$ the corresponding efficient influence function. Let μ be some one-dimensional submodel of $\mathcal{P}$ with score function $g(\cdot)$ that is non-trivial, that is, it contains some information about $\psi(P)$. Formally, this is equivalent to the requirement that $I(g) = \int_E g^2 dP > 0$. Since $I(g) = 0$ implies that $g = 0$ P-a.e. by the vanishing property for non-negative functions, and P-almost surely equivalent functions are identified as members of $L^2(\mathcal{E}, P)$, it follows that any $g \in \dot{\mathcal{P}}_P$ such that $g \neq 0$ contains at least some information about $\psi(P)$.

For any non-trivial score function $g \in \dot{\mathcal{P}}_P$, since

$$\lim_{n\to\infty} \frac{\psi(\mu(t,\cdot)) - \psi(P)}{t} = \int_E \tilde{\psi}_P g \, dP,$$

for some one-dimensional submodel μ with score function g, the CRLB for μ is given as

$$\frac{1}{I(g)} \left(\int_E \tilde{\psi}_P g \, dP \right) \left(\int_E \tilde{\psi}_P g \, dP \right)'.$$

As indicated at the beginning of the current section, a sensible definition of information for the model $\mathcal{P}$ is as the infimum of the information contained in all finite-dimensional submodels. Therefore, we define the CRLB for $\mathcal{P}$ as the matrix V_P such that

$$V_P - \frac{1}{I(g)} \left(\int_E \tilde{\psi}_P g \, dP \right) \left(\int_E \tilde{\psi}_P g \, dP \right)'$$

is positive semidefinite for any $g \in \dot{\mathcal{P}}_P \setminus \{0\}$. The next result furnishes a positive semidefinite matrix that can serve the role of V_P:

Lemma 9.19 (The Generalized Information Lower Bound)
Let $(E, \mathcal{E})$ be a measurable space, and P a probability measure in some statistical model $\mathcal{P} \subset \mathcal{M}_+(\mathcal{E})$. Let $\dot{\mathcal{P}}_P \subset L^2(\mathcal{E}, P)$ be the tangent set of the model $\mathcal{P}$ at P, $\psi : \mathcal{P} \to \mathbb{R}^d$ a function that is differentiable at P relative to $\dot{\mathcal{P}}_P$, and $\tilde{\psi}_P(\cdot)$ the corresponding efficient influence function. Then,

$$\alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}_P' dP \right) \alpha = \sup_{g \in \mathrm{sp}(\dot{\mathcal{P}}_P) \setminus \{0\}} \frac{\left(\int_E (\alpha' \tilde{\psi}_P) g \, dP \right)^2}{\int_E g^2 dP}$$

for any $\alpha \in \mathbb{R}^d$. By implication,

$$\int_E \tilde{\psi}_P \tilde{\psi}_P' dP - \frac{1}{\int_E g^2 dP} \left(\int_E \tilde{\psi}_P g \, dP \right) \left(\int_E \tilde{\psi}_P g \, dP \right)'$$

is positive semidefinite for any non-trivial score function $g \in \dot{\mathcal{P}}_P$.

Proof) Fix any $\alpha \in \mathbb{R}^k$. Because the components of $\tilde{\psi}_P(\cdot)$ are contained in $L^2(\mathcal{E}, P)$, it follows that $\int_E \tilde{\psi}_P \tilde{\psi}'_P dP$ is a real-valued positive semidefinite matrix, and that

$$\alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}'_P dP \right) \alpha < +\infty.$$

By Hölder's inequality,

$$\left(\int_E (\alpha' \tilde{\psi}_P) g dP \right)^2 \leq \left[\alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}'_P dP \right) \alpha \right] \left(\int_E g^2 dP \right),$$

for any $g \in L^2(\mathcal{E}, P)$, so that

$$\frac{\left(\int_E (\alpha' \tilde{\psi}_P) g dP \right)^2}{\int_E g^2 dP} \leq \alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}'_P dP \right) \alpha$$

for any $g \in \mathrm{sp}(\dot{\mathcal{P}}_P) \setminus \{0\}$. It follows that

$$\sup_{g \in \mathrm{sp}(\dot{\mathcal{P}}_P) \setminus \{0\}} \frac{\left(\int_E (\alpha' \tilde{\psi}_P) g dP \right)^2}{\int_E g^2 dP} \leq \alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}'_P dP \right) \alpha.$$

To establish the opposite inequality, suppose that

$$c := \sup_{g \in \mathrm{sp}(\dot{\mathcal{P}}_P) \setminus \{0\}} \frac{\left(\int_E (\alpha' \tilde{\psi}_P) g dP \right)^2}{\int_E g^2 dP} < \alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}'_P dP \right) \alpha = \left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2.$$

It must then be the case that $\left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2 > 0$, since $c \geq 0$, and that $c < +\infty$. By implication, the relationship

$$c \left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2 < \left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2^2,$$

or equivalently

$$\left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2^2 - c \left\langle \alpha' \tilde{\psi}_P,\, \alpha' \tilde{\psi}_P \right\rangle_2 > 0,$$

holds. $\alpha' \tilde{\psi}_P$ is contained in $\overline{\mathrm{sp}(\dot{\mathcal{P}}_P)}$, so there exists a sequence $\{g_n\}_{n \in N_+}$ of functions in $\mathrm{sp}(\dot{\mathcal{P}}_P)$ that converges to $\alpha' \tilde{\psi}_P$ in mean square:

$$\lim_{n \to \infty} \left\| g_n - \alpha' \tilde{\psi}_P \right\|_2 = 0.$$

By the continuity of the mappings

$$g \mapsto \left\langle \alpha' \tilde{\psi}_P,\, g \right\rangle_2^2 - c \left\langle g,\, g \right\rangle_2$$

$$g \mapsto \Big\langle g, g \Big\rangle_2$$

on $L^2(\mathcal{E}, P)$ with respect to the L^2-norm, it follows that there exists some $N \in N_+$ such that

$$0 < \Big\langle \alpha' \tilde{\psi}_P,\, g_N \Big\rangle_2^2 - c \Big\langle g_N,\, g_N \Big\rangle_2.$$

and $0 < \Big\langle g_N,\, g_N \Big\rangle_2$. In other words,

$$c < \frac{\Big\langle \alpha' \tilde{\psi}_P,\, g_N \Big\rangle_2^2}{\Big\langle g_N,\, g_N \Big\rangle_2} = \frac{\Big(\int_E (\alpha' \tilde{\psi}_P) g_N dP\Big)^2}{\int_E g_N^2 dP}.$$

This contradicts the fact that $g_N \in \mathrm{sp}(\dot{\mathcal{P}}_P)$. It follows that

$$c = \alpha' \left(\int_E \tilde{\psi}_P \tilde{\psi}_P' dP \right) \alpha,$$

which is exactly what we wanted to show.

Q.E.D.

Note that this implies that the integral $\int_E \tilde{\psi}_P \tilde{\psi}_P' dP$ should serve as the lower bound for the asymptotic variance of estimators of $\psi(P)$ in the model $\mathcal{P}$. Before we formalize this notion for regular estimators via a generalization of the convolution theorem, a few remarks are in order. Recall that, in the parametric case, the CRLB was also given as the integral of the outer product of the efficient influence function under iid data. In this sense, our result is a true generalization of the result for parametric models, and it suggests that estimators $\mathcal{T}_n$ that are asymptotically linear as

$$\sqrt{n}\,[\mathcal{T}_n(X^n) - \psi(P)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\psi}_P(X_i) + o_p(1),$$

where $\{X_i\}_{i \in N_+}$ is a sequence of random variables with distribution P, are asymptotically efficient among the class of regular statistics.

Below we formalize the above intuition. First, we define regular statistics in the current non-parametric context. We say that the measurable function $\mathcal{T}_n : E^n \to \mathbb{R}^d$ is a regular statistic relative to the tangent set $\dot{\mathcal{P}}_P$ if there exists a d-dimensional random vector $\mathcal{T}$ if

$$\sqrt{n}\,[\mathcal{T}_n - \psi(\mu_g(1/\sqrt{n}, \cdot))] \overset{\mu_{g,n}(1/\sqrt{n},\cdot)}{\Rightarrow} \mathcal{T}$$

as $n \to \infty$ for any $g \in \dot{\mathcal{P}}_P$, where μ_g is a one-dimensional submodel of $\mathcal{P}$ with score function $g(\cdot)$ and $\{\mu_{g,n}\}_{n \in N_+}$ is the statistical model generated by μ_g. Note that $\mathcal{T}_n$ is regular in the sense that the limiting vector $\mathcal{T}$ does not depend on the choice of score function $g(\cdot)$. This is an

appropriate generalization of regular statistics in parametric models because the score function, in some sense, represents the direction from which the corresponding one-dimensional submodel passes through P. As such, the limiting distributions of regular statistics in the model $\mathcal{P}$ do not depend on $g(\cdot)$ much like the limiting distributions of regular statistics in parametric models do not depend on the direction $h \in \mathbb{R}^k$ from which the local alternative approaches the true parameter value.

The next theorem shows us that the limiting distribution L for any regular statistic in the model $\mathcal{P}$ can be expressed as the sum of two random vectors, one of which has distribution $\mathcal{N}\left[O_{d\times 1},\ \int_E \tilde{\psi}_P \tilde{\psi}_P' dP\right]$. In light of Anderson's lemma, this shows us that the normal distribution with mean 0 and variance $\int_E \tilde{\psi}_P \tilde{\psi}_P' dP$ is the optimal limiting distribution among the class of regular statistics.

Theorem 9.20 (The Generalized Convolution Theorem)
Let $(E, \mathcal{E})$ be a measurable space, and P a probability measure in some statistical model $\mathcal{P} \subset \mathcal{M}_+(\mathcal{E})$. Let $\dot{\mathcal{P}}_P \subset L^2(\mathcal{E}, P)$ be the tangent set of the model $\mathcal{P}$ at P, $\psi : \mathcal{P} \to \mathbb{R}^d$ a function that is differentiable at P relative to $\dot{\mathcal{P}}_P$, and $\tilde{\psi}_P(\cdot)$ the corresponding efficient influence function. Suppose $\{\mathcal{T}_n\}_{n\in N_+}$ is a sequence such that each $\mathcal{T}_n : E^n \to \mathbb{R}^d$ is measurable and there exists a d-dimensional random vector $\mathcal{T}$ such that

$$\sqrt{n}\left[\mathcal{T}_n - \psi(\mu_g(1/\sqrt{n}, \cdot))\right] \overset{\mu_{g,n}(1/\sqrt{n},\cdot)}{\Rightarrow} \mathcal{T} \sim L$$

as $n \to \infty$ for any $g \in \dot{\mathcal{P}}_P$, where μ_g is a one-dimensional submodel of $\mathcal{P}$ with score function $g(\cdot)$ and $\{\mu_{g,n}\}_{n\in N_+}$ is the statistical model generated by μ_g. Suppose that

$$\int_E \tilde{\psi}_P \tilde{\psi}_P' dP$$

is a positive definite matrix, and that $\dot{\mathcal{P}}_P$ is a linear subspace of $L^2(\mathcal{E}, P)$. Then, there exists some probability measure λ on $\mathbb{R}^d$ such that

$$L \sim \mathcal{N}\left[O_{d\times 1},\ \int_E \tilde{\psi}_P \tilde{\psi}_P' dP\right] * \lambda.$$

By implication, if $\int_E |x|^2 dL(x) < +\infty$, then

$$\int_E xx' dL(x) - \int_E \tilde{\psi}_P \tilde{\psi}_P' dP$$

is positive semidefinite.

Proof) The proof proceeds in a similar fashion to the proof of the convolution theorem by considering finite-dimensional subspaces of the tangent space $\dot{\mathcal{P}}_P$. We first show that such a finite-dimensional approximation works for our present purposes, then prove the convolution theorem for the finite-dimensional subspaces of interest.

Approximation by Finite Orthonormal Sets of Score Functions

Because each component of $\tilde{\psi}_P$ is contained in the closure of $\dot{\mathcal{P}}_P = \mathrm{sp}(\dot{\mathcal{P}}_P)$ with respect to the metric induced by the L^2-norm on $L^2(\mathcal{E}, P)$, it follows that there exists some sequence $\{k_m\}_{m \in N_+}$ of d-dimensional functions on E such that each component of k_m is contained in $\mathrm{sp}(\dot{\mathcal{P}}_P)$ and

$$\lim_{m \to \infty} \int_E \left| \tilde{\psi}_P - k_m \right|^2 dP = 0.$$

Each component of k_m is a score function in $\dot{\mathcal{P}}_P$, so it is contained in some finite-dimensional linear subspace of $\dot{\mathcal{P}}_P$ that contains k_m. Since every finite-dimensional linear subspace has an orthonormal basis, it follows that there exists a collection $\{g_1^m, \cdots, g_{s_m}^m\}$ of orthonormal vectors in $\dot{\mathcal{P}}_P$ such that the components of k_m are contained in the span of $\{g_1^m, \cdots, g_{s_m}^m\}$.

Defining $G_m = (g_1^m, \cdots, g_{s_m}^m)'$, let $H_m \in \mathbb{R}^{d \times s_m}$ be a non-random matrix such that $H_m G_m$ collects the orthogonal projections of the components of $\tilde{\psi}_P$ onto the span of $\{g_1^m, \cdots, g_{s_m}^m\}$. By the definition of the orthogonal projection,

$$\int_E \left| \tilde{\psi}_P - H_m G_m \right|^2 dP \leq \lim_{m \to \infty} \int_E \left| \tilde{\psi}_P - k_m \right|^2 dP$$

for any $m \in N_+$, so that

$$\lim_{m \to \infty} \int_E \left| \tilde{\psi}_P - H_m G_m \right|^2 dP = 0$$

as well. Furthermore, by the characterization of orthogonal projections, we can see that the identity

$$\int_E \left(\tilde{\psi}_P - H_m G_m \right) G_m' dP = O_{d \times s_m}$$

holds for any $m \in N_+$. By implication,

$$\left(\int_E \tilde{\psi}_P G_m' dP \right) \left(\int_E \tilde{\psi}_P G_m' dP \right)' = H_m \left(\int_E G_m G_m' dP \right) \left(\int_E G_m G_m' dP \right)' H_m' = H_m H_m',$$

where we used the fact that the vectors contained in G_m are orthonormal, so that $\int_E G_m G_m' dP = I_{s_m}$. The orthogonality condition also implies that

$$\begin{aligned} \int_E \tilde{\psi}_P \tilde{\psi}_P' dP - H_m H_m' &= \int_E \left(\tilde{\psi}_P \tilde{\psi}_P' - H_m G_m G_m' H_m \right) dP \\ &\leq \int_E \left(\tilde{\psi}_P - H_m G_m \right) \left(\tilde{\psi}_P - H_m G_m \right)' dP, \end{aligned}$$

so that

$$\left\| \int_E \tilde{\psi}_P \tilde{\psi}_P' dP - H_m H_m' \right\| \leq \int_E \left| \tilde{\psi}_P - H_m G_m \right|^2 dP \to 0$$

as $m \to \infty$. By implication,

$$\lim_{m\to\infty} \left(\int_E \tilde{\psi}_P G_m' dP\right)\left(\int_E \tilde{\psi}_P G_m' dP\right)' = \int_E \tilde{\psi}_P \tilde{\psi}_P' dP.$$

In particular, because $\int_E \tilde{\psi}_P \tilde{\psi}_P' dP$ is positive definite by assumption (in other words, it has a positive determinant), for large enough m the positive semidefinite matrix

$$\left(\int_E \tilde{\psi}_P G_m' dP\right)\left(\int_E \tilde{\psi}_P G_m' dP\right)'$$

is also positive definite.

Convolution Theorem for Finite-Dimensional Subspaces of $\dot{\mathcal{P}}_P$

Now fix any $m \in N_+$ that is large enough in the above sense. Then, because $\dot{\mathcal{P}}_P$ is a vector space and $h'G_m$ is a linear combination of score functions, for any $h \in \mathbb{R}^{s_m}$ we have $g = h'G_m \in \dot{\mathcal{P}}_P$. Furthermore, by the assumption that $\psi(\cdot)$ is differentiable at P relative to $\dot{\mathcal{P}}_P$ and the definition of the efficient influence function $\tilde{\psi}_P$, there exists some one-dimensional submodel μ_g of $\mathcal{P}$ with score function g satisfying

$$\lim_{n\to\infty} \frac{\psi(\mu_{h'G_m}(1/\sqrt{n},\cdot)) - \psi(P)}{\frac{1}{\sqrt{n}}} = \left(\int_E \tilde{\psi}_P G_m' dP\right) h,$$

or equivalently

$$\sqrt{n}\left[\psi(\mu_{h'G_m}(1/\sqrt{n},\cdot)) - \psi(P)\right] \overset{\mu_{h'G_m,n}}{\Rightarrow} {}^{(1/\sqrt{n},\cdot)} \left(\int_E \tilde{\psi}_P G_m' dP\right) h.$$

By the regularity assumption, we also have

$$\sqrt{n}\left[\mathcal{T}_n - \psi\left(\mu_{h'G_m}(1/\sqrt{n},\cdot)\right)\right] \overset{\mu_{h'G_m,n}}{\Rightarrow} {}^{(1/\sqrt{n},\cdot)} \mathcal{T},$$

and as such, by Slutsky's theorem,

$$\sqrt{n}\left[\mathcal{T}_n - \psi(P)\right] \overset{\mu_{h'G_m,n}}{\Rightarrow} {}^{(1/\sqrt{n},\cdot)} \mathcal{T} + \left(\int_E \tilde{\psi}_P G_m' dP\right) h.$$

Letting $\{X_i\}_{i\in N_+}$ be a sequence of random variables with distribution P, we saw in lemma 9.18 that

$$\frac{d\mu_{h'G_m,n}(1/\sqrt{n},\cdot)}{dP^n}(X_i) - \exp\left(h'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n G_m(X_i)\right) - \frac{1}{2}h'h\right) = o_p(1),$$

where

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_m(X_i) \xrightarrow{d} \Delta_0 \sim \mathcal{N}[O_{s_m \times 1},\, I_{s_m}]$$

by the Lindeberg-Levy CLT and the fact that the components of the d-dimensional function G_m are orthonormal.

This holds for any $h \in \mathbb{R}^{s_m}$, so by a process identical to the proof of the main approximation theorem, theorem 9.9, this implies that there exists some randomized statistic $\mathcal{S}_m : \mathbb{R}^{s_m} \times [0,1] \to \mathbb{R}^d$ from the Gaussian location model with variance equal to the identity matrix I_{s_m} such that

$$\mathcal{T} + \left(\int_E \tilde{\psi}_P G_m' dP\right) h \overset{h}{\sim} \mathcal{S}_m(X,U)$$

for any $h \in \mathbb{R}^{s_m}$, where X is a single observation from the Gaussian location model with variance equal to the identity matrix I_{s_m} and U is a uniformly distributed random variable independent of X. In particular,

$$\mathcal{S}_m(X,U) - \left(\int_E \tilde{\psi}_P G_m' dP\right) h \overset{h}{\sim} \mathcal{T},$$

for any $h \in \mathbb{R}^k$, so that $\mathcal{S}_m(\cdot)$ is an equivariant randomized statistic from the Gaussian location model. It follows from result ii) of lemma 9.14 that there exist s_m-dimensional independent random vectors V_m and W_m such that

$$V_m \sim \mathcal{N}\left[O_{s_m \times 1},\, \left(\int_E \tilde{\psi}_P G_m' dP\right)\left(\int_E \tilde{\psi}_P G_m' dP\right)'\right]$$

$$\mathcal{T} \sim V_m + W_m.$$

The Main Result

We have shown that, for any $r \in \mathbb{R}^d$ and $m \in N_+$,

$$\begin{aligned}\varphi(r) = \mathbb{E}\left[e^{ir'\mathcal{T}}\right] &= \mathbb{E}\left[e^{ir'V_m}\right]\mathbb{E}\left[e^{ir'W_m}\right] \\ &= \mathbb{E}\left[e^{ir'W_m}\right] \cdot e^{-\frac{1}{2}r'\left(\int_E \tilde{\psi}_P G_m' dP\right)\left(\int_E \tilde{\psi}_P G_m' dP\right)'r}.\end{aligned}$$

We saw above that

$$\lim_{m\to\infty}\left(\int_E \tilde{\psi}_P G_m' dP\right)\left(\int_E \tilde{\psi}_P G_m' dP\right)' = \int_E \tilde{\psi}_P \tilde{\psi}_P' dP,$$

so that

$$\lim_{m\to\infty} \mathbb{E}\left[e^{ir'W_m}\right] = \varphi(r)e^{\frac{1}{2}r'(\int_E \tilde{\psi}_P\tilde{\psi}'_P dP)r}$$

The mapping

$$r \mapsto \varphi(r)e^{\frac{1}{2}r'(\int_E \tilde{\psi}_P\tilde{\psi}'_P dP)r}$$

is continuous on $\mathbb{R}^k$, and in particular at $O_{d\times 1}$, so by the continuity theorem there exists a d-dimensional random vector W such that $W_m \xrightarrow{d} W$ as $m \to \infty$. By implication,

$$\begin{aligned}
\varphi(r) &= \lim_{m\to\infty} \mathbb{E}\left[e^{ir'W_m}\right] \cdot e^{-\frac{1}{2}r'(\int_E \tilde{\psi}_P G'_m dP)(\int_E \tilde{\psi}_P G'_m dP)'r} \\
&= \mathbb{E}\left[e^{ir'W}\right] \cdot e^{-\frac{1}{2}r'(\int_E \tilde{\psi}_P\tilde{\psi}'_P dP)r} \\
&= \mathbb{E}\left[e^{ir'(V+W)}\right],
\end{aligned}$$

where $V \sim \mathcal{N}\left[O_{d\times 1},\ \int_E \tilde{\psi}_P\tilde{\psi}'_P dP\right]$ and $V \perp W$. This holds for any $r \in \mathbb{R}^k$, so

$$\mathcal{T} \sim V + W.$$

Letting λ be the distribution of W, this implies that

$$L \sim \mathcal{N}\left[O_{d\times 1},\ \int_E \tilde{\psi}_P\tilde{\psi}'_P dP\right] * \lambda.$$

The claim that $\int_E xx'dL(x) - \int_E \tilde{\psi}_P\tilde{\psi}'_P dP$ is positive semidefinite if $\int_E |x|^2 dL(x) < +\infty$ now follows in the same manner as in theorem 9.16 via Anderson's lemma.

Q.E.D.

The preceding theorem shows us that, in the special case where the tangent set $\dot{\mathcal{P}}_P$ is itself a linear subspace of $L^2(\mathcal{E}, P)$, the positive definite matrix $\int_E \tilde{\psi}_P\tilde{\psi}'_P dP$ serves as the lower bound for the asymptotic variance of regular statistics. Furthermore, the convolution theorem, together with Anderson's lemma, tells us that the asymptotic risk under a normal distribution centered at the origin with variance $\int_E \tilde{\psi}_P\tilde{\psi}'_P dP$ is always lower than that of any regular statistic, so that we can once again formalize the notion of $\mathcal{N}\left[O_{d\times 1},\ \int_E \tilde{\psi}_P\tilde{\psi}'_P dP\right]$ as the optimal limiting distribution among regular statistics.

Naturally, we can also furnish asymptotically efficient regular estimators of $\psi(P)$ using the efficient influence function $\tilde{\psi}_P$ in the same manner through which we defined asymptotically efficient estimators in the parametric case. This is the content of the next theorem:

Theorem 9.21 (Asymptotically Efficient Estimators of $\psi(P)$)
Let $(E,\mathcal{E})$ be a measurable space, and P a probability measure in some statistical model $\mathcal{P}\subset\mathcal{M}_+(\mathcal{E})$. Let $\dot{\mathcal{P}}_P\subset L^2(\mathcal{E},P)$ be the tangent set of the model $\mathcal{P}$ at P, $\psi:\mathcal{P}\to\mathbb{R}^d$ a function that is differentiable at P relative to $\dot{\mathcal{P}}_P$, and $\tilde{\psi}_P(\cdot)$ the corresponding efficient influence function.Suppose that

$$\int_E \tilde{\psi}_P\tilde{\psi}_P' dP$$

is a positive definite matrix.

Let $\{\mathcal{T}_n^*\}_{n\in N_+}$ be a sequence of statistics satisfying

$$\sqrt{n}\,[\mathcal{T}_n^*(X_1,\cdots,X_n)-\psi(P)]=\frac{1}{\sqrt{n}}\sum_{i=1}^n\tilde{\psi}_P(X_i)+o_p(1)$$

for any sequence $\{X_i\}_{i\in N_+}$ of iid random variables with distribution P. Then, $\mathcal{T}_n^*$ is a regular statistic relative to $\dot{\mathcal{P}}_P$ such that

$$\sqrt{n}\,[\mathcal{T}_n^*-\psi(\mu_g(1/\sqrt{n},\cdot))]\overset{\mu_{g,n}(1/\sqrt{n},\cdot))}{\Rightarrow}\mathcal{T}^*\sim\mathcal{N}\left[O_{d\times 1},\int_E\tilde{\psi}_P\tilde{\psi}_P'dP\right]$$

for any $g\in\dot{\mathcal{P}}_P$, where μ_g is a one-dimensional submodel of $\mathcal{P}$ with score function $g(\cdot)$ and $\{\mu_{g,n}\}_{n\in N_+}$ is the statistical model generated by μ_g. If, in addition, $\dot{\mathcal{P}}_P$ is a linear subspace of $L^2(\mathcal{E},P)$, then $\mathcal{T}_n^*$ has the optimal limiting distribution among regular statistics for estimating $\psi(P)$.

Proof) We first show that $\mathcal{T}_n^*$ is a regular statistic. Choose any non-zero $g\in\dot{\mathcal{P}}_P$ and let μ_g be a one-dimensional submodel with score function g. Denote

$$Q_n=\mu_{g,n}(1/\sqrt{n},\cdot)$$

for any $n\in N_+$. By lemma 9.18, the submodel μ_g satisfies the LAN property: letting $\{X_i\}_{i\in N_+}$ be a sequence $\{X_i\}_{i\in N_+}$ of iid random variables with distribution P and denoting $X^n=(X_1,\cdots,X_n)$ for any $n\in N_+$, we have

$$\frac{dQ_n}{dP^n}(X^n)=\exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i)-\frac{1}{2}I(g)\right)+o_p(1),$$

where $I(g)=\int_E g^2dP>0$. By the Lindberg-Levy CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\begin{pmatrix}\tilde{\psi}_P(X_i)\\ g(X_i)\end{pmatrix}\xrightarrow{d}\mathcal{N}\left[O_{(d+1)\times 1},\begin{pmatrix}\int_E\tilde{\psi}_P\tilde{\psi}_P'dP & \int_E\tilde{\psi}_P g dP\\ \left(\int_E\tilde{\psi}_P g dP\right)' & \int_E g^2dP\end{pmatrix}\right].$$

Since

$$\left(\sqrt{n}\,[\mathcal{T}_n^*(X^n)-\psi(P)]\,,\quad \frac{dQ_n}{dP^n}(X^n)\right)=\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\tilde{\psi}_P(X_i),\quad \exp\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n g(X_i)-\frac{1}{2}I(g)\right)\right)+o_p(1)$$

for any $n\in N_+$, it follows from Slutsky's theorem that

$$\left(\sqrt{n}\,[\mathcal{T}_n^*(X^n)-\psi(P)]\,,\quad \log\frac{dQ_n}{dP^n}(X^n)\right)\xrightarrow{d}\mathcal{N}\left[\begin{pmatrix}O_{d\times 1}\\ -\frac{1}{2}\int_E g^2dP\end{pmatrix},\begin{pmatrix}\int_E\tilde{\psi}_P\tilde{\psi}_P'dP & \int_E\tilde{\psi}_P g dP\\ \left(\int_E\tilde{\psi}_P g dP\right)' & \int_E g^2dP\end{pmatrix}\right],$$

or equivalently,

$$\left(\sqrt{n}\,[\mathcal{T}_n^*-\psi(P)]\,,\quad \log\frac{dQ_n}{dP^n}\right)\overset{P^n}{\Rightarrow}\mathcal{N}\left[\begin{pmatrix}O_{d\times 1}\\ -\frac{1}{2}\int_E g^2dP\end{pmatrix},\begin{pmatrix}\int_E\tilde{\psi}_P\tilde{\psi}_P'dP & \int_E\tilde{\psi}_P g dP\\ \left(\int_E\tilde{\psi}_P g dP\right)' & \int_E g^2dP\end{pmatrix}\right].$$

By Le Cam's third lemma,

$$\sqrt{n}\,[\mathcal{T}_n^*-\psi(P)]\overset{Q_n}{\Rightarrow}\mathcal{N}\left[\int_E\tilde{\psi}_P g dP,\ \int_E\tilde{\psi}_P\tilde{\psi}_P'dP\right].$$

Meanwhile, by the definition of the efficient influence function,

$$\sqrt{n}\,[\psi(Q_n)-\psi(P)]\to\int_E\tilde{\psi}_P g dP$$

as $n\to\infty$, so by Slutsky's theorem,

$$\begin{aligned}\sqrt{n}\,[\mathcal{T}_n^*-\psi(Q_n)]&=\sqrt{n}\,[\mathcal{T}_n^*-\psi(P)]-\sqrt{n}\,[\psi(Q_n)-\psi(P)]\\ &\overset{Q_n}{\Rightarrow}\mathcal{N}\left[\int_E\tilde{\psi}_P g dP,\ \int_E\tilde{\psi}_P\tilde{\psi}_P'dP\right]-\int_E\tilde{\psi}_P g dP\\ &\sim\mathcal{N}\left[O_{d\times 1},\ \int_E\tilde{\psi}_P\tilde{\psi}_P'dP\right].\end{aligned}$$

This holds for any $g\in\dot{\mathcal{P}}_P$, so $\mathcal{T}_n^*$ is a regular statistic.

If, in addition, $\dot{\mathcal{P}}_P$ is a linear space, then by the generalized convolution theorem, the asymptotic variance of $\mathcal{T}_n^*$, $\int_E\tilde{\psi}_P\tilde{\psi}_P'dP$, is the lower bound on the asymptotic variance of regular statistics for estimating $\psi(P)$.

Q.E.D.

9.4.3 Application: The Parametric Model

Here we show that, in the special case that the model of interest is actually a parametric model, the usual convolution results for parametric models follows from the generalized convolution theorem. This demonstrates how the generalized notions of DQM, differentiability and the convolution theorem are truly generalizations of these concepts from a parametric to a nonparametric context.

Model Setup

Suppose that we are actually interested in a parametric model with k-dimensional parameter space (Θ, ϑ) and transition probability kernel μ from (Θ, ϑ) into $(E, \mathcal{E})$. Let $P = \mu(\theta_0, \cdot)$ for some $\theta_0 \in \Theta^o$, so that $\mathcal{P} = \{\mu(\theta, \cdot) \mid \theta \in \Theta\}$. Assume μ satisfies the DQM property at θ_0 with dominating measure λ and score function $g_0(\cdot)$:

$$\lim_{|h|\to 0} \frac{1}{|h|^2} \int_E \left(\sqrt{d\mu(\theta_0 + h, \cdot)} - \sqrt{d\mu(\theta_0, \cdot)} - \frac{1}{2} h' g_0 \cdot \sqrt{d\mu(\theta_0, \cdot)} \right)^2 d\lambda = 0.$$

Assume in addition that $\int_E g_0 g_0' d\mu(\theta_0, \cdot) = I_0$ is positive definite. We are interested in the value $\Psi(\theta_0)$, where $\Psi : \Theta \to \mathbb{R}^d$ is some function that is differentiable at θ_0 with derivative denoted by $\dot{\Psi}(\theta_0)$.

The Tangent Space

In this model, the maximal tangent set $\dot{\mathcal{P}}_P$ collects all linear combinations of the coordinates of g_0. To see this, note that any one-dimensional submodel v of $\mathcal{P}$ must be defined as

$$v(t, \cdot) = \mu(\theta_0 + th, \cdot)$$

for some $h \in \mathbb{R}^k$. Since

$$\int_E \left(\frac{\sqrt{dv(t,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} h' g_0 \sqrt{dP} \right)^2 d\lambda = \frac{1}{t^2} \int_E \left(\sqrt{d\mu(\theta_0 + th, \cdot)} - \sqrt{d\mu(\theta_0, \cdot)} - \frac{1}{2} (th)' g_0 \cdot \sqrt{d\mu(\theta_0, \cdot)} \right)^2 d\lambda$$

for any $t > 0$, it follows from the DQM property of μ that

$$\lim_{t\to 0} \int_E \left(\frac{\sqrt{dv(t,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} h' g_0 \sqrt{dP} \right)^2 d\lambda = 0,$$

so that $h' g_0$ is the score function corresponding to the one-dimensional submodel v; it follows that $\dot{\mathcal{P}}_P \subset \{h' g_0 \mid h \in \mathbb{R}^k\}$.

Conversely, choose any $h \in \mathbb{R}^k$. Defining $v(t, \cdot) = \mu(\theta_0 + th, \cdot)$ for any $t \geq 0$, it follows that v is a one-dimensional submodel of $\mathcal{P}$ with score function $h' g_0$, as we showed above. Therefore,

$h'g_0 \in \dot{\mathcal{P}}_P$, which shows that

$$\dot{\mathcal{P}}_P = \{h'g_0 \mid h \in \mathbb{R}^k\}.$$

This proves that $\dot{\mathcal{P}}_P$ is a linear subspace of $L^2(\mathcal{E}, P)$; the zero function is contained in $\dot{\mathcal{P}}_P$ (put $h = O_{k\times 1}$), while for any $a \in \mathbb{R}$ and $g_1, g_2 \in \dot{\mathcal{P}}_P$, letting $h_1, h_2 \in \mathbb{R}^k$ be chosen so that $g_1 = h_1'g_0$ and $g_2 = h_2'g_0$, we have

$$ag_1 + g_2 = (ah_1 + h_2)'g_0 \in \dot{\mathcal{P}}_P$$

because $ah_1 + h_2 \in \mathbb{R}^k$.

Differentiability of $\Psi(\cdot)$

Define the function $\psi : \mathcal{P} \to \mathbb{R}^d$ as

$$\psi(\mu(\theta, \cdot)) = \Psi(\theta)$$

for any $\theta \in \Theta$. By assumption, $\Psi(\cdot)$ is differentiable at θ_0 with derivative $\dot{\Psi}(\theta_0)$. We will show that this implies the function $\psi(\cdot)$ is differentiable at P relative to $\dot{\mathcal{P}}_P$ in the sense defined above.

To this end, choose any non-zero $g \in \dot{\mathcal{P}}_P$; by definition, there exists some non-zero $h \in \mathbb{R}^k$ such that $g = h'g_0$. g is the score function corresponding to the one-dimensional submodel v defined as

$$v(t, \cdot) = \mu(\theta_0 + th, \cdot)$$

for any $t \geq 0$, in the sense that

$$\int_E \left(\frac{\sqrt{dv(1/\sqrt{n}, \cdot)} - \sqrt{dP}}{t} - \frac{1}{2} g \sqrt{dP} \right)^2 dP$$

$$= |h|^2 \left[\frac{1}{\left(\frac{|h|}{\sqrt{n}}\right)^2} \int_E \left(\sqrt{d\mu(\theta_0 + h/\sqrt{n}, \cdot)} - \sqrt{d\mu(\theta_0, \cdot)} - \frac{1}{2} \frac{h'}{\sqrt{n}} g_0 \cdot \sqrt{d\mu(\theta_0, \cdot)} \right)^2 dP \right]$$

for any $n \in N_+$, so that

$$\lim_{n\to\infty} \int_E \left(\frac{\sqrt{dv(1/\sqrt{n}, \cdot)} - \sqrt{dP}}{t} - \frac{1}{2} g \sqrt{dP} \right)^2 dP = 0.$$

It follows that

$$\lim_{t\to 0} \frac{\psi(v(t, \cdot)) - \psi(P)}{t} = \lim_{t\to 0} \frac{\Psi(\theta_0 + th) - \Psi(\theta_0)}{t} = \dot{\Psi}(\theta_0)h = \int_E \left(\dot{\Psi}(\theta_0) I_0^{-1} g_0 \right) (h'g_0) dP$$

by the formula for the directional derivative. Defining the efficient influence function as

$$\tilde{\psi}_P(\cdot) = \dot{\Psi}(\theta_0) I_0^{-1} g_0(\cdot),$$

it follows that each component of $\tilde{\psi}_P$ is contained in the linear space $\dot{\mathcal{P}}_P$ and

$$\lim_{t \to 0} \frac{\psi(v(t,\cdot)) - \psi(P)}{t} = \int_E \tilde{\psi}_P g \, dP.$$

Meanwhile, if $g = 0$, then g is the score function corresponding to the one-dimensional submodel v defined as $v(t,\cdot) = \mu(\theta_0, \cdot)$ for any $t \geq 0$, so that

$$\lim_{t \to 0} \frac{\psi(v(t,\cdot)) - \psi(P)}{t} = 0 = \int_E \tilde{\psi}_P g \, dP.$$

In any case, for any $g \in \dot{\mathcal{P}}_P$ and one-dimensional submodel v of $\mathcal{P}$ with score function g, we have

$$\lim_{t \to 0} \frac{\psi(v(t,\cdot)) - \psi(P)}{t} = \int_E \tilde{\psi}_P g \, dP,$$

so that $\psi(\cdot)$ is differentiable at P relative to $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P(\cdot) = \dot{\Psi}(\theta_0) I_0^{-1} g_0(\cdot)$.

The Optimal Limiting Distribution for Regular Statistics

Note that

$$\int_E \tilde{\psi}_P \tilde{\psi}_P' dP = \dot{\Psi}(\theta_0) I_0^{-1} \dot{\Psi}(\theta_0)'.$$

Since $\dot{\Psi}(\theta_0)$ has rank d by assumption, $\int_E \tilde{\psi}_P \tilde{\psi}_P' dP$ is positive definite. $\dot{\mathcal{P}}_P$ is also a linear subspace of $L^2(\mathcal{E}, P)$, it follows from the convolution theorem that

$$\mathcal{N}\left[O_{d \times 1},\ \dot{\Psi}(\theta_0) I_0^{-1} \dot{\Psi}(\theta_0)'\right]$$

is the optimal limiting distribution for regular statistics. This shows us that the generalized convolution theorem is indeed a generalization of the same result for parametric models.

9.4.4 Application: A Completely Non-Parametric Model

Now we move onto the other extreme and consider a completely non-parametric model, that is, the case where the model $\mathcal{P}$ collects every probability measure on $(E,\mathcal{E})$. The only restriction we impose is that the true distribution P is dominated by some σ-finite measure λ on $(E,\mathcal{E})$, so that the DQM property is well-defined locally around P. Let $p_0(\cdot)$ be the density of P with respect to λ.

In this case, the maximal tangent set $\dot{\mathcal{P}}_P$ is the collection of functions in $g \in L^2(\mathcal{E},P)$ satisfying $\int_E g dP = 0$, where the latter condition must necessarily be satisfied by any score function (note lemma 9.18). We formally state and prove this below. First, we prove a preliminary lemma:

Lemma 9.22 Let $(E,\mathcal{E},\lambda)$ be a measure space and $\{f_n\}_{n\in N_+}$ a sequence of real-valued $\mathcal{E}$-measurable functions that converges pointwise to some $\mathcal{E}$-measurable function f. If

$$\limsup_{n\to\infty} \int_E |f_n|^2 d\lambda \leq \int_E |f|^2 d\lambda < +\infty,$$

then

$$\lim_{n\to\infty} \int_E |f_n - f|^2 d\lambda = 0.$$

Proof) Since $\limsup_{n\to\infty} \int_E |f_n|^2 d\lambda < +\infty$, there must exist some $N \in N_+$ such that $\int_E |f_n|^2 d\lambda < +\infty$ for any $n \geq N$. As such, for any $n \geq N$,

$$|f_n - f|^2 \leq (|f_n| + |f|)^2 \leq 2|f_n|^2 + 2|f|^2,$$

where the mapping on the right hand side is λ-integrable. It follows that

$$\int_E |f_n - f|^2 d\lambda \leq 2\left(\int_E |f_n|^2 d\lambda + \int_E |f|^2 d\lambda\right)$$

and as such

$$\limsup_{n\to\infty} \int_E |f_n - f|^2 d\lambda \leq 2\left(\limsup_{n\to\infty} \int_E |f_n|^2 d\lambda + \int_E |f|^2 d\lambda\right) \leq 4\int_E |f|^2 d\lambda < +\infty.$$

Now we can apply Fatou's lemma to see that

$$\begin{aligned}
4\int_E |f|^2 d\lambda &= \int_E \liminf_{n\to\infty} \left(2|f_n|^2 + 2|f|^2 - |f_n - f|^2\right) d\lambda \\
&\leq \liminf_{n\to\infty} \int_E \left(2|f_n|^2 + 2|f|^2 - |f_n - f|^2\right) d\lambda \\
&= \liminf_{n\to\infty} \int_E \left(2|f_n|^2 + 2|f|^2\right) d\lambda - \limsup_{n\to\infty} \int_E |f_n - f|^2 d\lambda \\
&\leq \limsup_{n\to\infty} \left(2\int_E |f_n|^2 d\lambda + 2\int_E |f|^2 d\lambda\right) - \limsup_{n\to\infty} \int_E |f_n - f|^2 d\lambda
\end{aligned}$$

$$\leq 2\left(\limsup_{n\to\infty}\int_E |f_n|^2 d\lambda\right) + 2\int_E |f|^2 d\lambda - \limsup_{n\to\infty}\int_E |f_n - f|^2 d\lambda$$

$$\leq 4\int_E |f|^2 d\lambda - \limsup_{n\to\infty}\int_E |f_n - f|^2 d\lambda.$$

Note that the integrability of $2|f_n|^2 + 2|f|^2$ and $|f_n - f|^2$ for large enough n allows us to apply the linearity of integration for integrable functions and justify the second equality. The fact that the limit superior of a sum is majorized by the sum of limit superiors is used to justify the third inequality. It follows that

$$\limsup_{n\to\infty}\int_E |f_n - f|^2 d\lambda \leq 0,$$

so that $\int_E |f_n - f|^2 d\lambda \to 0$ as $n \to \infty$.

Q.E.D.

Theorem 9.23 (Tangent Space for Non-Parametric Models)
Let $(E, \mathcal{E})$ be a measurable space, and P a probability measure in the completely non-parametric statistical model $\mathcal{P} = \mathcal{M}_+(\mathcal{E})$. Let $\dot{\mathcal{P}}_P \subset L^2(\mathcal{E}, P)$ be the maximal tangent set of the model $\mathcal{P}$ at P. Suppose that there exists some σ-finite measure λ on $(E, \mathcal{E})$ such that $P << \lambda$. Then,

$$\dot{\mathcal{P}}_P = \{g \in L^2(\mathcal{E}, P) \mid \int_E g \, dP = 0\}.$$

Proof) In light of lemma 9.18, it is clear that

$$\dot{\mathcal{P}}_P \subset \{g \in L^2(\mathcal{E}, P) \mid \int_E g dP = 0\}.$$

To establish the reverse inclusion, choose any $g \in L^2(\mathcal{E}, P)$ with $\int_E g dP = 0$. We will explicitly construct a one-dimensional submodel μ_g that satisfies the DQM property with score function g.

To this end, we consider the function $k : \mathbb{R} \to \mathbb{R}_+$ defined as

$$k(x) = \frac{2}{1 + e^{-2x}}$$

for any $x \in \mathbb{R}$. $k(\cdot)$ is bounded above by 2, and note that:

1. The mapping $x \mapsto k(tg(x))$ is $\mathcal{E}$-measurable and P-integrable for any $t \in \mathbb{R}$
2. The mapping $t \mapsto k(tg(x))g(x)$ is continuously differentiable on $\mathbb{R}$ with derivative

$$\frac{\partial}{\partial t} k(tg(x)) = \frac{4g(x)}{(1 + e^{-2tg(x)})^2}$$

for any $x \in E$.

3. For any $x \in E$,

$$\left| \frac{\partial}{\partial t} k(tg(x)) \right| \leq 4.$$

By the Leibniz rule for interchanging integration with differentiation, it now follows that the mapping $x \mapsto \frac{\partial}{\partial t} k(tg(x))$ is P-integrable for any $t \in \mathbb{R}$ and that

$$\frac{\partial}{\partial t} \int_E k(tg(x)) dP(x) = \int_E \left(\frac{\partial}{\partial t} k(tg(x)) \right) dP(x) = \int_E \frac{4g(x)}{(1+e^{-2tg(x)})^2} dP(x).$$

Note that, for any $t \in \mathbb{R}$, $\int_E k(tg(x)) dP(x) = 0$ implies by the vanishing property for non-negative functions that $k(tg(x)) = 0$ P-almost surely. Since $k(\cdot)$ is always positive valued, this is a contradiction. Thus, $\int_E k(tg(x)) dP(x) > 0$, and we can now define the mapping $c : \mathbb{R} \to \mathbb{R}_+$ as

$$c(t) = \left(\int_E k(tg(x)) dP(x) \right)^{-1}$$

for any $t \in \mathbb{R}$. Since $k(0) = 1$, it follows that $c(0) = 1$. It follows from the results we have shown so far that

$$c'(t) = -\left(\int_E k(tg(x)) dP(x) \right)^{-2} \left(\frac{\partial}{\partial t} \int_E k(tg(x)) dP(x) \right) = -\frac{1}{c(t)^2} \int_E \frac{4}{(1+e^{-2tg(x)})^2} g(x) dP(x)$$

for any $t \in \mathbb{R}$ by the chain rule. This proves that $c(\cdot)$ is continuous on $\mathbb{R}$; together with the BCT, this proves that $c'(\cdot)$ is continuous on $\mathbb{R}$ as well. In particular,

$$c'(0) = -\frac{1}{c(0)^2} \int_E g dP = 0.$$

Denote the density of P with respect to λ by $p_0 : E \to \mathbb{R}_+$, and define the one-dimensional submodel μ_g as

$$\mu_g(t, A) = \int_A c(t) k(tg(x)) p_0(x) \, d\lambda(x)$$

for any $(t, A) \in \mathbb{R} \times \mathcal{E}$. When the true parameter is $t \in \mathbb{R}$, the density of the model with respect to λ is given as $c(t)k(tg(\cdot))p_0(\cdot)$; the distribution P corresponds to the model with parameter $t = 0$. Defining

$$p(t, \cdot) = d\mu_g(t, \cdot) = c(t)k(tg(\cdot))p_0(\cdot)$$

for any $(t, x) \in \mathbb{R} \times E$, each $p(t, \cdot)$ is $\mathcal{E}$-measurable, and each $p(\cdot, x)$ is continuously

differentiable on $\mathbb{R}$ with derivative given as

$$\frac{\partial p(t,x)}{\partial t} = c'(t)k(tg(x))p_0(x) + c(t)k'(tg(x))g(x)p_0(x)$$

for any $(t,x) \in \mathbb{R} \times E$. Define $E_0 = \{p_0 > 0\}$. For any $x \in E_0$,

$$\begin{aligned}\frac{\partial}{\partial t}\sqrt{p(t,x)} &= \frac{1}{2}\frac{1}{\sqrt{p(t,x)}}\frac{\partial p(t,x)}{\partial t} \\ &= \frac{1}{2}\left[\frac{c'(t)}{\sqrt{c(t)}} + k(tg(x))g(x)\sqrt{c(t)}\right]\sqrt{k(tg(x))p_0(x)}\end{aligned}$$

for any $t \in \mathbb{R}$. Meanwhile, if $x \in E_0^c$, then $p(t,x) = 0$ for any $t \in \mathbb{R}$, so that

$$\lim_{s \to t}\frac{\sqrt{p(s,x)} - \sqrt{p(t,x)}}{s-t} = 0 = \frac{1}{2}\left[\frac{c'(t)}{\sqrt{c(t)}} + k(tg(x))g(x)\sqrt{c(t)}\right]\sqrt{k(tg(x))p_0(x)}$$

for any $t \in \mathbb{R}$. In any case, defining $s : \mathbb{R} \times E \to \mathbb{R}$ as

$$s(t,x) = \frac{1}{2}\left[\frac{c'(t)}{\sqrt{c(t)}} + k(tg(x))g(x)\sqrt{c(t)}\right]\sqrt{k(tg(x))p_0(x)}$$

for any $(t,x) \in \mathbb{R} \times E$, so that $s(\cdot)$ is jointly measurable, $s(\cdot,x)$ is continuous for any $x \in E$, and

$$\lim_{s \to t}\frac{\sqrt{p(s,x)} - \sqrt{p(t,x)}}{s-t} = s(t,x) = \frac{\partial}{\partial t}\sqrt{p(t,x)}$$

for any $(t,x) \in \mathbb{R} \times E$, so that $p(\cdot,x)$ is continuously differentiable on $\mathbb{R}$ for any $x \in E$. In particular,

$$\lim_{t \to 0}\frac{\sqrt{p(t,x)} - \sqrt{p_0(x)}}{t} = \frac{1}{2}g(x)\sqrt{p_0(x)}$$

for any $x \in E$.

Fix a sequence $\{t_n\}_{n \in N_+}$ of positive reals decreasing to 0, and define the functions $\{f_n\}_{n \in N_+}$ and f as

$$f_n(x) = \frac{\sqrt{p(t_n,x)} - \sqrt{p_0(x)}}{t_n}, \quad f(x) = \frac{1}{2}g(x)\sqrt{p_0(x)}$$

for any $n \in N_+$ and $x \in E$. Then, $f_n \to f$ pointwise on E. Furthermore, for any $x \in E$, since the mapping $\sqrt{p(\cdot,x)}$ is continuously differentiable on $\mathbb{R}$, for any $t > 0$ the second fundamental theorem of calculus implies that

$$\sqrt{p(t,x)} - \sqrt{p_0(x)} = \int_0^t s(z,x)dz \leq \sqrt{t}\left(\int_0^t s(z,x)^2 dz\right)^{\frac{1}{2}},$$

where we used Hölder's inequality to justify the last inequality. As such,

$$|f_n(x)|^2 = \left(\frac{\sqrt{p(t_n,x)}-\sqrt{p_0(x)}}{t_n}\right)^2 \leq \frac{1}{t_n}\left(\int_0^{t_n} s(z,x)^2 dz\right)$$

$$= \frac{1}{4t_n}\left[\int_0^{t_n}\left(\frac{c'(z)}{\sqrt{c(z)}}+g(x)\sqrt{c(z)}\right)^2 k(zg(x))p_0(x)dz\right]$$

for any $n \in N_+$, so that

$$\int_E |f_n|^2 d\lambda \leq \frac{1}{4t_n}\left[\int_0^{t_n}\int_E\left(\frac{c'(z)}{\sqrt{c(z)}}+k(zg(x))g(x)\sqrt{c(z)}\right)^2 k(zg(x))p_0(x)\, d\lambda(x)dz\right]$$

$$= \frac{1}{4t_n}\int_0^{t_n}\left[\frac{c'(z)^2}{c(z)}+2c'(z)\left(\int_E k(zg(x))^2 g(x)dP(x)\right)+c(z)\left(\int_E k(zg(x))^3 g(x)^2 dP(x)\right)\right]dz.$$

Since $k(\cdot)$ is a bounded function, the dominated convergence theorem implies that the mapping

$$z \mapsto \frac{c'(z)^2}{c(z)}+2c'(z)\left(\int_E k(zg(x))^2 g(x)dP(x)\right)+c(z)\left(\int_E k(zg(x))^3 g(x)^2 dP(x)\right)$$

is continuous on $\mathbb{R}$; therefore, by the first fundamental theorem of calculus,

$$\lim_{n\to\infty}\frac{1}{t_n}\int_0^{t_n}\left[\frac{c'(z)^2}{c(z)}+2c'(z)\left(\int_E k(zg(x))^2 g(x)dP(x)\right)+c(z)\left(\int_E k(zg(x))^3 g(x)^2 dP(x)\right)\right]dz$$

$$= \frac{\partial}{\partial t}\int_0^t\left[\frac{c'(z)^2}{c(z)}+2c'(z)\left(\int_E k(zg(x))^2 g(x)dP(x)\right)+c(z)\left(\int_E k(zg(x))^3 g(x)^2 dP(x)\right)\right]dz\,|_{t=0}$$

$$= \frac{c'(0)^2}{c(0)}+2c'(0)\left(\int_E k(0)^2 g(x)dP(x)\right)+c(0)\left(\int_E k(0)^3 g(x)^2 dP(x)\right)$$

$$= \int_E g^2 dP = 4\int_E |f|^2 d\lambda.$$

Therefore,

$$\limsup_{n\to\infty}\int_E |f_n|^2 d\lambda \leq \int_E |f|^2 d\lambda < +\infty,$$

so that, by the preceding lemma, we have

$$\lim_{n\to\infty}\int_E |f_n - f|^2 d\lambda = 0.$$

Equivalently,

$$\lim_{n\to\infty}\int_E\left(\frac{\sqrt{p(t_n,x)}-\sqrt{p_0(x)}}{t_n}-\frac{1}{2}g(x)\sqrt{p_0(x)}\right)^2 d\lambda = 0,$$

so by definition g is the score function corresponding to the one-dimensional submodel μ_g. This proves that $g \in \dot{\mathcal{P}}_P$, and the proof is complete.

Q.E.D.

Because the tangent space $\dot{\mathcal{P}}_P$ is so large for a completely non-parametric model, we can unfortunately say very little in general about the efficient influence function $\tilde{\psi}_P$ corresponding to any mapping $\psi : \mathcal{P} \to \mathbb{R}^d$ that is differentiable at P relative to $\dot{\mathcal{P}}_P$. Indeed, $\tilde{\psi}_P$ is determined entirely by the function $\psi(\cdot)$.

The most important takeaway from the preceding theorem is actually the way we constructed one-dimensional submodels passing through P with the desired score function $g(\cdot)$. The one-dimensional submodel μ_g was constructed as a parametric model satisfying

$$d\mu_g(t, \cdot) = c(t)k(tg(\cdot))p_0(\cdot)$$

for any $t \in \mathbb{R}$, where $p_0(\cdot)$ is the density of P with respect to some dominating measure λ, $c(t)$ is the normalizing constant, and $k(\cdot)$ is a positive, bounded and infinitely differentiable function such that

$$k(0) = 1, \quad k'(0) = 1, \quad k'(t) = k(t)^2$$

for any $t \in \mathbb{R}$. These conditions for $k(\cdot)$ ensure that μ_g satisfies the DQM property with score function $g(\cdot)$. This convenient construction for one-dimensional submodels will also appear later on when studying semiparametric models.

9.4.5 Application: Models Characterized by Moment Conditions

In the preceding sections, we discussed the two extremes of a parametric model determined by a finite-dimensional parameter and a completely non-parametric model where the parameter space is the collection of all probability measures. Often we work on something of a middle ground, that is, we impose some restrictions on the set of admissible probability measures, but nevertheless refrain from specifying an explicit parametric form. One popular restrictions that are imposed in the econometric literature are moment conditions; think of GMM methods. In this section we derive the tangent space for models that are subject only to moment conditions using methods analogous to those employed in the last section.

The mathematical framework is the following. As usual, let $(E,\mathcal{E})$ be a measurable space, and $\mathcal{P}$ a collection of probability measures that we treat as our model. Let $P \in \mathcal{P}$ be the true distribution, and $m : E \to \mathbb{R}^k$ a measurable function that is P-square integrable. We say that the model $\mathcal{P}$ satisfies the moment conditions for $m(\cdot)$ if

$$\mathcal{P} = \{Q \in \mathcal{M}_+(\mathcal{E}) \mid \int_E m \, dQ = O_{k\times 1}\}.$$

In other words, letting $\{X_i\}_{i\in N_+}$ be a random sample drawn from the distribution P, we must have $\mathbb{E}[m(X_i)] = O_{k\times 1}$.

Unlike in the completely non-parametric case discussed earlier, we cannot obtain a closed form expression for the actual tangent set $\dot{\mathcal{P}}_P$. However, it is possible to characterize the closure of the linear span of the tangent set, $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ for some subset $\dot{\mathcal{P}}_P$ of the maximal tangent set. Since the efficient influence function is contained in $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$, knowledge of the form of $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ suffices to recover the efficient influence function in many cases.

To obtain some intuition for how to characterize the tangent set, consider the simple case of regular parametric models. Let μ be a transition probability kernel from some d-dimensional parameter space Θ into E, and suppose that each $\mu(\theta,\cdot)$ is dominated by some measure λ on $(E,\mathcal{E})$ with density $p(\cdot \mid \theta)$. Assume that the moment condition $\int_E m d\mu(\theta,\cdot) = O_{k\times 1}$ holds for any $\theta \in \Theta$. Suppose the usual regularity conditions hold, so that the score function g_0 is given as the derivative of the log likelihood $\log p(\cdot \mid \theta)$ at $\theta = \theta_0$. Then, assuming that $p_0 > 0$ on E and the order of integration and differentiation can be interchanged (via a Leibniz rule-type result), we have

$$\begin{aligned}
\int_E (m g_0') dP &= \int_E m \left[\frac{\partial}{\partial \theta} \log p(\cdot \mid \theta_0)\right]' p(\cdot \mid \theta_0) d\lambda \\
&= \int_E m \left(\frac{\partial}{\partial \theta} p(\cdot \mid \theta_0)\right)' d\lambda \\
&= \frac{\partial}{\partial \theta'} \left[\int_E m p(\cdot \mid \theta_0) d\lambda\right] = \frac{\partial}{\partial \theta'} O_{k\times 1} = O_{k\times d}.
\end{aligned}$$

In other words, the components of the moment function m and the score function g_0 are orthogonal with respect to the inner product on $L^2(\mathcal{E},P)$. The next theorem generalizes this result and shows us that $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ is precisely the orthogonal complement of the linear subspace spanned by

the components of the moment function m.

Theorem 9.24 (Tangent Space for Models with Moment Conditions)
Let $(E,\mathcal{E})$ be a measurable space, and P a probability measure in the statistical model $\mathcal{P} \subset \mathcal{M}_+(\mathcal{E})$. We assume that there exists a function m whose components are contained in $L^2(\mathcal{E},P)$ and $\int_E (1,m)(1,m)'dP$ is positive definite such that $\mathcal{P}$ collects the probability measures that satisfy the moment condition with respect to $m=(m_1,\cdots,m_k)$ and are dominated by some σ-finite measure λ on $(E,\mathcal{E})$:

$$\mathcal{P} = \{Q \in \mathcal{M}_+(\mathcal{E}) \mid Q << \lambda, \quad \int_E m\, dQ = O_{k\times 1}\}.$$

For any $Q \in \mathcal{P}$, let $dQ \in \mathcal{E}_+$ denote the density of Q with respect to λ.

Let $\dot{\mathcal{P}}_P \subset L^2(\mathcal{E},P)$ be a tangent set of the model $\mathcal{P}$ at P that collects the score functions of any one-dimensional submodel μ of $\mathcal{P}$ that satisfies the DQM property under the dominating measure λ in addition to the following conditions:

i) The density $d\mu(t,\cdot)$ converges pointwise to the density dP as $t \to 0$.

ii) $\mu(t,\cdot)$ converges to P in squared moments:

$$\limsup_{t\to 0} \int_E m_i^2 d\mu(t,\cdot) \le \int_E m_i^2 dP \quad \forall 1 \le i \le k.$$

Then, the L^2-closure of the span of $\dot{\mathcal{P}}_P$ is exactly equal to the orthogonal complement of the span of the components of the moment function m:

$$\overline{\mathrm{sp}(\dot{\mathcal{P}}_P)} = (\mathrm{sp}\{1, m_1, \cdots, m_k\})^{\perp}.$$

Proof) Denote $W = \mathrm{sp}\{1, m_1, \cdots, m_k\}$. We first show that any score function $g \in \dot{\mathcal{P}}_P$ must be contained in $W^{\perp}$. Choose any $g \in \dot{\mathcal{P}}_P$ and let μ_g be the one-dimensional submodel of $\mathcal{P}$ with score function g. We can immediately see from lemma 9.18 that $\int_E g dP = 0$. Since every probability measure in $\mathcal{P}$ satisfies the moment condition with respect to m, it follows that

$$\int_E m\, d\mu_g(t,\cdot) = O_{k\times 1}$$

for any $t \ge 0$. In particular, $\int_E m\, dP = O_{k\times 1}$. Since any submodel associated with a score function in $\dot{\mathcal{P}}_P$ satisfies the DQM property under the dominating measure λ, we have

$$\lim_{t\to 0} \int_E \left(\frac{\sqrt{d\mu_g(t,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} g \sqrt{dP} \right)^2 d\lambda = 0.$$

By definition,

$$\frac{\sqrt{d\mu_g(t,\cdot)} - \sqrt{dP}}{t} \overset{L^2(\mathcal{E},\lambda)}{\to} \frac{1}{2} g \sqrt{dP},$$

as $t \to 0$, and multiplying both sides by t shows us that

$$\sqrt{d\mu_g(t,\cdot)} \overset{L^2(\mathcal{E},\lambda)}{\to} \sqrt{dP}$$

as $t \to 0$.

Now choose $1 \leq i \leq k$, and let $\{t_n\}_{n \in N_+}$ be a sequence of positive reals converging to 0. By the way we chose the tangent set $\dot{\mathcal{P}}_P$,

$$\lim_{n \to \infty} m_i \sqrt{d\mu_g(t_n,\cdot)} = m_i \sqrt{dP}$$

pointwise on E, and

$$\limsup_{n \to \infty} \int_E m_i^2 d\mu_g(t_n,\cdot) \leq \int_E m_i^2 dP < +\infty.$$

By lemma 9.22, this implies that

$$\lim_{n \to \infty} \int_E \left(m_i \sqrt{d\mu_g(t_n,\cdot)} - m_i \sqrt{dP} \right)^2 d\lambda = 0,$$

so that

$$m_i \sqrt{d\mu_g(t_n,\cdot)} \overset{L^2(\mathcal{E},\lambda)}{\to} m_i \sqrt{dP}$$

Let $\left\langle \cdot, \cdot \right\rangle_2$ be the usual inner product on $L^2(\mathcal{E},\lambda)$ and choose any $1 \leq i \leq k$. It now follows from the continuity of the inner product that

$$\begin{aligned}
\int_E (m_i g) dP &= \int_E \left(\frac{1}{2} g \sqrt{dP} \right) \left(2 m_i \sqrt{dP} \right) d\lambda = \left\langle \frac{1}{2} g \sqrt{dP},\, 2 m_i \sqrt{dP} \right\rangle_2 \\
&= \lim_{t \to 0} \left\langle \frac{\sqrt{d\mu_g(t,\cdot)} - \sqrt{dP}}{t},\, m_i \left(\sqrt{d\mu_g(t_n,\cdot)} + \sqrt{dP} \right) \right\rangle_2 \\
&= \lim_{t \to 0} \int_E \left(\frac{\sqrt{d\mu_g(t,\cdot)} - \sqrt{dP}}{t} \right) \left(\sqrt{d\mu_g(t_n,\cdot)} + \sqrt{dP} \right) m_i d\lambda.
\end{aligned}$$

For any $t \geq 0$,

$$\begin{aligned}
\int_E \left(\frac{\sqrt{d\mu_g(t,\cdot)} - \sqrt{dP}}{t} \right) \left(\sqrt{d\mu_g(t_n,\cdot)} + \sqrt{dP} \right) m_i d\lambda &= \int_E \frac{d\mu_g(t,\cdot) - dP}{t} m_i d\lambda \\
&= \frac{1}{t} \left(\int_E m_i d\mu_g(t,\cdot) - \int_E m_i dP \right) = 0
\end{aligned}$$

by the moment condition. Therefore, $\int_E (m_i g) dP = 0$, and because this holds for any $1 \leq i \leq k$, by definition $g \in W^\perp$. This proves that $\dot{\mathcal{P}}_P \subset W^\perp$.

Now choose any bounded $g \in W^\perp$, so that

$$\int_E mg \; dP = O_{k \times 1}.$$

We will to show that there exists a one-dimensional submodel μ of $\mathcal{P}$ with score function $g \in \dot{\mathcal{P}}_P$. To this end, we use a similar construction to that of theorem 9.23. Because g is bounded, there exists some $M > 0$ such that $|g| \leq M$ on E. Letting $p_0 = \frac{\partial P}{\partial \lambda}$, define the one-dimensional submodel μ_g as

$$\mu_g(t, A) = \int_A \underbrace{(1 + tg(x))p_0(x)}_{p(t,x)} d\lambda(x)$$

for any $t \in \Theta = \left[-\frac{1}{M}, \frac{1}{M}\right]$ and $A \in \mathcal{E}$. In other words, μ_g is a one-dimensional parametric model with parameter space Θ and density $x \mapsto (1 + tg(x))p_0(x)$ with respect to λ under the parameter $t \in \Theta$. The fact that $\int_E g dP = 0$ implies that $\mu_g(t, \cdot)$ is a well-defined probability measure under each $t \in \Theta$. Furthermore,

$$\begin{aligned} \int_E m \; d\mu_g(t, \cdot) &= \int_E m(1 + tg(x))p_0(x) d\lambda(x) \\ &= \int_E m \; dP + t \cdot \int_E mg \; dP = O_{k \times 1} \end{aligned}$$

for any $t \in \Theta$ by the way we chose g and the fact that P satisfies the moment condition. This proves that μ_g is a one-dimensional submodel of $\mathcal{P}$.

Note also that

$$d\mu_g(t, \cdot) = (1 + t \cdot g)p_0 \to p_0 = dP$$

as $t \to 0$ pointwise on E. For any $1 \leq i \leq k$ and $t \in \Theta$,

$$\int_E m_i^2 d\mu_g(t, \cdot) = \int_E m_i^2 (1 + tg) dP = \int_E m_i^2 dP + t \cdot \int_E m_i^2 g dP.$$

Here, since $|g| \leq M$,

$$\left| \int_E m_i^2 g dP \right| \leq \int_E m_i^2 |g| dP \leq M \left(\int_E m_i^2 dP \right) < +\infty,$$

so that

$$\lim_{t \to 0} \int_E m_i^2 d\mu_g(t, \cdot) = \int_E m_i^2 dP.$$

It remains to show that μ_g satisfies the DQM property with score function g to conclude that $g \in \dot{\mathcal{P}}_P$.

We can again make use of lemma 9.22 for this purpose. Define $E_0 = \{p_0 > 0\}$. Then, for any $x \in E_0$ and $t \in \Theta^o$, since $q(t,x) > 0$, we have

$$\begin{aligned}\frac{\partial}{\partial t}\sqrt{p(t,x)} &= \frac{1}{2}\left(p(t,x)\right)^{-\frac{1}{2}}\left(\frac{\partial}{\partial t}p(t,x)\right)\\ &= \frac{1}{2}\left(p(t,x)\right)^{-\frac{1}{2}} g(x)p_0(x) = \frac{1}{2\sqrt{1+tg(x)}}g(x)\sqrt{p_0(x)}.\end{aligned}$$

Meanwhile, if $x \notin E_0^c$, then $p(t,x) = 0$ for any $t \in \Theta$, so that

$$\lim_{s\to t}\frac{\sqrt{p(s,x)}-\sqrt{p_0(x)}}{s-t} = 0 = \frac{1}{2\sqrt{1+tg(x)}}g(x)\sqrt{p_0(x)}$$

for any $t \in \Theta^o$. In any case,

$$\lim_{s\to t}\frac{\sqrt{p(s,x)}-\sqrt{p_0(x)}}{s-t} = s(t,x)$$

for any $t \in \Theta^o$ and $x \in E$, where we define $s(t,x) = \frac{1}{2\sqrt{1+tg(x)}}g(x)\sqrt{p_0(x)}$.

Let $\{t_n\}_{n\in N_+}$ be a sequence of positive reals converging to 0, and define $\{f_n\}_{n\in N_+}$ as

$$f_n(x) = \frac{\sqrt{p(t_n,x)}-\sqrt{p_0(x)}}{t_n}$$

for any $n \in N_+$ and $x \in E$. We already saw that $f_n \to s(0,\cdot) = \frac{1}{2}g\sqrt{p_0}$ pointwise as $n \to \infty$. Furthermore, because $t \mapsto \sqrt{p(t,x)}$ is continuously differentiable on Θ^o for any $x \in E$, by the second fundamental theorem of calculus we have

$$\sqrt{p(t,x)}-\sqrt{p_0(x)} = \int_0^t s(r,x)dr \le \sqrt{t}\left(\int_0^t s(r,x)^2 dr\right)^{\frac{1}{2}}$$

for any $x \in E$, where the inequality is Hölder's. By implication,

$$\begin{aligned}|f_n|^2 = \left(\frac{\sqrt{p(t_n,\cdot)}-\sqrt{p_0}}{t_n}\right)^2 &\le \frac{1}{t_n}\int_0^{t_n} s(r,\cdot)^2 dr\\ &= \frac{1}{t_n}\left(\int_0^{t_n}\frac{1}{4(1+r\cdot g)}dr\right)(g^2 p_0),\end{aligned}$$

and by Fubini's theorem,

$$\int_E |f_n|^2 d\lambda \le \frac{1}{t_n}\int_0^{t_n}\left(\int_E \frac{g^2}{4(1+r\cdot g)}dP\right)dr.$$

The mapping

$$r \mapsto \int_E \frac{g^2}{4(1+r\cdot g)} dP$$

is continuous on the set $[-\frac{1}{2M}, \frac{1}{2M}] \subset \Theta^o$ by the DCT, so by the first fundamental theorem of calculus,

$$\lim_{n\to\infty} \frac{1}{t_n} \int_0^{t_n} \left(\int_E \frac{g^2}{4(1+r\cdot g)} dP \right) dr = \frac{\partial}{\partial t} \int_0^t \left(\int_E \frac{g^2}{4(1+r\cdot g)} dP \right) dr|_{t=0} = \frac{1}{4} \int_E g^2 dP.$$

In other words,

$$\limsup_{n\to\infty} \int_E |f_n|^2 d\lambda \leq \frac{1}{4} \int_E g^2 dP = \int_E \left(\frac{1}{2} g\sqrt{p_0} \right)^2 d\lambda < +\infty.$$

By lemma 9.22, this implies that

$$\lim_{n\to\infty} \int_E \left(f_n - \frac{1}{2} g\sqrt{p_0} \right)^2 d\lambda = 0,$$

and since this holds for any $t_n \searrow 0$, we can conclude that

$$\lim_{t\to 0} \int_E \left(\frac{\sqrt{p(t,\cdot)} - \sqrt{p_0}}{t} - \frac{1}{2} g\sqrt{p_0} \right)^2 d\lambda = 0.$$

Since this holds for any bounded $g \in W^\perp$, we have

$$W^\perp \cap B(E,\mathbb{R}) \subset \dot{\mathcal{P}}_P,$$

where $B(E,\mathbb{R})$ denotes the set of all bounded real valued functions on E. So far, we have shown that

$$W^\perp \cap B(E,\mathbb{R}) \subset \dot{\mathcal{P}}_P \subset \mathrm{sp}(\dot{\mathcal{P}}_P) \subset W^\perp.$$

For any sequence $\{g_n\}_{n\in N_+} \in W^\perp$ that converges to some $g \in L^2(\mathcal{E}, P)$ in the L^2-norm,

$$\int_E m_i g \, dP = \lim_{n\to\infty} \int_E m_i g_n dP = 0$$

for each $1 \leq i \leq k$ by the continuity of the inner product on $L^2(\mathcal{E}, P)$. This proves that $g \in W^\perp$ as well, and as such that $W^\perp$ is a closed subset of $L^2(\mathcal{E}, P)$. As such,

$$\overline{W^\perp \cap B(E,\mathbb{R})} \subset W^\perp.$$

To establish the reverse inclusion, note first that $B(E,\mathbb{R})$ is dense in $L^2(\mathcal{E}, P)$. To see

this, let $g \in L^2(\mathcal{E}, P)$ and define $\{g_n\}_{n \in N_+}$ as $g_n = g \cdot I_{\{|g| \leq n\}}$ for each $n \in N_+$. Then,

$$\lim_{n \to \infty} \int_E |g - g_n|^2 dP = \lim_{n \to \infty} \int_{\{|g|^2 > n^2\}} |g|^2 dP = 0,$$

where the final equality follows from the fact that the singleton $\{g^2\}$ is uniformly integrable.

Now choose any $g \in W^\perp$ and define the function $\tilde{m} : E \to \mathbb{R}^{k+1}$ as $\tilde{m}(\cdot) = (1, m(\cdot))$. Since g and each of the components of $\tilde{m}$ are contained in $L^2(\mathcal{E}, P)$, from what we just showed there exists a sequence $\{g_n\}_{n \in N_+}$ of bounded functions on E and a sequence $\{\tilde{m}_n\}_{n \in N_+}$ of bounded k-dimensional functions on E such that

$$\lim_{n \to \infty} \int_E |g_n - g|^2 dP = 0$$

$$\lim_{n \to \infty} \int_E |\tilde{m}_n - \tilde{m}|^2 dP = 0.$$

By the continuity of the inner product on $L^2(\mathcal{E}, P)$, we have

$$\lim_{n \to \infty} \int_E \tilde{m} \tilde{m}_n' dP = \int_E \tilde{m} \tilde{m}' dP$$

$$\lim_{n \to \infty} \int_E \tilde{m} g_n dP = \int_E \tilde{m} g dP = O_{(k+1) \times 1},$$

where the last equality follows because g is orthogonal to the elements of $\tilde{m}$. Since $\int_E \tilde{m} \tilde{m}' dP$ is positive definite by assumption, it follows that $\int_E \tilde{m} \tilde{m}_n' dP$ is also positive definite for large enough n.

For n large enough in the sense above, define

$$\beta_n = \left(\int_E \tilde{m} \tilde{m}_n' dP \right)^{-1} \left(\int_E \tilde{m} g_n dP \right).$$

Then, $\beta_n \to O_{(k+1) \times 1}$ and

$$\begin{aligned} \int_E \tilde{m} \left(g_n - \beta_n' \tilde{m}_n \right) dP &= \int_E \tilde{m} g_n dP - \left(\int_E \tilde{m} \tilde{m}_n' dP \right) \beta_n \\ &= \int_E \tilde{m} g_n dP - \int_E \tilde{m} g_n dP = O_{(k+1) \times 1} \end{aligned}$$

for any $n \in N_+$. It follows that $\{g_n - \beta_n' \tilde{m}_n\}_{n \in N_+}$ is a sequence of bounded functions in $W^\perp$ such that

$$\left(\int_E |g_n - \beta_n' \tilde{m}_n - g|^2 dP \right)^{\frac{1}{2}} \leq \left(\int_E |g_n - g|^2 dP \right)^{\frac{1}{2}} + \left(\int_E |\beta_n' \tilde{m}_n|^2 dP \right)^{\frac{1}{2}}$$

(Minkowski's Inequality)

$$\leq \left(\int_E |g_n - g|^2 dP \right)^{\frac{1}{2}} + |\beta_n| \left(\int_E |\tilde{m}_n|^2 dP \right)^{\frac{1}{2}}$$

for each n. Since

$$\left(\int_E |g_n - g|^2 dP \right)^{\frac{1}{2}} + |\beta_n| \left(\int_E |\tilde{m}_n|^2 dP \right)^{\frac{1}{2}} \to 0 + \left| O_{(k+1)\times 1} \right| \cdot \left(\int_E |\tilde{m}|^2 dP \right)^{\frac{1}{2}} = 0$$

as $n \to \infty$, it follows that

$$g_n - \beta_n' \tilde{m}_n \overset{L^2(\mathcal{E},P)}{\to} g.$$

In other words, g is a limit point of $W^\perp \cap B(E, \mathbb{R})$, and since this holds for any $g \in W^\perp$, we have

$$W^\perp = \overline{W^\perp \cap B(E, \mathbb{R})}.$$

It follows that

$$W^\perp = \overline{W^\perp \cap B(E, \mathbb{R})} \subset \overline{\mathrm{sp}(\dot{\mathcal{P}}_P)} \subset W^\perp,$$

so that

$$\overline{\mathrm{sp}(\dot{\mathcal{P}}_P)} = W^\perp,$$

as we desired.

Q.E.D.

9.4.6 Application: Asymptotic Efficiency for Semiparametric Models

We now move onto the case of semiparametric models, which are models with infinite-dimensional parameters where we are interested only in a finite subset of the parameters. This affords additional structure to the problem at hand, and we are able to say much more about the efficient influence function and the optimal limiting distribution compared to completely non-parametric models.

The basic statistical framework is the same as the one introduced in the previous section. Let $(E, \mathcal{E})$ be a measurable space, $\mathcal{P}$ a collection of probability measures on $(E, \mathcal{E})$ that represents the statistical model, and $P \in \mathcal{P}$ the true distribution of the data. Now we assume that each probability measure on $\mathcal{P}$ can be indexed with a parameter lying in the parameter space $\Theta = \mathbb{R}^k \times H$, where H is some infinite-dimensional set. Letting $(\theta_0, \eta_0) \in \Theta$ be the true parameters, the statistical model $\mathcal{P}$ can be written as

$$\mathcal{P} = \{P_{\theta,\eta} \mid \theta \in \mathbb{R}^k, \quad \eta \in H\} \subset \mathcal{M}_+(\mathcal{E})$$

and $P = P_{\theta_0,\eta_0}$. We are interested in the finite-dimensional parameter θ; formally, we want to conduct inference for $\psi(P)$, where the function $\psi : \mathcal{P} \to \mathbb{R}^k$ is defined as

$$\psi(P_{\theta,\eta}) = \theta$$

for any $(\theta, \eta) \in \Theta$. This is why we call η the nuisance parameter, and H the nuisance parameter space.

As before, we start by characterizing the tangent set $\dot{\mathcal{P}}_P$ of $\mathcal{P}$ at P, and by extension, the efficient influence function of P relative to $\dot{\mathcal{P}}_P$. To determine which subset of the maximal tangent set to work with, we first study the case where H is also finite-dimensional; put $H = \mathbb{R}^d$. In this case, the model reduces to a parametric model, and the one-dimensional submodels μ of $\mathcal{P}$ are of the form

$$\mu(t, \cdot) = P_{\theta_0 + th_1,\, \eta_0 + th_2}$$

for any $t \geq 0$ and $h_1 \in \mathbb{R}^k, h_2 \in \mathbb{R}^d$. Assuming that $p_0 : \mathbb{R}^{k+d} \to \mathbb{R}_+$ is the density with respect to some measure λ and g_0 is the score function corresponding to this model, we saw above for the application to parametric models that μ satisfies the DQM property with score function equal to

$$h_1' g_{0,1} + h_2' g_{0,2},$$

where $g_{0,1}$ collects the first k coordinates of g_0 and $g_{0,2}$ the rest. Here, $h_1' g_{0,1}$ is the score function of the submodel

$$\mu_1(t, \cdot) = P_{\theta_0 + th_1,\, \eta_0}$$

and $h_2' g_{0,2}$ that of

$$\mu_2(t,\cdot) = P_{\theta_0,\, \eta_0 + t h_2}.$$

In other words, the score function of μ can be expressed as the sum of the score functions for the "partial" submodels μ_1 and μ_2 that correspond to keeping either θ or η fixed at their true values.

We can therefore generalize this approach to accommodate a wider variety of semiparametric models. Denote the submodel of $\mathcal{P}$ corresponding to the nuisance parameter by

$$\eta\mathcal{P}_{\theta_0} = \{P_{\theta_0,\eta} \mid \eta \in H\} \subset \mathcal{P},$$

and let $\eta\dot{\mathcal{P}}_P$ be a tangent set of $\mathcal{P}_{\theta_0}$ at P.

Similarly, let $\theta\mathcal{P}_{\eta_0}$ be the parametric submodel of $\mathcal{P}$ that corresponds to the parameter of interest; assuming that there exists a transition probability kernel Λ from $(E,\mathcal{E})$ into $\mathbb{R}^k$ such that $\Lambda(\theta,\cdot) = P_{\theta,\eta_0}$ for any $\theta \in \mathbb{R}^k$, we put

$$\theta\mathcal{P}_{\eta_0} = \{\Lambda(\theta,\cdot) \mid \theta \in \mathbb{R}^k\} \subset \mathcal{P}.$$

Since Λ is simply a parametric model, we can now assume that it satisfies the DQM property with score function $\dot{l}_{\theta,\eta} : E \to \mathbb{R}^k$ and dominating measure λ:

$$\lim_{|h|\to 0} \frac{1}{|h|^2} \int_E \left(\sqrt{d\Lambda(\theta_0 + h,\cdot)} - \sqrt{dP} - \frac{1}{2} h' \dot{l}_{\theta,\eta} \sqrt{dP} \right)^2 d\lambda = 0,$$

and in particular

$$\lim_{t\to 0} \int_E \left(\frac{\sqrt{d\Lambda(\theta_0 + th,\cdot)} - \sqrt{dP}}{t} - \frac{1}{2} h' \dot{l}_{\theta,\eta} \sqrt{dP} \right)^2 d\lambda = 0$$

for any $h \in \mathbb{R}^k$. The maximal tangent set for this submodel is given as

$$\theta\dot{\mathcal{P}}_P = \{h' \dot{l}_{\theta,\eta} \mid h \in \mathbb{R}^k\}.$$

We now define $\dot{\mathcal{P}}_P$ as

$$\begin{aligned} \dot{\mathcal{P}}_P &= \theta\dot{\mathcal{P}}_P + \eta\dot{\mathcal{P}}_P \\ &= \{h' \dot{l}_{\theta,\eta} + g \mid h \in \mathbb{R}^k,\ g \in \eta\dot{\mathcal{P}}_P\}. \end{aligned}$$

The following theorem furnishes sufficient conditions for $\dot{\mathcal{P}}_P$ to be a well-defined tangent set for $\mathcal{P}$ at P, and derives the corresponding efficient influence function for estimating $\psi(P)$ relative to $\dot{\mathcal{P}}_P$.

Theorem 9.25 (Efficient Influence Function for Semiparametric Models)
Let $(E,\mathcal{E})$ be a measurable space, and P a probability measure in the statistical model $\mathcal{P} \subset \mathcal{M}_+(\mathcal{E})$. Let $\Theta = \mathbb{R}^k \times H$, where H is the nuisance parameter space, and assume that $\mathcal{P}$ is a semiparametric model indexed by the parameters in Θ as follows:

$$\mathcal{P} = \{P_{\theta,\eta} \mid \theta \in \mathbb{R}^k, \eta \in H\}.$$

Suppose $P = P_{\theta_0,\eta_0}$ for some $(\theta_0,\eta_0) \in \Theta$. Let $\eta\mathcal{P}_{\theta_0}$ be the submodel of $\mathcal{P}$ corresponding to the nuisance parameter:

$$\eta\mathcal{P}_{\theta_0} = \{P_{\theta_0,\eta} \mid \eta \in H\} \subset \mathcal{P},$$

and let $\eta\dot{\mathcal{P}}_P$ be the tangent set of $\eta\mathcal{P}_{\theta_0}$ at P. Letting Π_0 be the projection mapping onto the closed convex subset $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}$ of $L^2(\mathcal{E},P)$, define

$$\tilde{l}_{\theta,\eta} = \dot{l}_{\theta,\eta} - \Pi_0(\dot{l}_{\theta,\eta})$$

and $\tilde{I}_0 = \int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP$, where the mapping $\Pi_0(\cdot)$ is applied elementwise. We make the following assumptions:

1) **(Baseline Score Function for θ)**
There exists a transition probability kernel Λ from $\mathbb{R}^k$ into $(E,\mathcal{E})$ such that $\Lambda(\theta,\cdot) = P_{\theta,\eta_0}$ for any $\theta \in \mathbb{R}^k$, and Λ satisfies the DQM property with score function $\dot{l}_{\theta,\eta} : E \to \mathbb{R}^k$ and dominating σ-finite measure λ.

2) **(Separability of θ and Nuisance Parameter)**
For any $h \in \mathbb{R}^k$ and $g \in \eta\dot{\mathcal{P}}_P$, there exists a one-dimensional path $t \mapsto \eta_t$ in H such that the one-dimensional submodel μ of $\mathcal{P}$ defined as

$$\mu(t,\cdot) = P_{\theta_0+th,\ \eta_t}$$

for any $t \geq 0$ satisfies the DQM property with score function $h'\dot{l}_{\theta,\eta} + g$.

3) **(Nonsingular Information Matrix)**
$\tilde{I}_0$ is a positive definite $k \times k$ matrix.

Then, the collection $\dot{\mathcal{P}}_P$ of functions in $L^2(\mathcal{E},P)$ defined as

$$\dot{\mathcal{P}}_P = \{h'\dot{l}_{\theta,\eta} + g \mid h \in \mathbb{R}^k, g \in \eta\dot{\mathcal{P}}_P\}$$

is a tangent set of $\mathcal{P}$ at P, and the function $\psi : \mathcal{P} \to \mathbb{R}^k$ defined as

$$\psi(P_{\theta,\eta}) = \theta$$

for any $(\theta, \eta) \in \Theta$ is differentiable at P relative to $\dot{\mathcal{P}}_P$ with efficient influence function

$$\tilde{\psi}_P(\cdot) = \tilde{I}_0^{-1} \tilde{l}_{\theta,\eta}(\cdot).$$

Proof) By assumption 2, $\dot{\mathcal{P}}_P = \theta \dot{\mathcal{P}}_P + \eta \dot{\mathcal{P}}_P$ is a well-defined tangent set of $\mathcal{P}$ at P. We can thus write

$$\dot{\mathcal{P}}_P = \text{sp}(\dot{l}_{\theta,\eta}) + \eta \dot{\mathcal{P}}_P,$$

where $\text{sp}(\dot{l}_{\theta,\eta})$ denotes the linear subspace of $L^2(\mathcal{E}, P)$ spanned by the k components of $\dot{l}_{\theta,\eta}$.

To establish the differentiability of $\psi(\cdot)$ at P relative to $\dot{\mathcal{P}}_P$, we need to show that there exists some function $\tilde{\psi}_P = (\tilde{\psi}_{P,1}, \cdots, \tilde{\psi}_{P,k})$ such that:

a) For each $1 \leq i \leq k$, $\tilde{\psi}_{P,i} \in \overline{\text{sp}(\dot{\mathcal{P}}_P)}$

b) For any $h \in \mathbb{R}^k$ and $g \in \eta \dot{\mathcal{P}}_P$, there exists some one-dimensional submodel μ with score function $h' \dot{l}_{\theta,\eta} + g$ satisfying

$$\lim_{t \to 0} \frac{\psi(\mu(t, \cdot)) - \psi(P)}{t} = \int_E \tilde{\psi}_P (h' \dot{l}_{\theta,\eta} + g) dP.$$

To this end, we choose $h \in \mathbb{R}^k$ and $g \in \eta \dot{\mathcal{P}}_P$, and we will show that the above relationship holds for the separable one-dimensional submodel μ with score function $h' \dot{l}_{\theta,\eta} + g$, that is, the submodel μ defined as

$$\mu(t, \cdot) = P_{\theta_0 + th, \eta_t}$$

for any $t \geq 0$ with score function $h' \dot{l}_{\theta,\eta} + g$. By definition, we can see that

$$\frac{\psi(\mu(t, \cdot)) - \psi(P)}{t} = h$$

for any $t \geq 0$, so that

$$\lim_{t \to 0} \frac{\psi(\mu(t, \cdot)) - \psi(P)}{t} = h.$$

It remains to find a function $\tilde{\psi}_P$ whose elements are in $\overline{\text{sp}(\dot{\mathcal{P}}_P)}$ such that

$$\int_E \tilde{\psi}_P(h'\dot{l}_{\theta,\eta} + g)dP = h.$$

For this equation to hold for any $h \in \mathbb{R}^k$ and $g \in \eta\dot{\mathcal{P}}_P$, the function $\tilde{\psi}_P$ must satisfy

$$\int_E \tilde{\psi}_P \dot{l}'_{\theta,\eta} dP = I_k \quad \text{and} \quad \int_E \tilde{\psi}_P g \, dP = O_{k\times 1}$$

for any $g \in \eta\dot{\mathcal{P}}_P$. The second condition tells us that the components of $\tilde{\psi}_P$ must be orthogonal to any function in $\eta\dot{\mathcal{P}}_P$; this will be the case if each component of $\tilde{\psi}_P$ is contained in the orthogonal complement of the closed convex subset $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}$ of $L^2(\mathcal{E}, P)$. One way to construct $\tilde{\psi}_P$ is therefore to choose any k-dimensional function $\varphi : E \to \mathbb{R}^k$ with components in $L^2(\mathcal{E}, P)$ and let $\tilde{\psi}_P$ be its orthogonal projection onto $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}^{\perp}$. Since

$$L^2(\mathcal{E}, P) = \overline{\text{sp}(\eta\dot{\mathcal{P}}_P)} \bigoplus \overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}^{\perp},$$

each component of φ can be decomposed as the unique sum of its projection onto $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}$ and its projection onto $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}^{\perp}$; it follows that

$$\tilde{\psi}_P = \varphi - \Pi_0(\varphi)$$

collects the projections of the elements of φ onto $\overline{\text{sp}(\eta\dot{\mathcal{P}}_P)}^{\perp}$.

We now proceed to narrow down possible candidates for the function φ to find the desired efficient influence function. Note first that the elements of $\tilde{\psi}_P$ must be contained in the closure of the subspace spanned by $\dot{\mathcal{P}}_P = \text{sp}(\dot{l}_{\theta,\eta}) + \eta\dot{\mathcal{P}}_P$. Since

$$\text{sp}(\dot{l}_{\theta,\eta}) + \eta\dot{\mathcal{P}}_P \subset \text{sp}(\dot{l}_{\theta,\eta}) + \text{sp}\left(\eta\dot{\mathcal{P}}_P\right),$$

and the sum of two vector spaces over the real field is itself a vector space over the real field, it follows that

$$\text{sp}\left(\dot{\mathcal{P}}_P\right) \subset \text{sp}(\dot{l}_{\theta,\eta}) + \text{sp}\left(\eta\dot{\mathcal{P}}_P\right).$$

On the other hand, for any $f \in \text{sp}(\dot{l}_{\theta,\eta}) + \text{sp}\left(\eta\dot{\mathcal{P}}_P\right)$, there exists some $h_1 \in \mathbb{R}^k$, $h_2 \in \mathbb{R}^m$ and $G = (g_1, \cdots, g_m) : E \to \mathbb{R}^m$ with components in $\eta\dot{\mathcal{P}}_P$ such that

$$f = h_1'\dot{l}_{\theta,\eta} + h_2'G.$$

Assuming without loss of generality that h_2 does not contain any elements equal to 0,

we can see that

$$f = \sum_{i=1}^{m} h_{2,i} \left(\frac{h_1'}{h_{2,i}} \dot{l}_{\theta,\eta} + g_m \right) \in \mathrm{sp}\left(\dot{\mathcal{P}}_P\right).$$

By implication,

$$\mathrm{sp}(\dot{l}_{\theta,\eta}) + \mathrm{sp}\left(\eta\dot{\mathcal{P}}_P\right) = \mathrm{sp}\left(\dot{\mathcal{P}}_P\right) \subset \overline{\mathrm{sp}(\dot{\mathcal{P}}_P)}.$$

Since the sum of closures is contained in the closure of the sum,

$$\begin{aligned}\mathrm{sp}(\dot{l}_{\theta,\eta}) + \overline{\mathrm{sp}\left(\eta\dot{\mathcal{P}}_P\right)} &\subset \overline{\mathrm{sp}(\dot{l}_{\theta,\eta})} + \overline{\mathrm{sp}\left(\eta\dot{\mathcal{P}}_P\right)} \\ &\subset \overline{\mathrm{sp}(\dot{l}_{\theta,\eta}) + \mathrm{sp}\left(\eta\dot{\mathcal{P}}_P\right)} \subset \overline{\mathrm{sp}(\dot{\mathcal{P}}_P)}.\end{aligned}$$

In other words, if the components of $\tilde{\psi}_P$ are constructed as the sum of the linear combination of elements of $\dot{l}_{\theta,\eta}$ and a function in $\overline{\mathrm{sp}(\eta\dot{\mathcal{P}}_P)}$, then they are also contained in $\overline{\mathrm{sp}(\dot{\mathcal{P}}_P)}$. Thus, if $\varphi = A\dot{l}_{\theta,\eta}$ for some non-random $k \times k$ matrix A, then the components of

$$\tilde{\psi}_P = \varphi - \Pi_0(\varphi) = A\left(\dot{l}_{\theta,\eta} - \Pi_0(\dot{l}_{\theta,\eta})\right) = A\tilde{l}_{\theta,\eta}$$

are contained in $\overline{\mathrm{sp}(\dot{\mathcal{P}}_P)}$, where we used the linearity of the projection operator Π_0 to justify the second equality.

Now it remains to choose A so that the first condition $\int_E \tilde{\psi}_P \dot{l}_{\theta,\eta}' dP = I_k$ is satisfied. A must therefore satisfy

$$A\int_E \tilde{l}_{\theta,\eta}\dot{l}_{\theta,\eta}' dP = I_k.$$

Here, because the components of $\Pi_0(\dot{l}_{\theta,\eta})$ are contained in $\overline{\mathrm{sp}(\eta\dot{\mathcal{P}}_P)}$, while the components of $\tilde{l}_{\theta,\eta}$ are contained in $\overline{\mathrm{sp}(\eta\dot{\mathcal{P}}_P)}^{\perp}$, it follows that

$$\int_E \tilde{l}_{\theta,\eta}\Pi_0(\dot{l}_{\theta,\eta})' dP = O_{k\times k},$$

so that

$$A\int_E \tilde{l}_{\theta,\eta}\dot{l}_{\theta,\eta}' dP = A\int_E \tilde{l}_{\theta,\eta}\tilde{l}_{\theta,\eta}' dP = A\tilde{I}_0.$$

In other words, A must satisfy the equation $A\tilde{I}_0 = I_k$; thus, $A = \tilde{I}_0^{-1}$ and our efficient influence function is given as

$$\tilde{\psi}_P(\cdot) = \tilde{I}_0^{-1}\tilde{l}_{\theta,\eta}(\cdot).$$

Q.E.D.

In light of theorem 9.21 and the preceding theorem, it follows that the sequence of statistics $\{\mathcal{T}_n^*\}_{n\in N_+}$ satisfying

$$\sqrt{n}[\mathcal{T}_n^*(X_1,\cdots,X_n)-\theta_0] = \tilde{I}_0^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{l}_{\theta,\eta}(X_i)\right)+o_p(1)$$

for any $n\in N_+$, where $\{X_i\}_{i\in N_+}$ is a random sample with distribution P, defines a regular statistic relative to $\dot{\mathcal{P}}_P$ such that

$$\sqrt{n}[\mathcal{T}_n^*-\theta_0] \overset{\mu_{h'\dot{l}_{\theta,\eta}+g,n}(1/\sqrt{n},\cdot)}{\Rightarrow} \mathcal{N}\left[O_{k\times 1},\ \tilde{I}_0^{-1}\right]$$

for any $h\in\mathbb{R}^k$, $g\in\eta\dot{\mathcal{P}}_P$ and submodel μ of $\mathcal{P}$ with score function $h'\dot{l}_{\theta,\eta}+g$. This closely mirrors the formula for obtaining asymptotically efficient estimators in parametric models, except here the efficient influence function is given in terms of the residual function $\tilde{l}_{\theta,\eta}=\dot{l}_{\theta,\eta}-\Pi_0(\dot{l}_{\theta,\eta})$ instead of the score function $\dot{l}_{\theta,\eta}$. We refer to $\tilde{l}_{\theta,\eta}$ as the efficient score function, and $\dot{l}_{\theta,\eta}$ as the baseline score.

Heuristically, the use of $\tilde{l}_{\theta,\eta}$ instead of $\dot{l}_{\theta,\eta}$ represents how information is lost in a semiparametric model with nuisance parameter η compared to a parametric model in which the value of η is known. More precisely, the added uncertainty in η means that we must partial out the part of the baseline score $\dot{l}_{\theta,\eta}$ that depends on η; this is why the efficient score $\tilde{l}_{\theta,\eta}$ is constructed by orthogonalizing $\dot{l}_{\theta,\eta}$ to the subspace spanned by the score functions corresponding to η. Mathematically, this results in a larger lower bound for the asymptotic variance of regular estimators: since

$$\int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP = \int_E \tilde{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP - \int_E \tilde{l}_{\theta,\eta}\Pi_0(\dot{l}_{\theta,\eta})'dP = \int_E \tilde{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP$$

and

$$\begin{aligned}\int_E \Pi_0(\dot{l}_{\theta,\eta})\Pi_0(\dot{l}_{\theta,\eta})'dP &= \int_E \Pi_0(\dot{l}_{\theta,\eta})\dot{l}'_{\theta,\eta}dP - \int_E \Pi_0(\dot{l}_{\theta,\eta})\tilde{l}'_{\theta,\eta}dP \\ &= \int_E \Pi_0(\dot{l}_{\theta,\eta})\dot{l}'_{\theta,\eta}dP\end{aligned}$$

because the components of $\Pi_0(\dot{l}_{\theta,\eta})$ and those of $\tilde{l}_{\theta,\eta}$ are orthogonal, we have

$$\begin{aligned}\int_E \dot{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP - \int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP &= \int_E \left(\dot{l}_{\theta,\eta}-\tilde{l}_{\theta,\eta}\right)\dot{l}'_{\theta,\eta}dP \\ &= \int_E \Pi_0(\dot{l}_{\theta,\eta})\dot{l}'_{\theta,\eta}dP = \int_E \Pi_0(\dot{l}_{\theta,\eta})\Pi_0(\dot{l}_{\theta,\eta})'dP.\end{aligned}$$

By implication, $\int_E \dot{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP - \int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP$ is positive semidefinite. If $\tilde{I}_0=\int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP$ is positive

definite, this implies that $\int_E \dot{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP$ is also positive definite and that

$$\left(\int_E \tilde{l}_{\theta,\eta}\tilde{l}'_{\theta,\eta}dP\right)^{-1} - \left(\int_E \dot{l}_{\theta,\eta}\dot{l}'_{\theta,\eta}dP\right)^{-1}$$

is also positive semidefinite. This shows that the CRLB for semiparametric models cannot be lower than that of the nested parametric model where the value of the nuisance parameter is known.

Chapter 10

Empirical Process Theory

10.1 Stochastic Convergence without Measurability

Empirical process theory mostly deals with stochastic convergence of random quantities of interest uniformly across some index set. When this index set is a subset of the real line, or the random quantities are continuous functions of some sort, then we can usually establish that these quantities are measurable using a combination of Berge's maximum theorem and some measurable selection theorem—this is exactly what is used to show that, for instance, the maximum likelihood estimator is a measurable function of the data when the objective function is continuous in the data as well.

However, for general index sets and cases where the quantities of interest need not be continuous, it is often impossible to establish measurability. In this case, we require the use of outer probabilities, which take values on all possible subsets of the sample space, to generalize the notion of convergence in probability and distribution. It can be shown that measurability of the limit is sufficient for the usual properties of stochastic convergence, such as the Portmanteau lemma and the continuous mapping theorem, to hold for stochastic convergence under outer probabilities as well.

The usefulness of this more generalized notion of stochastic convergence can be seen most clearly when it comes to the Glivenko-Cantelli theorem. Recall in the chapter on non-parametric estimation that we spent a great deal of effort to establish that $\sup_{x\in\mathbb{R}^k}\left|\frac{1}{n}\sum_{i=1}^n I_{X_i\leq x} - F(x)\right|$ is a random variable for a random sample $X_1,\cdots,X_n$ in $\mathbb{R}^k$ drawn from the distribution function F. We show below that these concerns are rendered irrelevant when we use outer probabilities to define almost sure convergence—leading to a much simplified proof of the Glivenko-Cantelli theorem.

10.1.1 Outer Measures and Integrals

Let $(E,\mathcal{E},\mu)$ be a measure space. Recall from the Hahn-Kolmogorov extension theorem that the function $\mu^*: 2^E \to [0,+\infty]$ defined as

$$\mu^*(A) = \inf\Big\{ \sum_{n=1}^{\infty} \mu(V_n) \mid \{V_n\}_n \subset \mathcal{E},\ A \subset \bigcup_n V_n \Big\}$$

for any $A \subset E$ is an outer measure on E, and the collection $\mathcal{M}$ of outer measurable sets defined as

$$\mathcal{M} = \{A \subset E \mid \mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c) \quad \forall B \subset E\}$$

is a σ-algebra on E containing $\mathcal{E}$. The restriction of μ^* to $\mathcal{M}$ constitutes a complete measure on $(E,\mathcal{M})$.

Note that an equivalent formulation for the outer measure is

$$\mu^*(A) = \inf\{\mu(V) \mid V \in \mathcal{E},\ A \subset V\} \quad \forall A \subset E.$$

To see the equivalence, note that, for any $V \in \mathcal{E}$ such that $A \subset V$, letting $V_1 = V$ and $V_n = \emptyset$ for any $n \geq 2$, $\{V_n\}_n \subset \mathcal{E}$ and $A \subset \bigcup_n V_n = V$, so that

$$\mu(V) = \sum_{n=1}^{\infty} \mu(V_n) \geq \mu^*(A).$$

This establishes that $\inf_{V \in \mathcal{E}, A \subset V} \mu(V) \geq \mu^*(A)$. For the converse inclusion, choose any $\{V_n\}_n \subset \mathcal{E}$ such that $A \subset \bigcup_n V_n$. Since $\mathcal{E}$ is a σ-algebra, $\bar{V} = \bigcup_n V_n \in \mathcal{E}$ as well, so that

$$\inf_{V \in \mathcal{E}, A \subset V} \mu(V) \leq \mu(\bar{V}) \leq \sum_{n=1}^{\infty} \mu(V_n)$$

by the subadditivity of μ. This show sus that $\inf_{V \in \mathcal{E}, A \subset V} \mu(V) \leq \mu^*(A)$.

Analogously to the definition of the outer measure, we can define the outer integral of a real-valued function $f: E \to \mathbb{R}$ that is not necessarily $\mathcal{E}$-measurable as follows:

$$\int_E f d\mu^* = \inf\Big\{ \int_E g d\mu \mid g \in \mathcal{E}/\mathcal{B}(\mathbb{R}),\ f \leq g,\ \int_E g d\mu \text{ exists} \Big\}.$$

If the set on the right hand side is empty, we adopt the convention that $\int_E f d\mu^* = +\infty$. Note that, if f is $\mathcal{E}$-measurable and μ-integrable, then $\int_E f d\mu^* = \int_E f d\mu$, so that the definition above is a suitable generalization of the usual abstract integral. We can now show that the outer integral possesses the following properties:

Lemma 10.1.1. *Let $(E,\mathcal{E},\mu)$ be a measure space and μ^* the outer measure induced by μ. Then,*

i) ***Monotonicity.*** *For functions $f,g: E \to \mathbb{R}$ such that $f \leq g$, $\int_E f d\mu^* \leq \int_E g d\mu^*$.*

ii) For any $f : E \to \mathbb{R}$ and $a \in [0, +\infty)$,

$$\int_E (af) d\mu^* = a \int_E f d\mu^*.$$

iii) ***Finite Subadditivity.*** *For any $f, g : E \to \mathbb{R}$,*

$$\int_E (f+g) d\mu^* \leq \int_E f d\mu^* + \int_E g d\mu^*.$$

iv) ***Additivity with Measurable Functions.*** *For any $f, g : E \to \mathbb{R}$ such that $g(\cdot)$ is an $\mathcal{E}$-measurable integrable function,*

$$\int_E (f+g) d\mu^* = \int_E f d\mu^* + \int_E g d\mu.$$

v) ***MCT.*** *For any sequence $\{f_n\}_{n \in N_+}$ of non-negative functions on E such that $f_n \nearrow f$,*

$$\lim_{n \to \infty} \int_E f_n d\mu^* = \int_E f d\mu^*.$$

vi) ***Countable Subadditivity.*** *For any sequence $\{f_n\}_{n \in N_+}$ of non-negative functions on E, letting their sum be $f = \sum_{n=1}^{\infty} f_n$, we have*

$$\int_E f d\mu^* \leq \sum_{n=1}^{\infty} \int_E f_n d\mu^*.$$

vii) ***BCT.*** *Suppose μ is a finite measure. For any sequence $\{f_n\}_{n \in N_+}$ of uniformly bounded functions on E that converge pointwise to some $\mathcal{E}$-measurable function $f : E \to \mathbb{R}$ and satisfies the asymptotic measurability condition*

$$\int_E f_n d\mu^* + \int_E (-f_n) d\mu^* \to 0 \quad \text{as } n \to \infty,$$

$\int_E f d\mu$ exists and

$$\lim_{n \to \infty} \int_E f_n d\mu^* = \int_E f d\mu.$$

Proof. i) To show monotonicity, choose any real-valued functions $f, g : E \to \mathbb{R}$ such that $f \leq g$. If $\int_E f d\mu^* = +\infty$, then there does not exist any measurable real valued function h such that $f \leq h$ and $\int_E h d\mu$; and by implication $\int_E g d\mu^* = +\infty$ as well. Now let $\int_E f d\mu^* < +\infty$. Suppose for the sake of contradiction that $\int_E f d\mu^* > \int_E g d\mu^*$. Then, by the definition of the outer integral, there exists some $h \in \mathcal{E}/\mathcal{B}(\mathbb{R})$ such that $g \leq h$ and

$$\int_E g d\mu^* \leq \int_E h d\mu < \int_E f d\mu^*.$$

However, since $f \leq h$ as well, it must be the case that $\int_E f d\mu^* \leq \int_E h d\mu$, a contradiction.

ii) Next, let $f : E \to \mathbb{R}$ and $a \in \mathbb{R}$. If $a = 0$, then it is trivial that $\int_E (af) d\mu^* = a \int_E f d\mu^* = 0$. Suppose that $a > 0$. In the case that $\int_E f d\mu^* = +\infty$, if there exists some measurable h such that $af \leq h$ with finite integral, then $\frac{1}{a}h$ is a measurable function dominating f with finite integral, which contradicts the assumption that $\int_E f d\mu^* = +\infty$. It follows that $\int_E (af) d\mu^* = +\infty = a \int_E f d\mu^*$ in this case.

Thus, let $\int_E f d\mu^* < +\infty$. If $\int_E (af) d\mu^* < a \int_E f d\mu^*$, then by definition there exists some measurable $h : E \to \mathbb{R}$ such that $af \leq h$ and

$$\int_E h d\mu < a \int_E f d\mu^*.$$

Dividing both sides by a yields $\int_E \left(\frac{1}{a}h\right) d\mu < \int_E f d\mu^*$, but since $\frac{1}{a}h$ is a measurable real valued function that majorizes f, we must have $\int_E f d\mu^* \leq \int_E \left(\frac{1}{a}h\right) d\mu$, a contradiction. Thus, it must be the case that

$$a \int_E f d\mu^* \leq \int_E (af) d\mu^*.$$

The same holds with af in place of f and $\frac{1}{a}$ in place of a, which establishes that $\int_E (af) d\mu^* \leq a \int_E f d\mu^*$. This shows us that $\int_E (af) d\mu^* = a \int_E f d\mu^*$.

iii) Let $f, g : E \to \mathbb{R}$ be arbitrary real-valued functions. If either $\int_E f d\mu^*$ or $\int_E g d\mu^*$ are $+\infty$, the result is trivial. Thus, we let both integrals be finite. Let h_f, h_g be measurable real valued functions such that $f \leq h_f$ and $g \leq h_g$, whose integrals exist. Then, $h_f + h_g$ is a measurable real valued function such that $f + g \leq h_f + h_g$ and whose integral with respect to μ exists by the additivity of integration; as such,

$$\int_E (f+g) d\mu^* \leq \int_E (h_f + h_g) d\mu = \int_E h_f d\mu + \int_E h_g d\mu.$$

Taking the infimum on both sides with respect to h_f, h_g then yields the desired subadditivity result.

iv) Now suppose that $g(\cdot)$ is $\mathcal{E}$-measurable and integrable while $f(\cdot)$ may not be measurable. Suppose $\int_E f d\mu^* = +\infty$. Then, if there exists some measurable h such that $f + g \leq h$ and h is integrable, then $f \leq h - g$ for the integrable function $h - g$, which contradicts $\int_E f d\mu^* = +\infty$. Thus, in this case we have

$$\int_E (f+g) d\mu^* = +\infty = \int_E f d\mu^*.$$

Suppose $\int_E f d\mu^* < +\infty$. Then, for any measurable and integrable h such that $f + g \leq h$, it

follows that $h-g$ is a measurable and integrable function that majorizes f. By definition,

$$\int_E f d\mu^* \leq \int_E (h-g)d\mu = \int_E h d\mu - \int_E g d\mu.$$

This holds for any measurable and integrable function h that majorizes $f+g$, so

$$\int_E f d\mu^* \leq \int_E (f+g)d\mu^* - \int_E g d\mu,$$

and rearranging we have $\int_E f d\mu^* + \int_E g d\mu \leq \int_E (f+g)d\mu^*$. The reverse inequality was established above, so we immediately have our desired result.

v) Let $\{f_n\}_{n\in N_+}$ be a sequence of non-negative functions on E such that $f_n \nearrow f$. Since $\int_E f_n d\mu^* \leq \int_E f d\mu^*$ by the monotonicity of the outer integral, and $\int_E f_n d\mu^*$ is increasing in n, its limit is well-defined as the supremum and we have

$$\lim_{n\to\infty} \int_E f_n d\mu^* \leq \int_E f d\mu^*.$$

If the limit on the left hand side is $+\infty$, then the result follows trivially, so let it be finite. For any $\epsilon > 0$, since $\lim_{n\to\infty} \int_E f_n d\mu^* = \sup_{n\in N_+} \int_E f_n d\mu^*$ we have

$$\int_E f_n d\mu^* < \lim_{n\to\infty} \int_E f_n d\mu^* + \epsilon \quad \forall n \in N_+.$$

By definition, for each $n \in N_+$ there exists some measurable non-negative valued function g_n such that $f_n \leq g_n$ and

$$\int_E g_n d\mu < \lim_{n\to\infty} \int_E f_n d\mu^* + \epsilon.$$

For any $k \in N_+$, define $G_k = \inf_{n\geq k} g_n$. Then, $\{G_k\}_{k\in N_+}$ is an increasing sequence of non-negative $\mathcal{E}$-measurable functions with limit/supremum G that is also $\mathcal{E}$-measurable. Furthermore, since

$$f_k \leq f_n \leq g_n \quad \forall n \geq k,$$

it follows that $f_k \leq \inf_{n\geq k} g_n = G_k$. Taking $k \to \infty$ on both sides shows us that $f \leq G$ pointwise on E. Furthermore,

$$\int_E G_k d\mu \leq \inf_{n\geq k} \int_E g_n d\mu \leq \lim_{n\to\infty} \int_E f_n d\mu^* + \epsilon$$

by the monotonicity of μ, and by the MCT for ordinary integrals, taking $k \to \infty$ yields

$$\int_E G d\mu \leq \lim_{n\to\infty} \int_E f_n d\mu^* + \epsilon < +\infty.$$

Thus, G is a non-negative measurable function such that $f \leq G$ with a finite integral; by

the definition of the outer integral, $\int_E f d\mu^* \leq \int_E G d\mu$. It follows that

$$\int_E f d\mu^* \leq \lim_{n\to\infty} \int_E f_n d\mu^* + \epsilon.$$

This in turn holds for any $\epsilon > 0$, so

$$\int_E f d\mu^* \leq \lim_{n\to\infty} \int_E f_n d\mu^*,$$

and the proof is complete.

vi) For countable subadditivity, note that, by finite subadditivity,

$$\int_E \left(\sum_{m=1}^{n} f_m \right) d\mu^* \leq \sum_{m=1}^{n} \int_E f_m d\mu^*$$

for any $n \in N_+$. Defining $g_n = \sum_{m=1}^{n} f_m$, $\{g_n\}_{n\in N_+}$ is a sequence of functions on E that increase to f, and as such, the MCT implies

$$\int_E f d\mu^* = \lim_{n\to\infty} \int_E g_n d\mu^* \leq \sum_{m=1}^{\infty} \int_E f_m d\mu^*.$$

vii) Suppose that μ is a finite measure. Let $\{f_n\}_{n\in N_+}$ be a sequence of real-valued functions on E such that $|f_n(x)| \leq m$ for any $n \in N_+$ and $x \in E$. Since $f_n \to f$ pointwise, It follows that $|f(x)| \leq m$ for any $x \in E$ as well, so that $\int_E f d\mu$ exists. Define $g_n = \inf_{k\geq n} f_k$ for any $n \in N_+$; then, $\{g_n + m\}_{n\in N_+}$ is a sequence of non-negative functions on E that increases to $f + m$. By the MCT,

$$\lim_{n\to\infty} \int_E (g_n + m) d\mu^* = \int_E f d\mu + m \cdot \mu(E).$$

Similarly, define $h_n = \sup_{k\geq n} f_k$ for any $n \in N_+$. Then, $\{f - h_n + 2m\}_{n\in N_+}$ is a sequence of non-negative functions that increases to $2m \geq 0$. By the MCT again,

$$\lim_{n\to\infty} \int_E (f - h_n + 2m) d\mu^* = 2m \cdot \mu(E).$$

Now, since the constant function is always measurable, the additivity condition with measurable functions tells us that

$$\int_E (g_n + m) d\mu^* = \int_E g_n d\mu^* + m \cdot \mu(E)$$

for any $n \in N_+$. Likewise, the measurability of $f + 2m$ implies that

$$\int_E (f - h_n + 2m) d\mu^* = \int_E (-h_n) d\mu^* + \int_E f d\mu + 2m \cdot \mu(E).$$

It follows that

$$\lim_{n\to\infty}\int_E g_n d\mu^* = \int_E f d\mu = -\lim_{n\to\infty}\int_E (-h_n) d\mu^*.$$

Since $g_n \leq f_n \leq h_n$ for any $n \in N_+$, the monotonicity of outer integration shows us that

$$\int_E f d\mu = \lim_{n\to\infty}\int_E g_n d\mu^* \leq \liminf_{n\to\infty}\int_E f_n d\mu^* \leq \limsup_{n\to\infty}\int_E f_n d\mu^*.$$

By the asymptotic measurability condition,

$$\limsup_{n\to\infty}\int_E f_n d\mu^* \leq \limsup_{n\to\infty}\left(-\int_E (-f_n) d\mu^*\right) \leq \lim_{n\to\infty}\left(-\int_E (-h_n) d\mu^*\right) = \int_E f d\mu,$$

and it follows that

$$\lim_{n\to\infty}\int_E f_n d\mu^* = \int_E f d\mu.$$

□

The fact that linearity does not hold for outer integrals makes them much harder to work with. For example, the failure of linearity means that we cannot establish the DCT for outer integrals, and that the outer integral for real- and complex-valued functions cannot be defined through the outer integral for non-negative valued functions.

10.1.2 Stochastic Convergence without Measurability

Let $(\Omega_n, \mathcal{H}_n, \mathbb{P}_n)$ be a sequence of probability spaces and (E, d) a metric space. For each $n \in N_+$, let $\mathbb{P}_n^*$ be the outer measure constructed from $\mathbb{P}_n$. Let $\{X_n\}_{n\in N_+}$ be a sequence of mappings $X_n : \Omega_n \to E$, and X a random variable taking values in E under the probability space $(\Omega, \mathcal{H}, \mathbb{P})$. We define the following forms of convergence:

- **Almost Sure Convergence**
  Suppose $(\Omega_n, \mathcal{H}_n, \mathbb{P}_n) = (\Omega, \mathcal{H}, \mathbb{P})$ for any $n \in N_+$. We say $X_n \overset{a.s.*}{\to} X$ if there exists a sequence $\{\Delta_n\}_{n\in N_+}$ of non-negative functions on Ω_n that are $\mathcal{H}_n$-measurable such that

  $$d(X_n, X) \leq \Delta_n \quad \forall n \in N_+, \qquad \Delta_n \overset{a.s.}{\to} 0.$$

- **Convergence in Probability**
  Suppose $(\Omega_n, \mathcal{H}_n, \mathbb{P}_n) = (\Omega, \mathcal{H}, \mathbb{P})$ for any $n \in N_+$. We say $X_n \overset{p}{\to} X$ if

  $$\lim_{n\to\infty} \mathbb{P}_n^* \left[d(X_n, X) > \epsilon\right] = 0$$

  for any $\epsilon > 0$.

- **Convergence in Distribution**
  $X_n \xrightarrow{d} X$ if, for any $f \in C_b(E)$,

$$\lim_{n\to\infty} \mathbb{E}_n^* [f(X_n)] = \mathbb{E}[f(X)].$$

Note that the definitions are extended naturally for mappings X_n that need not be measurable through the use of the outer probability and expectations, but the limit X is nonetheless taken to be measurable. The definition of convergence in distribution cannot be stated in terms of the distributions of X_n because the lack of measurability means that the pushforward measure is undefined. The next theorem shows us that the Portmanteau theorem for convergence in distribution continues to hold:

Lemma 10.1.2 (Portmanteau Lemma for Outer Expectations)**.** *Let $(\Omega_n, \mathcal{H}_n, \mathbb{P}_n)$ be a sequence of probability spaces, (E,d) a metric space with corresponding Borel σ-algebra $\mathcal{E}$, and $\{X_n\}_{n\in N_+}$ a sequence of mappings $X_n : \Omega_n \to E$. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the limiting probability space and $X : \Omega \to E$ a $\mathcal{H}/\mathcal{E}$-measurable mapping. Then, the following are equivalent:*

i) $X_n \xrightarrow{d} X$.

ii) $\mathbb{E}_n^* [f(X_n)] \to \mathbb{E}[f(X)]$ *for any bounded Lipschitz function* $f : E \to \mathbb{R}$.

iii) For any closed subset F of E,

$$\limsup_{n\to\infty} \mathbb{P}_n^* [X_n \in F] \leq \mathbb{P}(X \in F).$$

iv) For any open subset G of E,

$$\liminf_{n\to\infty} \mathbb{P}_{n,*} [X_n \in G] \geq \mathbb{P}(X \in G).$$

v) For any Borel subset B of E with $\mathbb{P}(X \in \partial B) = 0$,

$$\lim_{n\to\infty} \mathbb{P}_n^* [X_n \in B] = \mathbb{P}(X \in B).$$

Proof. **i) → ii)**

This follows trivially because any bounded Lipschitz function $f : E \to \mathbb{R}$ is automatically a bounded continuous function on E.

ii) → iii)

Suppose that $\mathbb{E}_n^* [f(X_n)] \to \mathbb{E}[f(X)]$ for any bounded Lipschitz function $f : E \to \mathbb{R}$. Let F be an closed subset of E. Then, the distance function $x \mapsto d(x,F)$ is a continuous function on E, so that we can define the sequence $\{F_k\}_{k\in N_+}$ of closed subsets of E as

$$F_k = (d(\cdot, F))^{-1} \left(\left[\frac{1}{k}, +\infty \right) \right)$$

for any $k \in N_+$, where the closedness of F_k follows from the fact that the inverse image of a closed set with respect to a continuous function is closed. Since $F_k \nearrow F^c$ as $k \to \infty$, by the sequential continuity of measures we have

$$\lim_{k\to\infty} \mathbb{P}(X \in F_k) = \mathbb{P}(X \in F^c) = 1 - \mathbb{P}(X \in F).$$

Choose any $\epsilon > 0$, and choose $k \in N_+$ so that

$$|\mathbb{P}(X \in F_k^c) - \mathbb{P}(X \in F)| < \frac{\epsilon}{2}.$$

Defining $V = F_k^c$, V is an open subset of E that contains F, and such that $d(x,y) \geq \frac{1}{k}$ for any $x \notin V$ and $y \in F$. It follows from Urysohn's lemma for metric spaces that there exists a Lipschitz continuous function $f : E \to \mathbb{R}$ such that $I_F \leq f \leq I_V$. Thus, $f(\cdot)$ is also bounded, and we can apply ii) to see that there exists some $N \in N_+$ such that

$$|\mathbb{E}_n^*[f(X_n)] - \mathbb{E}[f(X)]| < \frac{\epsilon}{2} \quad \forall n \geq N.$$

Now, by the monotonicity of (outer) measures,

$$\begin{aligned}
\mathbb{P}_n^*[X_n \in F] = \mathbb{E}_n^*[I_F \circ X_n] &\leq \mathbb{E}_n^*[f(X_n)] \\
&\leq \frac{\epsilon}{2} + \mathbb{E}[f(X)] \leq \frac{\epsilon}{2} + \mathbb{E}[I_V \circ X] \\
&\leq \epsilon + \mathbb{P}(X \in F)
\end{aligned}$$

for any $n \geq N$. It follows that

$$\limsup_{n\to\infty} \mathbb{P}_n^*[X_n \in F] \leq \mathbb{P}(X \in F) + \epsilon,$$

and since this holds for any $\epsilon > 0$, we have the desired result.

iii) $\to$ iv)

Choose any open subset G of E. Then, letting $F = G^c$, F is a closed subset of E. Choose any $H \in \mathcal{H}_n$ such that $\{X_n \in F\} \subset H$. Then, $H^c \subset \{X_n \in G\}$, so that

$$\mathbb{P}_{n,*}[X_n \in G] \geq \mathbb{P}_n(H^c) = 1 - \mathbb{P}_n(H),$$

and rearranging we have

$$\mathbb{P}_n(H) \geq 1 - \mathbb{P}_{n,*}[X_n \in G].$$

Taking infima on both sides over $H \in \mathcal{H}_n$ such that $\{X_n \in F\} \subset H$, we have

$$\mathbb{P}_n^*[X_n \in F] \geq 1 - \mathbb{P}_{n,*}[X_n \in G].$$

This holds for any $n \in N_+$, so

$$\limsup_{n\to\infty} (1 - \mathbb{P}_{n,*}[X_n \in G]) \leq \mathbb{P}(X \in F) = 1 - \mathbb{P}(X \in G).$$

Rearranging terms yields the desired result.

iii) → v)

We saw above that iii) implies iv). Choose any subset B of E with $\mathbb{P}(X \in \partial B) = 0$. Then, letting G be the interior of B and F its closure, $\mathbb{P}(X \in G) = \mathbb{P}(X \in F) = \mathbb{P}(X \in B)$. By iii) and iv), we can see that

$$\liminf_{n\to\infty} \mathbb{P}_{n,*}[X_n \in G] \geq \mathbb{P}(X \in G) = \mathbb{P}(X \in F) \geq \limsup_{n\to\infty} \mathbb{P}_n^*[X_n \in F].$$

Since $G \subset B \subset F$, it follows that

$$\liminf_{n\to\infty} \mathbb{P}_n^*[X_n \in B] \geq \mathbb{P}(X \in B) \geq \limsup_{n\to\infty} \mathbb{P}_n^*[X_n \in B],$$

and we have the desired result.

v) → i)

Let $f : E \to \mathbb{R}$ be a bounded and continuous function on E. Suppose that $|f(x)| \leq M$ for any $x \in E$, and define $h(x) = \frac{1}{2M}(f(x) + M)$, so that $h(\cdot)$ is a non-negative continuous function bounded above by 1. For any $t \geq 0$, we showed in the proof of the ordinary Portmanteau lemma that $B_t = \{h > t\}$ is a Borel set such that $\mathbb{P}(X \in \partial B_t) > 0$ only for $t \in T_0$, where T_0 is a countably infinite subset of $[0,1]$. As such, by v),

$$\mathbb{P}_n^*[X_n \in B_t] \to \mathbb{P}(X \in B_t) \quad \forall t \in [0,1] \setminus T_0.$$

Note that the real-valued mapping $t \mapsto \mathbb{P}_n^*[X_n \in B_t]$ is decreasing in t by the monotonicity of the outer measure; as such, it has at most countably many discontinuity points and is Borel measurable. This mapping is also bounded above by 1, which is integrable with respect to the Lebesgue measure on $[0,1]$, so the almost sure version of the DCT tells us that

$$\lim_{n\to\infty} \int_0^1 \mathbb{P}_n^*[X_n \in B_t]\, dt = \int_0^1 \mathbb{P}(X \in B_t)\, dt = \mathbb{E}[h(X)],$$

where the last equality follows from Fubini's theorem.

Now, for any $n \in N_+$, let $Y_n : \Omega_n \to \mathbb{R}$ be $\mathcal{H}_n/\mathcal{B}(\mathbb{R})$ measurable with finite expectation $\mathbb{E}_n[Y_n] < +\infty$ under $\mathbb{P}_n$ and $0 \leq h(X_n) \leq Y_n$. Since $h(X_n) \leq \min(Y_n, 1)$ and

$$\mathbb{E}_n[\min(Y_n, 1)] \leq \mathbb{E}_n[Y_n],$$

it follows that

$$\mathbb{E}_n^*[h(X_n)] = \inf\left\{\mathbb{E}_n[Y_n] \mid Y_n \in \mathcal{H}_n/\mathcal{B}(\mathbb{R}),\ \mathbb{E}_n[Y_n] < +\infty,\ h(X_n) \leq Y_n\right\}$$

$$= \inf\left\{\mathbb{E}_n[Y_n] \mid Y_n \in \mathcal{H}_n/\mathcal{B}(\mathbb{R}),\ \mathbb{E}_n[Y_n] < +\infty,\ h(X_n) \leq Y_n \leq 1\right\},$$

and we can restrict our attention to Y_n bounded above by 1. By Fubini's theorem,

$$\begin{aligned}\mathbb{E}_n[Y_n] &= \int_0^1 \mathbb{P}_n[Y_n > t]\,dt \\ &\geq \int_0^1 \mathbb{P}_n^*[h(X_n) > t]\,dt \\ &= \int_0^1 \mathbb{P}_n^*[X_n \in B_t]\,dt.\end{aligned}$$

As such,

$$\int_0^1 \mathbb{P}_n^*[X_n \in B_t]\,dt \leq \inf\left\{\mathbb{E}_n[Y_n] \mid Y_n \in \mathcal{H}_n/\mathcal{B}(\mathbb{R}),\ \mathbb{E}_n[Y_n] < +\infty,\ h(X_n) \leq Y_n \leq 1\right\} = \mathbb{E}_n^*[h(X_n)].$$

Taking $n \to \infty$ on both sides yields

$$\mathbb{E}[h(X)] \leq \liminf_{n\to\infty} \mathbb{E}_n^*[h(X_n)].$$

To establish the converse, choose any $\epsilon > 0$ and an integer $k \in N_+$ such that $\frac{1}{k} < \epsilon$. Then,

$$h(x) \leq \frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} I_{\{h>j/k\}}(x)$$

for any $x \in E$. The monotonicity and subadditivity of the outer integral shows us that, for any $n \in N_+$,

$$\begin{aligned}\mathbb{E}_n^*[h(X_n)] &\leq \frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k} \mathbb{P}_n^*[h(X_n) > j/k] \\ &\leq \frac{1}{k} + \int_0^1 \mathbb{P}_n^*[h(X_n) > t]\,dt.\end{aligned}$$

Thus, taking $n \to \infty$ on both sides yields

$$\limsup_{n\to\infty} \mathbb{E}_n^*[h(X_n)] \leq \epsilon + \mathbb{E}[h(X)].$$

This in turn holds for any $\epsilon > 0$, so

$$\lim_{n\to\infty} \mathbb{E}_n^*[h(X_n)] = \mathbb{E}[h(X)].$$

Finally, since the outer integral is additive for a sum that includes a measurable function,

$$\mathbb{E}_n^*[h(X_n)] = \frac{1}{2M}\mathbb{E}_n^*[f(X_n)] + \frac{1}{2}$$

for any $n \in N_+$. Likewise, $\mathbb{E}[h(X)] = \frac{1}{2M}\mathbb{E}[f(X)] + \frac{1}{2}$, so it follows that

$$\lim_{n\to\infty} \mathbb{E}_n^*[f(X_n)] = \mathbb{E}[f(X)].$$

□

The usual relationships between the various forms of stochastic convergence continue to hold for outer expectations as well:

Lemma 10.1.3. *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, (E,d) a metric space with corresponding Borel σ-algebra $\mathcal{E}$, and $\{X_n\}_{n\in N_+}, \{Y_n\}_{n\in N_+}$ sequences of mappings $X_nY_n : \Omega \to E$. Let $X, Y : \Omega \to E$ be $\mathcal{H}/\mathcal{E}$-measurable mappings. The following hold true:*

i) $X_n \xrightarrow{a.s.*} X$ implies $X_n \xrightarrow{p} X$.

ii) If $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$.

iii) $X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$.

iv) For a scalar $c \in E$, $X_n \xrightarrow{p} c$ if and only if $X_n \xrightarrow{d} c$.

v) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$ in the box metric $d \times d$.

vi) For a scalar $c \in E$, if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$ in the box metric $d \times d$.

Proof. Suppose that $X_n \xrightarrow{a.s.*} X$. Then, there exists a sequence $\{\Delta_n\}_{n\in N_+}$ of non-negative $\mathcal{H}/\mathcal{B}(\mathbb{R})$ functions such that $d(X_n, X) \leq \Delta_n$ and $\Delta_n \xrightarrow{a.s.} 0$. For any $n \in N_+$ and $\epsilon > 0$, the event $\{d(X_n, X) > \epsilon\}$ is contained in $\{\Delta_n > \epsilon\}$, so that

$$\mathbb{P}_n^*(d(X_n, X) > \epsilon) \leq \mathbb{P}_n(\Delta_n > \epsilon).$$

Because $\Delta_n \xrightarrow{a.s.} 0$, $\Delta_n \xrightarrow{p} 0$ and thus the right hand side is $o(1)$ as $n \to \infty$. This proves that $X_n \xrightarrow{p} X$.

Suppose that $X_n \xrightarrow{d} X$ and $d(X_n, Y_n) \xrightarrow{p} 0$. Choose any bounded Lipschitz function $f : E \to \mathbb{R}$ such that $|f(x)| \leq M$ for any $x \in E$. Letting $L > 0$ be the associated Lipschitz constant, we can see that

$$\begin{aligned} |\mathbb{E}^*[f(Y_n)] - \mathbb{E}^*[f(X_n)]| &\leq |\mathbb{E}^*[f(Y_n) - f(X_n)]| && \text{(Subadditivity)} \\ &\leq \mathbb{E}^*|f(Y_n) - f(X_n)|. && \text{(Monotonicity)} \end{aligned}$$

Since $|f(x)-f(y)| \leq Ld(x,y)$ for any $x,y \in E$, it follows that

$$|f(x)-f(y)| \leq 2M \cdot I_{\{d(x,y)>\epsilon\}} + L\epsilon$$

for any $x,y \in E$ and $\epsilon > 0$. By the monotonicity and subadditivity of the outer integral, we now have

$$|\mathbb{E}^*[f(Y_n)] - \mathbb{E}^*[f(X_n)]| \leq 2M\mathbb{P}^*[d(X_n,Y_n) > \epsilon] + L\epsilon$$

for any $n \in N_+$. Taking $n \to \infty$ yields

$$\limsup_{n\to\infty} |\mathbb{E}^*[f(Y_n)] - \mathbb{E}^*[f(X_n)]| \leq L\epsilon,$$

and since this holds for any $\epsilon > 0$,

$$\mathbb{E}^*[f(Y_n)] - \mathbb{E}^*[f(X_n)] \to 0$$

as $n \to \infty$. By the definition of convergence in distribution, $\mathbb{E}^*[f(X_n)] \to \mathbb{E}[f(X)]$ as $n \to \infty$, so

$$\lim_{n\to\infty} \mathbb{E}^*[f(Y_n)] = \mathbb{E}[f(X)].$$

This in turn holds for any bounded Lipschitz function $f : E \to \mathbb{R}$, so by the Portmanteau lemma, $Y_n \xrightarrow{d} X$.

Suppose $X_n \xrightarrow{p} X$. Then, defining $Y_n = X$ for any $n \in N_+$, $d(X_n,Y_n) \xrightarrow{p} 0$ by the definition of convergence in probability. Since $Y_n \xrightarrow{d} X$ trivially, in light of the preceding result we can conclude that $X_n \xrightarrow{d} X$.

Now suppose $X_n \xrightarrow{d} c$ for some constant $c \in E$. Choose any $\epsilon > 0$ and define the Borel subset $B_\epsilon = \{x \in E \mid d(x,c) > \epsilon\}$ of E. Denote by X the degenerate random variable that assigns unit mass to c, and δ_c the Dirac measure sitting at c. Since $\delta_c(B_\epsilon) = 0$ $(c \notin B_\epsilon)$, it follows that $\mathbb{P}(X \in \partial B_\epsilon) = 0$ as well. By the Portmanteau lemma, we can see that

$$\lim_{n\to\infty} \mathbb{P}^*[X_n \in B_\epsilon] = \mathbb{P}(X \in B_\epsilon) = 0.$$

$X_n \xrightarrow{p} c$ then follows from the fact that $\{X_n \in B_\epsilon\} = \{d(X_n,c) > \epsilon\}$.

Suppose $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ individually. In this case, for any $n \in N_+$,

$$(d \times d)[(X_n,Y_n),\ (X,Y)] = \max(d(X_n,X),\ d(Y_n,Y)),$$

so that taking $n \to \infty$ on both sides implies that $(X_n,Y_n) \xrightarrow{p} (X,Y)$ in the box metric.

Finally, let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for some constant $c \in E$. The proof of the desired result follows

from a procedure similar to that of the proof of ii). First, choose any function $f : E \times E \to \mathbb{R}$ that is bounded and Lipschitz continuous with respect to the metric $d \times d$. Letting $|f(x)| \leq M$ for any $x \in E$, and letting $L > 0$ be the associated Lipschitz constant,

$$|\mathbb{E}^* [f(X_n, Y_n)] - \mathbb{E}^* [f(X_n, c)]| \leq \mathbb{E}^* |f(X_n, Y_n) - f(X_n, c)|.$$

For any $x, y \in E$ and $\epsilon > 0$, we have

$$|f(x, y) - f(x, c)| \leq 2M \cdot I_{\{d(y,c) > \epsilon\}} + L\epsilon,$$

so that

$$|\mathbb{E}^* [f(X_n, Y_n)] - \mathbb{E}^* [f(X_n, c)]| \leq 2M \cdot \mathbb{P}^* [d(Y_n, c) > \epsilon] + L\epsilon.$$

Therefore,

$$\limsup_{T \to \infty} |\mathbb{E}^* [f(X_n, Y_n)] - \mathbb{E}^* [f(X_n, c)]| \leq L\epsilon,$$

and since this holds for any $\epsilon > 0$,

$$\lim_{T \to \infty} |\mathbb{E}^* [f(X_n, Y_n)] - \mathbb{E}^* [f(X_n, c)]| = 0.$$

Now notice how the section $f(\cdot, c) : E \to \mathbb{R}$ is a bounded and continuous function on E. By the definition of convergence in distribution,

$$\mathbb{E}^* [f(X_n, c)] \to \mathbb{E}^* [f(X, c)]$$

as $n \to \infty$. Therefore,

$$\mathbb{E}^* [f(X_n, Y_n)] \to \mathbb{E}^* [f(X, c)]$$

as $n \to \infty$, and we can conclude that $(X_n, Y_n) \xrightarrow{d} (X, c)$.

□

The continuous mapping theorem also continues to hold for the more general forms of stochastic convergence. We present a refinement of the theorem that holds for more general sequences of functions.

Lemma 10.1.4. *Let $(\Omega_n, \mathcal{H}_n, \mathbb{P}_n)$ be a sequence of probability spaces, (E, d) a metric space with corresponding Borel σ-algebra $\mathcal{E}$, and $\{X_n\}_{n \in N_+}$ a sequence of mappings $X_n : \Omega_n \to E$. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the limiting probability space and $X : \Omega \to E$ a $\mathcal{H}/\mathcal{E}$-measurable mapping.*

Let (G, ρ) be another metric space with corresponding Borel σ-algebra $\mathcal{G}$, $\{E_n\}_{n \geq 0}$ subsets of E and $\{g_n\}_{n \geq 0}$ functions such that $g_n : E_n \to G$ for each $n \geq 0$. Suppose, for any sequence $\{x_n\}_{n \in N_+}$ that admits a subsequence converging to some $x \in E_0$, that $g_n(x_n) \to g_0(x)$ along this

subsequence, and that $g_0(X)$ is in $\mathcal{H}/\mathcal{G}$.

Then, $X_n \xrightarrow{d} X$ in the metric d implies $g_n(X_n) \xrightarrow{d} g_0(X)$ in the metric ρ.

Proof. We make use of the portmanteau lemma; choose any closed subset F of G. Then, for any $k \in N_+$ and $n \geq k$,

$$\{x \in E_n \mid g_n(x) \in F\} \subset \bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}.$$

By monotonicity and countable subadditivity,

$$\begin{aligned}\mathbb{P}^*\left[g_n(X_n) \in F\right] &\leq \mathbb{P}^*\left[X_n \in \bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}\right] \\ &\leq \mathbb{P}^*\left[X_n \in \overline{\bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}}\right]\end{aligned}$$

Taking $n \to \infty$ on both sides yields

$$\limsup_{n \to \infty} \mathbb{P}^*\left[g_n(X_n) \in F\right] \leq \limsup_{n \to \infty} \mathbb{P}^*\left[X_n \in \overline{\bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}}\right].$$

By the Portmanteau lemma applied to $X_n \xrightarrow{d} X$, we can see that

$$\limsup_{n \to \infty} \mathbb{P}^*\left[g_n(X_n) \in F\right] \leq \mathbb{P}\left(X \in \overline{\bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}}\right).$$

This holds for any $k \in N_+$.

Define $B_k = \overline{\bigcup_{m \geq k} \{x \in E_m \mid g_m(x) \in F\}}$. Then, $\{X^{-1}(B_k)\}_{k \in N_+}$ is a decreasing sequence of measurable sets in $\mathcal{H}$, so that, by sequential continuity,

$$\limsup_{n \to \infty} \mathbb{P}^*\left[g_n(X_n) \in F\right] \leq \lim_{k \to \infty} \mathbb{P}\left(X \in B_k\right) = \mathbb{P}\left(\bigcap_k X^{-1}(B_k)\right) = \mathbb{P}\left(X \in \bigcap_k B_k\right).$$

Choose any $x \in \bigcap_k B_k$. Then, for any $k \in N_+$, x is a limit point of the union $\bigcup_{m \geq k} \{x' \in E_m \mid g_m(x') \in F\}$, so that there exists some $x' \in E$ such that $g_m(x') \in F$ for at least one $m \geq k$ and $d(x', x) < \frac{1}{k}$. We use this property to construct the desired sequence. Choose any $x' \in E$ such that $d(x', x) < 1$ and $g_{m(1)}(x') \in F$; then, we put $x_i = x'$ for $1 \leq i \leq m(1)$. Suppose that we have now constructed the sequence up to the index $m(k) \in N_+$, where $d(x_{m(k)}, x) < \frac{1}{k}$ and $g_{m(k)}(x_{m(k)}) \in F$. Then, using the fact that $x \in B_{m(k)+1}$, we can find some $x' \in E$ such that $d(x', x) < \frac{1}{k+1}$ and $g_{m(k+1)}(x') \in F$ for $m(k+1) > m(k)$. Now put $x_{m(k)+1} = \cdots = x_{m(k+1)} = x'$.

This construction gives us a sequence $\{x_m\}_{m \in N_+}$ in E and a strictly increasing sequence $\{m(k)\}_{k \in N_+}$ in N_+ such that, for any $k \in N_+$, $d(x_{m(k)}, x) < \frac{1}{k}$ and $g_{m(k)}(x_{m(k)}) \in F$. It follows that $\{x_{m(k)}\}_{k \in N_+}$ is a subsequence that converges to x, and thus, if $x \in E_0$, then $g_{m(k)}(x_{m(k)}) \to g_0(x)$ as $k \to \infty$. Since $\{g_{m(k)}(x_{m(k)})\}_{k \in N_+}$ is a sequence contained in the closed set F, by the sequential

characterization of closed sets $g_0(x) \in F$ as well. We have just proven the inclusion

$$\bigcap_k B_k \subset g_0^{-1}(F) \cup E_0^c,$$

so that

$$\limsup_{n\to\infty} \mathbb{P}^* \left[g_n(X_n) \in F \right] \leq \mathbb{P}\left(X \in g_0^{-1}(F) \cup E_0^c \right).$$

Since $g_0(X)$ is well-defined on the sample space Ω, it follows that X must almost surely take values in E_0, so we have

$$\limsup_{n\to\infty} \mathbb{P}^* \left[g_n(X_n) \in F \right] \leq \mathbb{P}\left(g_0(X) \in F \right).$$

This in turn holds for any closed subset F of G, so the Portmanteau lemma implies that $g_n(X_n) \xrightarrow{d} g_0(X)$.

□

Note that the ordinary continuous mapping theorem follows immediately from the above generalization.

10.2 Convergence of Stochastic Processes

Here we are interested in the convergence of random bounded functions; it can be seen as a generalization of the results pertaining to random functions taking values in the space of continuous functions on a compact metric space. Our interest in the convergence of arbitrary random bounded functions elucidates the reason for focusing on outer expectations in the previous section. The space of continuous functions on compact metric spaces is itself complete and separable under the uniform metric, so that the usual notions of convergence in distribution and probability continue to hold. However, other classes of functions, such as the set of all bounded functions or cadlag functions, on spaces that may not be compact or metrizable, are generally not complete nor separable under the uniform metric. For the space of cadlag functions it is possible to construct the Skorokhod metric as an alternative to the uniform metric, under which completeness and separability are restored; but this construction is very complicated and the theoertical cost is often not worth the benefit. Instead, it is simpler to concede completeness and separability while continuing to work under the uniform metric—in which case the theory of stochastic convergence for non-measurable functions becomes indispensible.

10.2.1 Semimetric Spaces

Let E be an arbitrary set. A function $\rho : E \times E \to \mathbb{R}_+$ is called a semimetric if:

1) $\rho(x, x) = 0$ for any $x \in E$

2) $\rho(x,y) = \rho(y,x)$ for any $x,y \in E$

3) $\rho(x,y) \leq \rho(x,z) + \rho(z,y)$ for any $x,y,z \in E$.

Simply put, a semimetric is a metric under which it is possible to assign distance 0 to two distinct points. The topology τ induced by the semimetric ρ is the topology generated by the base of all open balls under ρ:

$$\mathbb{B} = \Big\{ \{y \in E \mid \rho(x,y) < \epsilon\} \mid x \in E,\ \epsilon > 0 \Big\}.$$

Note that $\mathbb{B}$ is a base on E because any $x \in E$ is contained in an open ball centered at x and, for any two open balls $B_\rho(x,\epsilon)$ and $B_\rho(y,\eta)$ for $x,y \in E$ and $\epsilon,\eta > 0$, if $z \in B_\rho(x,\epsilon) \cap B_\rho(y,\eta)$ then $z \in B_\rho(z,\delta) \subset B_\rho(x,\epsilon) \cap B_\rho(y,\eta)$ for $\delta = \min(\epsilon - \rho(x,z),\ \eta - \rho(y,z)) > 0$.

We say that a function $f : E \to \mathbb{R}$ is continuous on E if it is continuous relative to the topology induced by ρ and the standard euclidean topology on $\mathbb{R}$. It is easy to show that this is equivalent to requiring that, for any $x \in E$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon \quad \forall y \in E \text{ s.t. } \rho(x,y) < \delta.$$

Analogously to the case of metric spaces, we say that f is uniformly continuous on E if

$$\sup_{x,y \in E,\ \rho(x,y) < \delta} |f(x) - f(y)| \to 0$$

as $\delta \to 0$. Uniform continuity plays an especially important role for continuity with respect to semimetrics because it allows us to construct continuous extensions:

Lemma 10.2.1 (Uniformly Continuous Extensions)**.** *Let (E,ρ) be a semimetric space with corresponding topology τ and E_0 a subset of E that is dense in E. If $f : E_0 \to \mathbb{R}$ is uniformly continuous on E_0 with respect to the semimetric ρ, then there exists a unique continuous function $\bar{f} : E \to \mathbb{R}$ such that $\bar{f}(x) = f(x)$ for any $x \in E_0$. Furthermore, $\bar{f}$ is itself uniformly continuous.*

Proof. We start by defining $\bar{f}(x)$ for each $x \in E$. Choose $x \in E$. Because E_0 is dense in E, there exists a sequence $\{x_n\}_{n \in N_+}$ in E_0 such that $\rho(x_n,x) \to 0$ as $n \to \infty$. Choose any $\epsilon > 0$. By uniform continuity there exists a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for any $x,y \in E_0$ such that $\rho(x,y) < \delta$. Since $\{x_n\}_{n \in N_+}$ is Cauchy in ρ, there exists some $N \in N_+$ such that $\rho(x_n,x_m) < \delta$ for any $n,m \geq N$: and as such,

$$|f(x_n) - f(x_m)| < \epsilon \quad \forall n,m \geq N.$$

We can find such an N for any $\epsilon > 0$, so $\{f(x_n)\}_{n \in N_+}$ is a Cauchy sequence in $\mathbb{R}$. The real line with the euclidean metric is a complete metric space, so it follows that there exists some f_x such that $f(x_n) \to f_x$ as $n \to \infty$; we define $\bar{f}(x) = f_x$.

To show uniqueness, let $\{y_n\}_{n\in N_+}$ be another sequence in E_0 that converges to x. Then,

$$\left|f(y_n)-\bar{f}(x)\right| \leq |f(y_n)-f(x_n)| + \left|f(x_n)-\bar{f}(x)\right|,$$

and since $\rho(x_n,y_n) \leq \rho(x_n,x)+\rho(y_n,x)$ converges to 0 as $n\to\infty$, by uniform continuity of $f(\cdot)$ we can see that

$$\lim_{n\to\infty} f(y_n) = \bar{f}(x).$$

Thus, the value of $\bar{f}(x)$ does not depend on which approximating sequence in E_0 we use.

It remains to show that the function $\bar{f}$ defined thusly is uniformly continuous on E. For any $\epsilon>0$ choose $\delta>0$ so that $|f(x)-f(y)|<\frac{\epsilon}{3}$ for any $x,y\in E_0$ such that $\rho(x,y)<\delta$. Now choose any $x,y\in E$ and suppose that $\rho(x,y)<\frac{\delta}{3}$. We want to show that $\left|\bar{f}(x)-\bar{f}(y)\right|<\epsilon$.

Let $\{x_n\}_{n\in N_+}$ and $\{y_n\}_{n\in N_+}$ be sequences in E_0 such that $x_n\to x$ and $y_n\to y$ in ρ. Then, there exists an $N\in N_+$ such that

$$\max\left[\rho(x_n,x),\ \rho(y_n,y)\right] < \frac{\delta}{3} \quad \forall n\geq N.$$

By implication, $\rho(x_n,x_m),\ \rho(y_n,y_m)<\delta$ for any $n,m\geq N$, so that

$$\max\left[|f(x_n)-f(x_m)|,\ |f(y_n)-f(y_m)|\right] < \frac{\epsilon}{3}$$

for any $n,m\geq N$. Taking $m\to\infty$ on both sides of the inequality also shows us that

$$\max\left[\left|f(x_n)-\bar{f}(x)\right|,\ \left|f(y_n)-\bar{f}(y)\right|\right] < \frac{\epsilon}{3}$$

for any $n\geq N$ as well. Finally, we can see that

$$\rho(x_n,y_n) \leq \rho(x_n,x)+\rho(x,y)+\rho(y,y_n) < \delta$$

for any $n\geq N$, so that $|f(x_n)-f(y_n)|<\frac{\epsilon}{3}$. Putting this all together,

$$\left|\bar{f}(x)-\bar{f}(y)\right| \leq |f(x_N)-f(y_N)| + \left|f(x_N)-\bar{f}(x)\right| + \left|f(y_N)-\bar{f}(y)\right| < \epsilon.$$

We can find such a $\delta>0$ for any $\epsilon>0$, so by definition $\bar{f}$ is uniformly continuous on E.

□

10.2.2 The Main Theorem

Let E be an arbitrary set, and define $\ell^\infty(E)$ as the collection of all bounded real-valued functions on E. Then, defining $\|\cdot\|_\infty$ as

$$\|f\|_\infty = \sup_{x\in E}|f(x)| \quad \forall f\in\ell^\infty(E),$$

we know that $\|\cdot\|_\infty$ is a norm on $\ell^\infty(E)$. Let d_∞ be the metric on $\ell^\infty(E)$ induced by $\|\cdot\|_\infty$, so that $(\ell^\infty(E), d_\infty)$ is a well-defined metric space (that may or may not be complete and separable). Let $\mathcal{B}_\infty(E)$ be the Borel σ-algebra corresponding to d_∞.

Given a probability space $(\Omega, \mathcal{H}, \mathbb{P})$, we are interested in functions $X : \Omega \to \ell^\infty(E)$ that may or may not be measurable relative to $\mathcal{H}$ and $\mathcal{B}_\infty(E)$. Given this function, we can construct the corresponding stochastic process $\{X(t)\}_{t\in E}$ as

$$[X(t)](\omega) = [X(\omega)](t) \quad \forall t \in E, \quad \omega \in \Omega,$$

where the left hand side is the function $X(t) : \Omega \to \mathbb{R}$ evaluated at the outcome ω and the right hand side the path $X(\omega)$ of X corresponding to the outcome ω, evaluated at the ordinate $t \in E$. If X is $\mathcal{H}/\mathcal{B}_\infty(E)$ measurable, it is easy to show that each $X(t)$ is Borel measurable.

The following is the statement and proof of the main result. It shows us that, if a sequence $\{X_n\}_{n\in N_+}$ of functions taking values in $\ell^\infty(E)$ has weakly convergent finite-dimensional distributions and satisfies a tightness condition, then it converges in distribution to a random function whose paths are uniformly continuous on some well-defined semimetric space.

Theorem 10.2.1. *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, E an arbitrary set, and $\{X_n\}_{n\in N_+}$ a sequence of functions $X_n : \Omega \to \ell^\infty(E)$ with corresponding stochastic process $\{X_n(t)\}_{t\in E}$. Suppose that*

i) There exists a random function $X : \Omega \to \ell^\infty(E)$ with corresponding stochastic process $\{X(t)\}_{t\in E}$ such that, for any finite collection of points $(t_1, \cdots, t_k) \in E$,

$$(X_n(t_1), \cdots, X_n(t_k)) \xrightarrow{d} (X(t_1), \cdots, X(t_k)).$$

ii) For any $m \in N_+$ there exists a finite partition $E_1^m, \cdots, E_{k_m}^m$ of E such that

$$\limsup_{n\to\infty} \mathbb{P}^* \left[\sup_{1\le j\le k_m} \sup_{s,t\in E_j^m} |X_n(s) - X_n(t)| > 2^{-m} \right] < 2^{-m}.$$

Then, there exists a semimetric ρ on E and a $\mathcal{H}/\mathcal{B}_\infty$-measurable mapping $\bar{X} : \Omega \to \ell^\infty(E)$ with paths that are everywhere uniformly continuous with respect to ρ such that $X_n \xrightarrow{d} \bar{X}$ with respect to the uniform metric d_∞ on $\ell^\infty(E)$.

Suppose in addition that there exists a metric d on E such that we can choose $t_j^m \in E_j^m$ for any $1 \le j \le k_m$ and $m \in N_+$ such that

$$\left(\max_{1\le i\ne j\le k_m} \frac{\left|X(t_j^m) - X(t_i^m)\right|}{d(t_j^m, t_i^m)} \right) \left(\max_{1\le j\le k_m} diam_d(E_j^m) \right) \to 0$$

almost surely as $m \to \infty$. Then, $\bar{X}$ also has paths that are almost surely continuous and bounded with respect to d.

Proof. The proof proceeds roughly as follows. We know that the finite dimensional distributions of X_n coverge in distribution from condition i); the goal is to generalize this to convergence in distribution of X_n itself. Condition ii) allows us to construct finer and finer partitions of the space E, using which we can form a semimetric ρ based on whether two points are contained in the same partition. The axiom of choice then allows us to choose a point on each partition, and ρ will have been designed so that this collection of points E_0 forms a countable dense subset of E. Since the probability that the X_n-ordinates coincide for points in E that are contained in the same set in a given partition becomes smaller the finer the partition becomes, in the limit condition i) implies that the limiting process X has almost surely uniformly continuous paths on E_0 with respect to ρ. The continuous extension lemma proved above allows us to extend X to a process $\bar{X}$ with almost surely uniformly continuous paths on the entire set E with respect to ρ, which also inherits the measurability of X. Now we prove that X_n converges in distribution to $\bar{X}$ by first showing that, at each partition, we can construct a piecewise constant function from X_n which, because it can be reduced to the collection of a finite number of ordinates of X_n, converges in distribution to the same piecewise constant operation applied to $\bar{X}$. Afterward, we show that, as the partition becomes finer, the piecewise constant version of X_n converges to X_n itself in distribution, and likewise for $\bar{X}$.

Constructing the partitions and semimetric

By condition ii), for any $m \in N_+$ we can find a partition $\{E_1^m, \cdots, E_{k_m}^m\}$ of E such that

$$\limsup_{n\to\infty} \mathbb{P}^* \left[\sup_{1\le j\le k_m} \sup_{s,t\in E_j^m} |X_n(s) - X_n(t)| > 2^{-m} \right] < 2^{-m}.$$

Since the left hand side is smaller for finer partitions, we can let $\{E_1^{m+1}, \cdots, E_{k_{m+1}}^{m+1}\}$ be a refinement of $\{E_1^m, \cdots, E_{k_m}^m\}$ for each $m \in N_+$ without loss of generality. For each $m \in N_+$, we can define a semimetric ρ_m on E as

$$\rho_m(x,y) = \begin{cases} 0 & \text{if } x,y \in E_j^m \quad \text{for some } 1 \le j \le k_m \\ 1 & \text{otherwise} \end{cases}$$

for any $x, y \in E$; in other words, $\rho_m(x,y) = 1$ if x, y belong to different partitions and equals 0 otherwise. Note that the ϵ-ball $B_{\rho_m}(x,\epsilon)$ around some $x \in E$ equals the entire set E if $\epsilon \ge 1$ and equals the partition containing x if $\epsilon < 1$. Thus, the topology induced by ρ_m is the topology generated by the base $\{E_1^m, \cdots, E_{k_m}^m, E\}$.

Define the mapping $\rho : E \times E \to \mathbb{R}_+$ as

$$\rho(x,y) = \sum_{m=1}^{\infty} 2^{-m} \rho_m(x,y) \quad \forall x, y \in E.$$

Note that ρ is also a semimetric on E: $\rho(x,x) = 0$ and $\rho(x,y) = \rho(y,x)$ follow easily from the

same properties of ρ_m, and for any $x, y, z \in E$,

$$\rho(x,y) \le \sum_{m=1}^{\infty} 2^{-m} \left(\rho_m(x,z) + \rho_m(z,y)\right) = \rho(x,z) + \rho(z,y).$$

Note also that, for any $x, y \in E_j^m$, since $\rho_k(x,y) = 0$ for any $k \le m$,

$$\rho(x,y) \le \sum_{k=m+1}^{\infty} 2^{-k} = 2^{-m},$$

so that the ρ-diameter of the set E_j^m is bounded above by 2^{-m}. Construct the countable set E_0 as

$$E_0 = \{t_j^m\}_{1 \le j \le k_m,\ m \in N_+}$$

by choosing some t_j^m from each partition E_j^m. We can see that E_0 is dense for E under the semi-metric ρ: for any $x \in E$ and ρ-neighborhood $\mathcal{N}$ of x, by the definition of the topology induced by ρ there exists some $\epsilon > 0$ such that $B_\rho(x, \epsilon) \subset \mathcal{N}$. Now, choosing $m \in N_+$ so that $2^{-m} < \epsilon$, we can see that, if $x \in E_j^m$, then $\rho(x, t_j^m) \le 2^{-m} < \epsilon$, so that $t_j^m \in B_\rho(x, \epsilon) \subset \mathcal{N}$ and thus $t_j^m \in E_0 \cap \mathcal{N}$.

X has uniformly continuous paths on E_0

Since E_0 is countable, it can be enumerated as $E_0 = \{t_k^*\}_{k \in N_+}$. For any $k \in N_+$, define $E_{0,k} = \{t_1^*, \cdots, t_k^*\}$. Then, for any $m, k \in N_+$,

$$\left\{ \sup_{1 \le j \le k_m} \sup_{s,t \in E_j^m \cap E_{0,k}} |X_n(t) - X_n(s)| > 2^{-m} \right\} = \bigcup_{1 \le j \le k_m} \bigcup_{s,t \in E_j^m \cap E_{0,k}} \{|X_n(t) - X_n(s)| > 2^{-m}\},$$

and for any $s, t \in E$, $\{|X_n(t) - X_n(s)| > 2^{-m}\}$ can be expressed as

$$\left\{ (X_n(t), X_n(s)) \in \left(\bigcup_{r \in \mathbb{Q}} (r, +\infty) \times (-\infty, r - 2^{-m}) \right) \cap \left(\bigcup_{r \in \mathbb{Q}} (-\infty, r - 2^{-m}) \times (r, +\infty) \right) \right\}.$$

Since the set $\left(\bigcup_{r \in \mathbb{Q}} (r, +\infty) \times (-\infty, r - 2^{-m})\right) \cap \left(\bigcup_{r \in \mathbb{Q}} (-\infty, r - 2^{-m}) \times (r, +\infty)\right)$ is an open subset of $\mathbb{R}^2$, it follows that there exist some $t_1, \cdots, t_{l(m,k)} \in E$ and an open subset G_m^k of $\mathbb{R}^{l(m,k)}$ such that

$$\left\{ \sup_{1 \le j \le k_m} \sup_{s,t \in E_j^m \cap E_{0,k}} |X_n(t) - X_n(s)| > 2^{-m} \right\} = \{(X_n(t_1), \cdots, X_n(t_{l(m,k)})) \in G_m^k\}.$$

By assumption, $(X_n(t_1), \cdots, X_n(t_{l(m,k)})) \xrightarrow{d} (X(t_1), \cdots, X(t_{l(m,k)}))$, so the Portmanteau lemma implies that

$$\begin{aligned} \mathbb{P}\left((X(t_1), \cdots, X(t_{l(m,k)})) \in G_m^k\right) &\le \liminf_{n \to \infty} \mathbb{P}^* \left[(X_n(t_1), \cdots, X_n(t_{l(m,k)})) \in G_m^k\right] \\ &\le \limsup_{n \to \infty} \mathbb{P}^* \left[\sup_{1 \le j \le k_m} \sup_{s,t \in E_j^m} |X_n(t) - X_n(s)| > 2^{-m} \right] \le 2^{-m}. \end{aligned}$$

The definition of G_m^k now implies that

$$\mathbb{P}\left(\sup_{1\le j\le k_m}\sup_{s,t\in E_j^m\cap E_{0,k}}|X(t)-X(s)|>2^{-m}\right)\le 2^{-m}.$$

This holds for any $k\in N_+$, and defining

$$H_m^k=\Big\{\sup_{1\le j\le k_m}\sup_{s,t\in E_j^m\cap E_{0,k}}|X(t)-X(s)|>2^{-m}\Big\},$$

$\{H_m^k\}_{k\in N_+}$ are measurable sets in $\mathcal{H}$ that increase to

$$\Big\{\sup_{1\le j\le k_m}\sup_{s,t\in E_j^m\cap E_0}|X(t)-X(s)|>2^{-m}\Big\}.$$

By the MCT, taking $k\to\infty$ yields

$$\mathbb{P}\left(\sup_{1\le j\le k_m}\sup_{s,t\in E_j^m\cap E_0}|X(t)-X(s)|>2^{-m}\right)\le 2^{-m}.$$

For any $s,t\in E_0$, suppose that $\rho(s,t)<2^{-m}$. Then, $\rho_m(s,t)<1$ (otherwise, $\rho(s,t)\ge 2^{-m}\rho_m(s,t)=2^{-m}$), so that $s,t\in E_j^m\cap E_0$ for some $1\le j\le k_m$. As such,

$$\mathbb{P}\left(\sup_{s,t\in E_0,\ \rho(s,t)<2^{-m}}|X(t)-X(s)|>2^{-m}\right)\le\mathbb{P}\left(\sup_{1\le j\le k_m}\sup_{s,t\in E_j^m\cap E_0}|X(t)-X(s)|>2^{-m}\right)\le 2^{-m}.$$

Since

$$\sum_{m=1}^{\infty}\mathbb{P}\left(\sup_{s,t\in E_0,\ \rho(s,t)<2^{-m}}|X(t)-X(s)|>2^{-m}\right)\le 1<+\infty,$$

the Borel-Cantelli lemma implies that, for $\mathbb{P}$-almost every $\omega\in\Omega$, there exists some $N\in N_+$ such that

$$\sup_{s,t\in E_0,\ \rho(s,t)<2^{-m}}|[X(\omega)](t)-[X(\omega)](s)|\le 2^{-m}$$

for any $m\ge N$. By implication, for any $\epsilon>0$, choosing $m\ge N$ so that $2^{-m}<\epsilon$, $\sup_{s,t\in E_0,\ \rho(s,t)<\epsilon}|[X(\omega)](t)-[X(\omega)](s)|<\epsilon$, so that the function $X(\omega)$ is uniformly continuous on E_0. This proves that the limiting function X has paths that are almost surely uniformly continuous on E_0 with respect to the semimetric ρ. E_0 is dense in E under ρ.

Constructing measurable uniformly continuous extension

Let $\underline{X}(\omega)$ now represent the paths of X restricted to E_0, so that $\underline{X}:\Omega\to\ell^\infty(E_0)$. Let $d_{\infty,0}$ be the uniform metric on $\ell^\infty(E_0)$ and $\mathcal{B}_{\infty,0}$ the corresponding Borel σ-algebra. $\underline{X}$ is measurable relative to $\mathcal{H}$ and $\mathcal{B}_{\infty,0}$ because the norm $\|\cdot\|_{\infty,0}$ is strictly dominated by $\|\cdot\|_\infty$ and thus the transformation $X\to\underline{X}$ is continuous.

To find a measurable extension of $\underline{X}$, we first show that the subset $UC_b(E_0,\rho)$ of $\ell^\infty(E_0)$ that collects bounded functions on E that are uniformly continuous with respect to ρ is a closed subset of $\ell^\infty(E_0)$. Let x be a limit point of $UC_b(E_0,\rho)$ with respect to the uniform metric $d_{\infty,0}$, so that for any $\epsilon>0$ there exists some $y\in UC_b(E_0,\rho)$ such that $\|x-y\|_{\infty,0}<\epsilon$. Since y is uniformly continuous, there exists some $\delta>0$ such that $|y(t)-y(s)|<\epsilon$ for any $t,s\in E_0$ such that $\rho(t,s)<\delta$; now, for any $t,s\in E_0$ such that $\rho(t,s)<\delta$,

$$|x(t)-x(s)|\leq 2\|x-y\|_{\infty,0}+\|y(t)-y(s)\|<3\epsilon.$$

We can find such a $\delta>0$ for any $\epsilon>0$, so by definition $x\in UC_b(E_0,\rho)$ and thus $UC_b(E_0,\rho)$ is a closed subset of E_0.

Now define the mapping $\mathcal{T}:UC_b(E_0,\rho)\to\ell^\infty(E)$ so that $\mathcal{T}(x)$ is the uniformly continuous extension of x to E defined in the continuous extension lemma. Then, for any $x,y\in UC_b(E_0,\rho)$, the continuity of $\mathcal{T}(x),\mathcal{T}(y)$ as well as the fact that E_0 is ρ-dense in E implies that

$$\sup_{t\in E_0}|x(t)-y(t)|=\sup_{t\in E}|[\mathcal{T}(x)](t)-[\mathcal{T}(y)](t)|.$$

Thus, $\mathcal{T}$ is an isometry and by implication a continuous mapping. We now formally define $\bar{X}:\Omega\to\ell^\infty(E)$ as

$$\bar{X}=\mathcal{T}(X)\cdot I_{\Omega_0},$$

where $\Omega_0\in\mathcal{H}$ is the almost sure event on which X has uniformly continuous paths. It follows that, for any $A\in\mathcal{B}_\infty$,

$$\bar{X}^{-1}(A)=\left[\Omega_0\cap X^{-1}\left(\mathcal{T}^{-1}(A)\right)\right]\cup\left[\Omega_0^c\cap\{\mathbf{0}\in A\}\right],$$

where $\mathbf{0}$ is the zero function on E. All quantities on the right hand side are $\mathcal{H}$-measurable, so by definition $\bar{X}$ is $\mathcal{H}/\mathcal{B}_\infty$-measurable. $\bar{X}$ also has ρ-unifomrly continuous paths everywhere, since $\bar{X}(\omega)$ is ρ-uniformly continuous for $\omega\in\Omega_0$ by the continuous extension lemma and $\bar{X}(\omega)$ is constant for $\omega\in\Omega_0$.

Piecewise constant approximation

Now, for any $m\in N_+$, define the mapping $\pi_m:E\to E$ as

$$\pi_m(x)=t_j^m\quad\text{if } x\in E_j^m$$

for any $x\in E$, and define the mapping $Y_m:\Omega\to\ell^\infty(E)$ as

$$[Y_m(\omega)](t)=[\bar{X}(\omega)](\pi_m(t))\quad\forall\omega\in\Omega,\ t\in E.$$

We now establish that Y_m inherits the $\mathcal{H}/\mathcal{B}_\infty$ measurability of $\bar{X}$. Note that we can write

$$[Y_m(\omega)](t) = \sum_{j=1}^{k_m} [\bar{X}(\omega)](t_j^m) \cdot I_{E_j^m}(t),$$

so that $Y_m = \Phi_m \circ \bar{X}$, where we define the mapping $\Phi_m : \ell^\infty(E) \to \ell^\infty(E)$ is defined as

$$[\Phi_m(x)](t) = \sum_{j=1}^{k_m} x(t_j^m) \cdot I_{E_j^m}(t)$$

for any $x \in \ell^\infty(E)$ and $t \in E$. For $x, y \in \ell^\infty(E)$, we can see that

$$|[\Phi_m(x)](t) - [\Phi_m(y)](t)| \leq \sum_{j=1}^{k_m} \left| x(t_j^m) - y(t_j^m) \right| I_{E_j^m}(t) \leq \|x - y\|_\infty,$$

for any $t \in E$, so that

$$\|\Phi_m(x) - \Phi_m(y)\| \leq \|x - y\|_\infty$$

and Φ_m is a continuous mapping on $\ell^\infty(E)$. The continuity of Φ_m implies that it is Borel measurable, and since measurability is preserved across compositions of continuous functions, Y_m is also $\mathcal{H}/\mathcal{B}_\infty$-measurable.

For any $\omega \in \Omega$, note how

$$\left| [Y_m(\omega)](t) - [\bar{X}(\omega)](t) \right| \leq \sum_{j=1}^{k_m} \left| [\bar{X}(\omega)](t_j^m) - [\bar{X}(\omega)](t) \right| \cdot I_{E_j^m}(t)$$

for any $t \in E$. By uniform continuity, for any $\epsilon > 0$ there exists a $\delta > 0$ for which $|[X(\omega)](t) - [X(\omega)](s)| < \epsilon$ for any $s, t \in E$ such that $\rho(s,t) < \delta$. For large enough m, $2^{-m} < \delta$ and, for each $1 \leq j \leq k_m$ and $t \in E_j^m$, $\rho(t, t_j^m) < 2^{-m} < \delta$, so that

$$\left\| Y_m(\omega) - \bar{X}(\omega) \right\| \leq \epsilon.$$

We can always find an m large enough for any $\epsilon > 0$, so $Y_m(\omega) \to \bar{X}(\omega)$ in the uniform metric d_∞. This establishes that $Y_m \xrightarrow{a.s.} \bar{X}$ (this convergence actually holds everywhere), and by implication $Y_m \xrightarrow{d} \bar{X}$ as well.

Now, for each fixed $m \in N_+$ and $\omega \in \Omega$, $Y_m(\omega)$ is a piecewise constant function taking finitely many values $\{X(t_1^m), \cdots, X(t_{k_m}^m)\}$. Similarly, defining $Y_m^n : \Omega \to \ell^\infty(E)$ as

$$Y_m^n = \Phi_m \circ X_n$$

for any $n \in N_+$, $Y_m^n(\omega)$ is a piecewise constant function taking finitely many values $\{X_n(t_1^m), \cdots, X_n(t_{k_m}^m)\}$.

More precisely, defining the mapping $\phi_m : \mathbb{R}^{k_m} \to \ell^\infty(E)$ as

$$[\phi_m(v)](t) = \sum_{j=1}^{k_m} v_j \cdot I_{E_j^m}(t) \quad \forall t \in E,$$

we have $Y_m^n = \phi_m \circ (X_n(t_1^m), \cdots, X_n(t_{k_m}^m))$ for any $n \in N_+$. Note that ϕ_m is a continuous mapping:

$$|[\phi_m(v)](t) - [\phi_m(u)](t)| \leq \max_{1 \leq j \leq k_m} |v_j - u_j| \leq |v - u|$$

for any $v, u \in \mathbb{R}^{k_m}$ and $t \in E$, so that $\|\phi_m(v) - \phi_m(u)\|_\infty \leq |v - u|$. Therefore,

$$Y_m^n = \phi_m\left(X_n(t_1^m), \cdots, X_n(t_{k_m}^m)\right) \xrightarrow{d} \phi_m\left(X(t_1^m), \cdots, X(t_{k_m}^m)\right) = Y_m$$

as $n \to \infty$ by the continuous mapping theorem.

Completing the proof of convergence in distribution

Now choose any bounded Lipschitz function $f : \ell^\infty(E) \to \mathbb{R}$ with Lipschitz constant $L > 0$ and $|f(x)| \leq M$ for any $x \in \ell^\infty(E)$. We can see that

$$\left|\mathbb{E}^*[f(X_n)] - \mathbb{E}\left[f(\bar{X})\right]\right| \leq |\mathbb{E}^*[f(X_n)] - \mathbb{E}^*[f(Y_m^n)]| + |\mathbb{E}^*[f(Y_m^n)] - \mathbb{E}[f(Y_m)]| + \left|\mathbb{E}[f(Y_m)] - \mathbb{E}\left[f(\bar{X})\right]\right|$$

for any $n, m \in N_+$. For any $\epsilon > 0$, the first term can be bounded above by

$$\mathbb{E}^*|f(X_n) - f(Y_m^n)| \leq L\epsilon + 2M \cdot \mathbb{P}^*\left[\|X_n - X_m^n\|_\infty > \epsilon\right].$$

We saw above that

$$\|X_n - X_m^n\|_\infty \leq \sup_{1 \leq j \leq k_m} \sup_{t \in E_j^m} \left|X_n(t_j^m) - X_n(t)\right| \leq \sup_{1 \leq j \leq k_m} \sup_{s,t \in E_j^m} |X_n(s) - X_n(t)|,$$

and by construction we know that

$$\limsup_{n \to \infty} \mathbb{P}^*\left[\|X_n - X_m^n\|_\infty > \epsilon\right] \leq \limsup_{n \to \infty} \mathbb{P}^*\left[\sup_{1 \leq j \leq k_m} \sup_{s,t \in E_j^m} |X_n(s) - X_n(t)| > 2^{-m}\right] < 2^{-m}$$

for any m such that $2^{-m} < \epsilon$. As such, for any m such that $2^{-m} < \epsilon$,

$$\begin{aligned}\limsup_{n \to \infty} \left|\mathbb{E}^*[f(X_n)] - \mathbb{E}\left[f(\bar{X})\right]\right| \leq L\epsilon + 2M \cdot 2^{-m} + \left|\mathbb{E}[f(Y_m)] - \mathbb{E}\left[f(\bar{X})\right]\right| \\ + \limsup_{n \to \infty} |\mathbb{E}^*[f(Y_m^n)] - \mathbb{E}[f(Y_m)]|.\end{aligned}$$

We saw above that for any fixed m, the last term is equal to 0 by the Portmanteau lemma, so that

$$\limsup_{n \to \infty} \left|\mathbb{E}^*[f(X_n)] - \mathbb{E}\left[f(\bar{X})\right]\right| \leq L\epsilon + 2M \cdot 2^{-m} + \left|\mathbb{E}[f(Y_m)] - \mathbb{E}\left[f(\bar{X})\right]\right|.$$

Taking $m \to \infty$ at this point shows us that

$$\limsup_{n\to\infty} \left| \mathbb{E}^* \left[f(X_n) \right] - \mathbb{E}\left[f(\bar{X}) \right] \right| \leq L\epsilon,$$

since $Y_m \xrightarrow{d} \bar{X}$ as $m \to \infty$. This in turn holds for any $\epsilon > 0$, so we can conclude that

$$\lim_{n\to\infty} \mathbb{E}^* \left[f(X_n) \right] = \mathbb{E}\left[f(\bar{X}) \right],$$

and by the Portmanteau lemma we conclude that $X_n \xrightarrow{d} \bar{X}$ in the uniform metric on $\ell^\infty(E)$.

d-continuity of the paths of $\bar{X}$

Suppose now that d is a metric on E, and that $t_j^m \in E_j^m$ have been chosen so that

$$\left(\max_{1\leq i\neq j\leq k_m} \frac{\left| X(t_j^m) - X(t_i^m) \right|}{d(t_j^m, t_i^m)} \right) \left(\max_{1\leq j\leq k_m} \operatorname{diam}_d(E_j^m) \right) \xrightarrow{a.s.} 0$$

as $m \to \infty$. We show that $\bar{X}$ has continuous paths by first constructing a sequence $\{L_m\}_{m\in N_+}$ of mappings $L_m : \Omega \to \mathcal{C}_b(E,d)$, where $\mathcal{C}_b(E,d)$ is the set of all continuous and bounded functions on (E,d). If $L_m(\omega) \to \bar{X}(\omega)$ uniformly, then because $\mathcal{C}_b(E,d)$ is a complete and therefore closed subspace of $\ell^\infty(E)$ with respect to the uniform metric, it follows that $\bar{X}(\omega)$ takes values in $\mathcal{C}_b(E,d)$ as well.

Let Ω_0 be the almost sure set on which $[X(\omega)](t) = [\bar{X}(\omega)](t)$ for any $t \in E_0, \omega \in \Omega_0$ and the above condition holds. For any $m \in N_+$, we use the McShane extension to construct L_m. Choose any $\omega \in \Omega$ and let

$$\ell_m(\omega) = \max_{i\neq j\in\{1,\cdots,k_m\}} \frac{\left| [\bar{X}(\omega)](t_i^m) - [\bar{X}(\omega)](t_j^m) \right|}{d(t_i^m, t_j^m)} < +\infty.$$

Now define

$$[L_m(\omega)](t) = \min_{1\leq j\leq k_m} \left\{ [\bar{X}(\omega)](t_j^m) + \ell_m(\omega) d(t, t_j^m) \right\}$$

for any $t \in E$. By construction, for any $t, s \in E$, if $[L_m(\omega)](s) = [\bar{X}(\omega)](t_j^m) + \ell_m(\omega) d(s, t_j^m)$, then

$$[L_m(\omega)](t) - [L_m(\omega)](s) \leq \ell_m(\omega) \left(d(t, t_j^m) - d(s, t_j^m) \right) \leq \ell_m(\omega) d(t,s),$$

and since this holds for the order of t and s interchanged as well, we can see that $L_m(\omega)$ is Lipschitz on E with Lipschitz constant $\ell_m(\omega)$. We have not yet established that L_m has paths that are bounded.

Recall how we defined the mapping $Y_m : \Omega \to \ell^\infty(E)$ as a piecewise constant function. Now, for any $\omega \in \Omega$ and $t \in E_j^m$, suppose that $[L_m(\omega)](t) = [\bar{X}(\omega)](t_i^m) + \ell_m(\omega) d(t, t_i^m)$ for some $1 \leq i \leq k_m$.

Then,

$$[\bar{X}(\omega)](t_i^m) + \ell_m(\omega)d(t,t_i^m) \leq [\bar{X}(\omega)](t_j^m) + \ell_m(\omega)d(t,t_j^m)$$

by the definition of $[L_m(\omega)](t)$, so that

$$[L_m(\omega)](t) - [Y_m(\omega)](t) = [\bar{X}(\omega)](t_i^m) + \ell_m(\omega)d(t,t_i^m) - [\bar{X}(\omega)](t_j^m) \leq \ell_m(\omega)d(t,t_j^m).$$

Similarly,

$$\begin{aligned}
[Y_m(\omega)](t) - [L_m(\omega)](t) &= [\bar{X}(\omega)](t_j^m) - [\bar{X}(\omega)](t_i^m) - \ell_m(\omega)d(t,t_i^m) \\
&\leq \ell_m(\omega)\left(d(t_i^m,t_j^m) - d(t,t_i^m)\right) \\
&\leq \ell_m(\omega)d(t,t_j^m),
\end{aligned}$$

so that

$$|[Y_m(\omega)](t) - [L_m(\omega)](t)| \leq \ell_m(\omega)d(t,t_j^m).$$

As such,

$$\|Y_m(\omega) - L_m(\omega)\|_\infty \leq \ell_m(\omega)\left(\max_{1\leq j\leq k_m} \mathrm{diam}_d(E_j^m)\right) < +\infty,$$

which implies that $L_m : \Omega \to \mathcal{C}_b(E,d)$ because Y_m maps into the set of bounded functions $\ell^\infty(E)$ on E.

For any $\omega \in \Omega_0$,

$$\ell_m(\omega) = \max_{1\leq i\neq j\leq k_m} \frac{\left|[X(\omega)](t_j^m) - [X(\omega)](t_i^m)\right|}{d(t_j^m,t_i^m)}$$

and therefore $\ell_m(\omega)\left(\max_{1\leq j\leq k_m} \mathrm{diam}_d(E_j^m)\right) \to 0$ as $m \to \infty$ by assumption. It follows that $Y_m(\omega) - L_m(\omega) \to 0$ in the uniform metric, and since $Y_m(\omega) \to \bar{X}(\omega)$ in the uniform metric for any $\omega \in \Omega$, we have $L_m(\omega) \to \bar{X}(\omega)$ in d_∞ for any $\omega \in \Omega_0$. Since each $L_m(\omega)$ takes values in the closed subset $\mathcal{C}_b(E,d)$ of $\ell^\infty(E)$,$\bar{X}(\omega) \in \mathcal{C}_b(E,d)$ as well. The desired result follows now by noting that this holds for any ω in the almost sure set Ω_0.

□

10.2.3 Application: The Functional CLT

Let $\mathcal{D}[0,1]$ be the collection of all cadlag functions (right continuous with left limits) with respect to the euclidean metric on the interval $[0,1]$, and $\mathcal{C}[0,1]$ the collection of all continuous functions on the interval $[0,1]$ (again with respect to the euclidean metric). Note how $\mathcal{C}[0,1] \subset \mathcal{D}[0,1] \subset \ell^\infty([0,1])$.

For any $T \in N_+$, define the stochastic process $\{X_T(r)\}_{r\in[0,1]}$ as

$$X_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \epsilon_t \quad \forall r \in [0,1]$$

where $\{\epsilon_t\}_{t\in\mathbb{N}}$ is a martingale difference sequence with respect to some filtration $\mathcal{F}$ with mean 0 and variance 1. Letting $X_T : \Omega \to \mathcal{D}[0,1]$ be the corresponding function, we want to prove that $X_T \xrightarrow{d} W$, where $W : \Omega \to \ell^\infty(E)$ is a measurable function that takes values in $\mathcal{C}[0,1]$ and has stationary and independent Gaussian increments—W is exactly the random function corresponding to the standard Wiener process on $[0,1]$.

The proof proceeds by first showing that the finite-dimensional distributions of X_T converge to those of W, in what is ultimatley an application of the martingale difference sequence CLT. Afterward, we show that the limit $\bar{X}$ that has almost surely uniformly continuous paths with respect to the partition semimetric also has almost surely continuous paths with respect to the euclidean metric as well, which shows us that $\bar{X} = W$ almost surely.

Theorem 10.2.2 (Functional CLT)**.** *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space on which the standard Wiener random function $W : \Omega \to \mathcal{C}[0,1]$ and a sequence $\{\epsilon_t\}_{t\in\mathbb{Z}}$ of real random variables are supported. For any $T \in N_+$ define the stochastic process $\{X_T(r)\}_{r\in[0,1]}$ as*

$$X_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \epsilon_t \quad \forall r \in [0,1]$$

and let $X_T : \Omega \to \mathcal{D}[0,1]$ be the corresponding mapping into $\ell^\infty([0,1])$. Suppose that the following hold true:

i) ***MDS Property.*** *There exists a filtration $\mathcal{F}$ on $\mathbb{Z}$ such that $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a $\mathcal{F}$-martingale difference sequence.*

ii) ***Boundedness of Moments.*** *There exists some $\delta > 0$ such that $\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^{2+\delta} < +\infty$.*

iii) ***Convergence of Conditional Variances.*** *For any $r \in [0,1]$, $\frac{1}{T}\sum_{t=1}^{\lfloor Tr \rfloor} \mathbb{E}\left[\epsilon_t^2 \mid \mathcal{F}_{t-1}\right] \xrightarrow{p} r$ as $T \to \infty$.*

Then, $X_T \xrightarrow{d} W$ in the uniform metric on $\ell^\infty(E)$.

Proof. We need only show that:

- For any $r_1, \cdots, r_k \in [0,1]$,

$$(X_T(r_1), \cdots, X_T(r_k)) \xrightarrow{d} (W(r_1), \cdots, W(r_k))$$

- For any $m \in N_+$ there exists a finite partition $E_1^m, \cdots, E_{k_m}^m$ of $[0,1]$ such that

$$\limsup_{T\to\infty} \mathbb{P}^* \left[\sup_{1\le j\le k_m} \sup_{s,r\in E_j^m} |X_T(s) - X_T(r)| > 2^{-m} \right] < 2^{-m}.$$

in order to apply the preceding theorem and conclude that there exists some semimetric ρ on $[0,1]$ such that $X_T \xrightarrow{d} \bar{X}$ for some random function $\bar{X} : \Omega \to \ell^\infty(E)$ that takes values almost surely on $UC_b([0,1],\rho)$.

To show the first condition, we make use of the martingale difference CLT. Choose any $0 \le r_1 < \cdots < r_k \le 1$ and define $r_0 = 0$. Choose some $\alpha \in \mathbb{R}^k$, and define

$$A_T = \sum_{i=1}^k \alpha_i \left(X_T(r_i) - X_T(r_{i-1}) \right) = \frac{1}{\sqrt{T}} \sum_{i=1}^k \alpha_i \left(\sum_{t=\lfloor Tr_{i-1}\rfloor + 1}^{\lfloor Tr_i \rfloor} \epsilon_t \right).$$

Let $\{v_{t,T}\}_{1\le t\le \lfloor Tr_k\rfloor, T\in N_+}$ be the triangular array defined as

$$v_{t,T} = \frac{1}{\sqrt{T}} \alpha_i \epsilon_t \quad \text{if } \lfloor Tr_{i-1}\rfloor + 1 \le t \le \lfloor Tr_i \rfloor$$

for any $1 \le t \le \lfloor Tr_k \rfloor$. Then, $v_{t,T}$ is an $\mathcal{F}$-martingale difference array because $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an $\mathcal{F}$-MDS, and

$$\begin{aligned}
\sum_{t=1}^{\lfloor Tr_k\rfloor} \mathbb{E}\left[v_{t,T}^2 \mid \mathcal{F}_{t-1} \right] &= \sum_{i=1}^k \alpha_i^2 \left(\frac{1}{T} \sum_{t=\lfloor Tr_{i-1}\rfloor+1}^{\lfloor Tr_i\rfloor} \mathbb{E}\left[\epsilon_t^2 \mid \mathcal{F}_{t-1} \right] \right) \\
&= \sum_{i=1}^k \alpha_i^2 \left(\frac{1}{T} \sum_{t=1}^{\lfloor Tr_i\rfloor} \mathbb{E}\left[\epsilon_t^2 \mid \mathcal{F}_{t-1} \right] - \frac{1}{T} \sum_{t=1}^{\lfloor Tr_{i-1}\rfloor} \mathbb{E}\left[\epsilon_t^2 \mid \mathcal{F}_{t-1} \right] \right) \\
&\xrightarrow{p} \sum_{i=1}^k \alpha_i^2 \left(r_i - r_{i-1} \right)
\end{aligned}$$

by assumption iii). Finally, for each $1 \le t \le \lfloor Tr_i \rfloor$,

$$\mathbb{E}|v_{t,T}|^{2+\delta} \le \frac{1}{T^{1+\delta/2}} \left(\max_{1\le i\le k} |\alpha_i|^{2+\delta} \right) \left(\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^{2+\delta} \right),$$

so

$$\sum_{t=1}^{\lfloor Tr_k\rfloor} \mathbb{E}|v_{t,T}|^{2+\delta} \le \frac{\lfloor Tr_k\rfloor}{T} \times \frac{1}{T^{\delta/2}} \left(\max_{1\le i\le k} |\alpha_i|^{2+\delta} \right) \left(\sup_{t\in\mathbb{Z}} \mathbb{E}|\epsilon_t|^{2+\delta} \right) \to 0$$

as $T \to \infty$. This proves that $\{v_{t,T}\}_{t,T}$ satisfies the Lyapunov condition for $2+\delta$. It follows from the martingale difference array CLT that

$$A_T = \sum_{t=1}^{\lfloor Tr_k\rfloor} v_{t,T} \xrightarrow{d} \mathcal{N}\left[0, \sum_{i=1}^k \alpha_i^2 \left(r_i - r_{i-1} \right) \right].$$

This holds for any non-zero $\alpha \in \mathbb{R}^k$, so by the Cramer-Wold device we have

$$\begin{pmatrix} X_T(r_1) \\ X_T(r_2) - X_T(r_1) \\ \vdots \\ X_T(r_k) - X_T(r_{k-1}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left[O_{k\times 1}, \begin{pmatrix} r_1 & 0 & \cdots & 0 \\ 0 & r_2 - r_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_k - r_{k-1} \end{pmatrix} \right]$$

and

$$\begin{pmatrix} X_T(r_1) \\ X_T(r_2) - X_T(r_1) \\ \vdots \\ X_T(r_k) - X_T(r_{k-1}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} W(r_1) \\ W(r_2) - W(r_1) \\ \vdots \\ W(r_k) - W(r_{k-1}) \end{pmatrix}$$

by the uniqueness of weak limits. The continuous mapping theorem now allows us to conclude that $(X_T(r_1), \cdots, X_T(r_k)) \xrightarrow{d} (W(r_1), \cdots, W(r_k))$.

To show that the second condition holds, for any $m \in N_+$ construct a partition $\{E_1^m, \cdots, E_{k_m}^m\}$ of $[0,1]$ as $E_j^m = \left[\frac{j-1}{k_m}, \frac{j}{k_m}\right)$ for any $1 \le j \le k_m$, where $E_{k_m}^m$ contains the right boundary as well. Denote $a_{j,T}^m = \lfloor T\frac{j-1}{k_m} \rfloor$ for any $1 \le j \le k_m + 1$. Then, for any $r, s \in E_j^m$,

$$\begin{aligned} |X_T(r) - X_T(s)| &\le \left| X_T(r) - X_T\left(\frac{a_{j,T}^m}{T}\right) \right| + \left| X_T(s) - X_T\left(\frac{a_{j,T}^m}{T}\right) \right| \\ &\le \frac{2}{\sqrt{T}} \max_{0 \le q \le a_{j+1,T}^m - a_{j,T}^m} \left| \sum_{t=a_{j,T}^m+1}^{a_{j,T}^m+q} \epsilon_t \right| \\ &\le \frac{2}{\sqrt{T}} \max_{0 \le q \le \Delta_T^m} \left| \sum_{t=a_{j,T}^m+1}^{a_{j,T}^m+q} \epsilon_t \right| \end{aligned}$$

where $\Delta_T^m = a_{j+1,T}^m - a_{j,T}^m$. It follows that, for any $\epsilon > 0$,

$$\left\{ \sup_{1 \le j \le k_m} \sup_{s,r \in E_j^m} |X_T(s) - X_T(r)| > 2^{-m} \right\} \subset \bigcup_{j=1}^{k_m} \left\{ \frac{2}{\sqrt{T}} \max_{0 \le q \le \Delta_T^m} \left| M_q(a_{j,T}^m) \right| > 2^{-m} \right\}.$$

where

$$M_q(a) = \sum_{t=a+1}^{a+q} \epsilon_t$$

for any $q, a \in \mathbb{N}$. For each $1 \le j \le k_m$, the process $\{M_q(a_{j,T}^m)\}_{q \in \mathbb{N}}$ is a martingale, so by Doob's

maximal inequality,

$$\mathbb{P}\left(\max_{0\le q\le \Delta_T^m}\left|M_q(a_{j,T}^m)\right| > 2^{-m-1}\sqrt{T}\right) \le 2^{(m+1)(2+\delta)}T^{-1-\delta/2}\mathbb{E}\left|M_{\Delta_T^m}(a_{j,T}^m)\right|^{2+\delta}$$

$$= 2^{(m+1)(2+\delta)}T^{-1-\delta/2}\mathbb{E}\left|\sum_{t=a_{j,T}^m+1}^{a_{j+1,T}^m}\epsilon_t\right|^{2+\delta}$$

$$\le 2^{(m+1)(2+\delta)}T^{-1-\delta/2}C_{2+\delta}\cdot\mathbb{E}\left|\sum_{t=a_{j,T}^m+1}^{a_{j+1,T}^m}\epsilon_t^2\right|^{1+\frac{\delta}{2}}$$

where we used the Burkholder result to assert the existence of some constant $C_{2+\delta}$ satisfying the last inequality. By the summation version of Hölder's inequality,

$$\mathbb{E}\left|\sum_{t=a_{j,T}^m+1}^{a_{j+1,T}^m}\epsilon_t^2\right|^{1+\frac{\delta}{2}} \le \left(\sum_{t=a_{j,T}^m+1}^{a_{j+1,T}^m}\mathbb{E}|\epsilon_t|^{2+\delta}\right)(\Delta_T^m)^{\frac{\delta}{2}}$$

$$\le (\Delta_T^m)^{1+\frac{\delta}{2}}\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|\epsilon_t|^{2+\delta}\right).$$

Since $\Delta_T^m \le \frac{T}{k_m}+1$, it follows that

$$\mathbb{P}^*\left[\sup_{1\le j\le k_m}\sup_{s,r\in E_j^m}|X_T(s)-X_T(r)| > 2^{-m}\right] \le \sum_{j=1}^{k_m}\mathbb{P}\left(\max_{0\le q\le\Delta_T^m}\left|M_q(a_{j,T}^m)\right| > 2^{-m-1}\sqrt{T}\right)$$

$$\le C_{2+\delta}\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|\epsilon_t|^{2+\delta}\right)\times k_m 2^{(m+1)(2+\delta)}\left(\frac{1}{k_m}+\frac{1}{T}\right)^{1+\frac{\delta}{2}}.$$

As $T\to\infty$, we can see that

$$\limsup_{T\to\infty}\mathbb{P}^*\left[\sup_{1\le j\le k_m}\sup_{s,r\in E_j^m}|X_T(s)-X_T(r)| > 2^{-m}\right] \le C_{2+\delta}\left(\sup_{t\in\mathbb{Z}}\mathbb{E}|\epsilon_t|^{2+\delta}\right)2^{(m+1)(2+\delta)}k_m^{-\frac{\delta}{2}}.$$

A specific value of k_m under which condition ii) is satisfied is then

$$k_m \propto 2^{cm}$$

for some natural number c such that $c > \frac{6}{\delta}+2$.

We have thus shown that both conditions of the preceding theorem are satisfied by X_n and W: thus, there exists a semimetric ρ on $[0,1]$ and a random function $\bar{X}:\Omega\to UC_b([0,1],\rho)$ such that $X_T \xrightarrow{d} \bar{X}$.

Now it remains to show that $\bar{X}$ almost surely has paths that are continuous with respect to

the euclidean metric. Then, $\bar{X}$ will almost surely take values on $\mathcal{C}[0,1]$, and, its finite dimensional distributions will coincide with those of W by condition i) and the continuous mapping theorem. The fact that the collection of all inverse images of finite-dimensional projections is a π-system generating the Borel σ-algebra on $\mathcal{C}[0,1]$ then implies that $\bar{X}$ has the same distribution as W, and therefore we can assert $X_T \xrightarrow{d} W$.

To establish the requisite continuity property, in light of the preceding theorem it suffices to show that we can choose points $r_j^m \in E_j^m = \left[\frac{j-1}{k_m}, \frac{j}{k_m}\right)$ such that

$$\left(\max_{1 \le i \ne j \le k_m} \frac{\left| W(r_j^m) - W(r_i^m) \right|}{\left| r_j^m - r_i^m \right|} \right) \frac{1}{k_m} \xrightarrow{a.s.} 0$$

as $m \to \infty$. Let r_j^m be the left endpoint of the interval E_j^m, that is, $r_j^m = \frac{j-1}{k_m}$. Then, for any $\omega \in \Omega$ and $1 \le i < j \le k_m$, we have

$$\begin{aligned} \left| [W(\omega)](r_j^m) - [W(\omega)](r_i^m) \right| &= \left| \sum_{\ell=i}^{j-1} \left\{ [W(\omega)]\left(\frac{\ell}{k_m}\right) - [W(\omega)]\left(\frac{\ell-1}{k_m}\right) \right\} \right| \\ &\le \sum_{\ell=i}^{j-1} \left| [W(\omega)]\left(\frac{\ell}{k_m}\right) - [W(\omega)]\left(\frac{\ell-1}{k_m}\right) \right| \\ &\le (j-i) \max_{1 \le \ell \le k_m} \left| [W(\omega)]\left(\frac{\ell}{k_m}\right) - [W(\omega)]\left(\frac{\ell-1}{k_m}\right) \right|. \end{aligned}$$

It follows that

$$\frac{\left| [W(\omega)](r_j^m) - [W(\omega)](r_i^m) \right|}{\left| r_j^m - r_i^m \right|} \le k_m \max_{1 \le \ell \le k_m} \left| [W(\omega)]\left(\frac{\ell}{k_m}\right) - [W(\omega)]\left(\frac{\ell-1}{k_m}\right) \right|,$$

and since this holds for any $1 \le i < j \le k_m$,

$$\left(\max_{1 \le i \ne j \le k_m} \frac{\left| W(r_j^m) - W(r_i^m) \right|}{\left| r_j^m - r_i^m \right|} \right) \frac{1}{k_m} \le \max_{1 \le \ell \le k_m} \left| [W(\omega)]\left(\frac{\ell}{k_m}\right) - [W(\omega)]\left(\frac{\ell-1}{k_m}\right) \right|.$$

$W(\omega) : [0,1] \to \mathbb{R}$ is uniformly continuous on $[0,1]$ and $k_m \to \infty$ as $m \to \infty$, so the desired result follows and we can conclude that $\bar{X}$ almost surely has paths that are continuous with respect to the euclidean metric.

□

Note that the standard Wiener process on $[0,1]$ can be constructed using continuous function space methods, as in the Bilingsley text. The theorem then concerns the convergence of partial sum processes with paths in $\mathcal{D}[0,1]$ to this Wiener process.

10.3 Gilvenko-Cantelli and Donsker Classes

The machinery developed above is instrumental to the study of empirical processes. Let $\{X_n\}_{n\in N_+}$ be a random sample drawn from a distribution P on the measurable space $(E,\mathcal{E})$. Given any $\mathcal{E}$-measurable function $f : E \to \mathbb{R}$, we denote by Pf the integral $\int_E f dP$, granted it exists. The empirical analogue to this population value is defined as $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$. Given a class of functions $\mathcal{F}$ of interest, the goal in this section is to investigate the uniform convergence over $\mathcal{F}$ of two classes of objects:

$$|P_n f - Pf| \quad \text{and} \quad \sqrt{n}\,[P_n f - Pf]\,.$$

Given any $f \in L^1(P)$, we know from the SLLN that $P_n f \overset{a.s.}{\to} Pf$ pointwise. If

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \overset{a.s.^*}{\to} 0,$$

we call the class $\mathcal{F}$ of functions P-Gilvenko Cantelli in light of the Glivenko-Cantelli theorem for distribution functions.

Similarly, for any $f \in L^2(P)$, the Lindeberg-Levy CLT tells us that

$$\sqrt{n}\,[P_n f - Pf] = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(f(X_i) - \int_E f dP\right) \overset{d}{\to} \mathcal{N}\left[0,\ \int_E (f(x) - Pf)^2 dP(x)\right].$$

We say that a class of functions $\mathcal{F} \subset L^2(P)$ is P-Donsker if the sequence of mappings $G_n : \Omega \to \{E \to \mathbb{R}\}$ on Ω defined as

$$[G_n(\omega)](f) = \sqrt{n}\,[[P_n f](\omega) - Pf] \quad \forall f \in \mathcal{F}$$

converges in distribution to a Gaussian process on $\mathcal{F}$, that is, a random function taking values in $\ell^\infty(\mathcal{F})$ whose finite-dimensional distributions are jointly Gaussian. Here, $\{E \to \mathbb{R}\}$ denotes the set of all real-valued functions on E.

10.3.1 Bracketing Numbers and the Glivenko-Cantelli Theorem Revisited

Sufficient conditions for a class of functions to be Glivenko-Cantelli or Donsker are furnished through the use of bracketing numbers and entropy. For any pair of functions $\ell, u : E \to \mathbb{R}$, the bracket $[\ell, u]$ is defined as

$$[\ell, u] = \{f : E \to \mathbb{R} \mid \ell \le f \le u \text{ pointwise on } E\}.$$

We say that $[\ell, u]$ is an ϵ-bracket in $L^r(P)$ if

$$\|u - \ell\|_{P,r} = \left(\int_E |u - \ell|^r dP\right)^{\frac{1}{r}} < \epsilon.$$

Given a class $\mathcal{F}$ of real valued functions on E, the bracketing number $N_{[]}(\epsilon, \mathcal{F}, L^r(P))$ is the minimum number of ϵ-brackets in $L^r(P)$ required to cover $\mathcal{F}$. Note that the mapping $\epsilon \mapsto N_{[]}(\epsilon, \mathcal{F}, L^r(P))$ is necessarily decreasing in ϵ, since any collection of ϵ-brackets that covers $\mathcal{F}$ is also a collection of η-brackets covering $\mathcal{F}$ for $\eta > \epsilon$. Since monotonically decreasing functions on the real line have at most countably many discontinuity points, this mapping is Lebesgue measurable on $(0, +\infty)$.

It can be shown easily that a sufficient condition for a class of functions $\mathcal{F}$ to be P-Glivenko Cantelli is for $N_{[]}(\epsilon, \mathcal{F}, L^1(P))$ to be finite for any $\epsilon > 0$.

Theorem 10.3.1 (Glivenko-Cantelli Theorem)**.** *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space and $\{X_n\}_{n \in N_+}$ a random sample drawn from a distribution P on the measurable space $(E, \mathcal{E})$. Suppose $\mathcal{F}$ is a collection of real-valued $\mathcal{E}$-measurable functions. If $N_{[]}(\epsilon, \mathcal{F}, L^1(P)) < +\infty$ for any $\epsilon > 0$, then*

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \overset{a.s.*}{\to} 0.$$

Proof. Choose any $m \in N_+$, and let $\{[\ell_1^m, u_1^m], \cdots, [\ell_{k_m}^m, u_{k_m}^m]\}$ be the minimal collection of $\frac{1}{m}$-brackets in $L^1(P)$ that cover $\mathcal{F}$. Thus, for any $f \in \mathcal{F}$, there exists some $1 \leq j \leq k_m$ such that $\ell_j^m \leq f \leq u_j^m$. Furthermore, since $[\ell_j^m, u_j^m]$ is an $\frac{1}{m}$-bracket in $L^1(P)$, $\int_E \left| u_j^m - \ell_j^m \right| dP < \frac{1}{m}$ and we can see that

$$\int_E u_j^m dP < \int_E \ell_j^m dP + \frac{1}{m}.$$

Now we can see that

$$P_n f - Pf = \frac{1}{n} \sum_{i=1}^n f(X_i) - \int_E f dP \leq \frac{1}{n} \sum_{i=1}^n u_j^m(X_i) - \int_E u_j^m dP + \frac{1}{m},$$

while

$$Pf - P_n f \leq \int_E u_j^m dP - \frac{1}{n} \sum_{i=1}^n \ell_j^m(X_i) \leq \int_E \ell_j^m dP - \frac{1}{n} \sum_{i=1}^n \ell_j^m(X_i) + \frac{1}{m}.$$

It follows that

$$|P_n f - Pf| \leq \left(\left| P_n u_j^m - P u_j^m \right| \vee \left| P_n \ell_j^m - P \ell_j^m \right| \right) + \frac{1}{m}.$$

This holds for any $f \in \mathcal{F}$, so

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \max_{1 \leq j \leq k_m} \left(\left| P_n u_j^m - P u_j^m \right| \vee \left| P_n \ell_j^m - P \ell_j^m \right| \right) + \frac{1}{m}.$$

Since $u_j^m, \ell_j^m \in L^1(P)$ for each $1 \leq j \leq k_m$, the SLLN tells us that

$$\Delta_n^m := \max_{1 \leq j \leq k_m} \left(\left| P_n u_j^m - P u_j^m \right| \vee \left| P_n \ell_j^m - P \ell_j^m \right| \right) \overset{a.s.}{\to} 0$$

as $n \to \infty$. Let $\Omega_m \in \mathcal{H}$ collect the outcomes for which $\Delta_n^m \to 0$ as $n \to \infty$. Then, $\Omega_0 = \bigcap_m \Omega_m$ is also a $\mathcal{H}$-measurable event with probability 1 under $\mathbb{P}$. Now define

$$\Delta_n = \inf_{m \in N_+} \left\{ \Delta_n^m + \frac{1}{m} \right\}$$

for any $n \in N_+$. Since Δ_n is formed as the pointwise infimum of $\mathcal{H}/\mathcal{B}(\mathbb{R})$-measurable functions, it is also $\mathcal{H}/\mathcal{B}(\mathbb{R})$-measurable. Furthermore, for any $\omega \in \Omega_0$,

$$\Delta_n(\omega) \leq \Delta_n^m(\omega) + \frac{1}{m}$$

for any $m \in N_+$, so that, taking $n \to \infty$, we have $\limsup_{n \to \infty} \Delta_n(\omega) \leq \frac{1}{m}$. Taking $m \to \infty$ now shows us that $\Delta_n(\omega) \to 0$ as $n \to \infty$, and since this holds for any $\omega \in \Omega_0$, $\Delta_n \overset{a.s.}{\to} 0$. Finally, for any $n \in N_+$,

$$\sup_{f \in \mathcal{F}} |P_n f - P f| \leq \Delta_n^m + \frac{1}{m} \quad \forall m \in N_+,$$

so that

$$\sup_{f \in \mathcal{F}} |P_n f - P f| \leq \Delta_n.$$

This completes the proof.

□

The condition that bracketing numbers are finite also has implications for the mappings G_n. Specifically, it allows us to establish the integrability of a given class of functions and the boundedness of the paths of G_n.

Lemma 10.3.1. *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space and $\{X_n\}_{n \in N_+}$ a random sample drawn from a distribution P on the measurable space $(E, \mathcal{E})$. Suppose $\mathcal{F}$ is a collection of real-valued $\mathcal{E}$-measurable functions. If $N_{[]}(\epsilon, \mathcal{F}, L^r(P)) < +\infty$ for some $\epsilon > 0$ and $r \in [1, +\infty)$, then $\mathcal{F} \subset L^r(P)$ and $G_n(\omega) \in \ell^\infty(\mathcal{F})$ for every $\omega \in \Omega$.*

Proof. By assupmtion, there exists a collection $[\ell_1, u_1], \cdots, [\ell_k, u_k]$ of ϵ-brackets in $L^r(P)$ that covers $\mathcal{F}$. For any $f \in \mathcal{F}$, there exists some $1 \leq j \leq k$ such that $\ell_j \leq f \leq u_j$, so that $|f| \leq |u_j| \vee |\ell_j|$. Since the right hand side is the maximum of two functions in $L^r(P)$, $f \in L^r(P)$ as well, which establishes that $\mathcal{F} \subset L^r(P)$.

Now define $F = \max_{1 \leq i \leq k} (|u_j| \vee |\ell_j|)$, it follows that $F \in L^r(P)$ and $|f| \leq F$ pointwise for

any $f \in \mathcal{F}$. As such, for any $n \in N_+$ and $f \in \mathcal{F}$,

$$\begin{aligned}
|G_n(f)| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |f(X_i)| + \sqrt{n}\|f\|_{P,1} \\
&\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} F(X_i) + \sqrt{n}\|F\|_{P,1} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (F(X_i) - PF) + 2\sqrt{n}\|F\|_{P,1} \\
&\leq \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (F(X_i) - PF)}_{G_n(F)} + 2\sqrt{n}\|F\|_{P,r},
\end{aligned}$$

so that

$$\sup_{f \in \mathcal{F}} |G_n(f)| \leq |G_n(F)| + 2\sqrt{n}\|F\|_{P,r}.$$

It follows now that, for any $\omega \in \Omega$, and denoting the supremum norm on $\ell^\infty(\mathcal{F})$ by $\|\cdot\|_{\mathcal{F},\infty}$,

$$\|G_n(\omega)\|_{\mathcal{F},\infty} = \sup_{f \in \mathcal{F}} |[G_n(\omega)](f)| \leq |[G_n(\omega)](F)| + 2\sqrt{n}\|F\|_{P,r} < +\infty.$$

We have therefore verified that G_n maps into $\ell^\infty(\mathcal{F})$.

□

For a class of functions to be P-Donsker we require a stronger restriction than finiteness on the L^2-bracketing numbers $N_{[]}(\epsilon, \mathcal{F}, L^2(P))$. The bracketing entropy in $L^r(P)$ of the class $\mathcal{F}$ is defined as $\log N_{[]}(\epsilon, \mathcal{F}, L^r(P))$. The δ-bracketing integral in $L^r(P)$ is then given as

$$J_{[]}(\delta, \mathcal{F}, L^r(P)) = \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^r(P))} d\epsilon.$$

For $\mathcal{F}$ to be P-Donsker, it suffices that $J_{[]}(1, \mathcal{F}, L^2(P)) < +\infty$. Roughly speaking, this requires that $N_{[]}(\epsilon, \mathcal{F}, L^2(P)) < +\infty$ for $0 < \epsilon < 1$, so that each function in $\mathcal{F}$ is P-square integrable, and that the entropy $\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ grows at a slower speed than ϵ^{-2} as $\epsilon \to 0$. This is the sense in which the class $\mathcal{F}$ must be "small enough" for a uniform CLT over $\mathcal{F}$ to hold.

10.3.2 Maximal Inequalities

The proof of the general Donsker theorem requires a few complicated maximal inequalities that bounds the outer expectation of $\|G_n\|_{\mathcal{F},\infty} := \sup_{f \in \mathcal{F}} |G_n(f)|$. This supremum generally ranges over an infinite set, so to establish the desired bound, we make use of a chaining argument alongside maximal inequalities for finite maxima. First, we express each $G_n(f)$ as the telescoping sum of terms that are finite maxima—this is the chaining step.Afterward, we use an exponential

inequality for finite maxima to bound each of the terms in the telescoping sum.

Lemma 10.3.2 (Bernstein's Inequality)**.** *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, $(E, \mathcal{E})$ a measurable space and P a probability measure on $(E, \mathcal{E})$. Let $\{X_n\}_{n \in N_+}$ be an iid sequence of random variables taking values in E with distribution P. Then,*

$$\mathbb{P}\left(|G_n(f)| > x\right) \leq 2\exp\left(-\frac{1}{4}\frac{x^2}{\int_E f^2 dP + \frac{1}{\sqrt{n}}x\|f\|_\infty}\right)$$

for any $x > 0$ and $n \in N_+$.

Proof. Choose any bounded and $\mathcal{E}$-measurable $f : E \to \mathbb{R}$ such that $Pf^2 > 0$, $x > 0$ and $n \in N_+$. By Markov's inequality, for any $\lambda > 0$ we can see that

$$\begin{aligned}
\mathbb{P}\left(G_n(f) > x\right) = \mathbb{P}\left(e^{\lambda G_n(f)} > e^{\lambda x}\right) &\leq e^{-\lambda x}\mathbb{E}\left[e^{\lambda G_n(f)}\right] \\
&= e^{-\lambda x}\left(\mathbb{E}\left[e^{\lambda \frac{1}{\sqrt{n}}(f(X_1) - Pf)}\right]\right)^n \\
&= e^{-\lambda x}\left(\mathbb{E}\left[1 + \sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\sqrt{n}}\right)^k (f(X_1) - Pf)^k\right]\right)^n \\
&= e^{-\lambda x}\left(1 + \sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\sqrt{n}}\right)^k P(f - Pf)^k\right)^n
\end{aligned}$$

where the second line follows from the iid nature of $X_1, \cdots, X_n$, the third line from the definition of the exponential function, and the third from the DCT for series. Choose

$$\lambda = \frac{1}{2}\frac{x}{Pf^2 + \frac{1}{\sqrt{n}}x\|f\|_\infty}.$$

Defining $\lambda_1 = \frac{1}{2}\frac{x}{Pf^2}$ and $\lambda_2 = \frac{1}{2}\frac{\sqrt{n}}{\|f\|_\infty}$, note that $\lambda \leq \max(\lambda_1, \lambda_2)$ and

$$\lambda^k = \frac{1}{2^k}\frac{x^k}{\left(Pf^2 + \frac{1}{\sqrt{n}}x\|f\|_\infty\right)^k} \leq \lambda_1 \frac{1}{2^{k-1}}\frac{x^{k-1}}{\left(Pf^2 + \frac{1}{\sqrt{n}}x\|f\|_\infty\right)^{k-1}} \leq \lambda_1 \lambda_2^{k-1}$$

for any $k \geq 2$. By implication, $\lambda^k = \lambda\lambda^{k-1} \leq \lambda_1\lambda_2^{k-2}\lambda$ for any $k \geq 2$, so that

$$\begin{aligned}
\sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{\lambda}{\sqrt{n}}\right)^k P(f - Pf)^k &\leq \frac{\lambda\lambda_1}{n}\sum_{k=2}^{\infty}\frac{1}{k!}\left(\frac{\lambda_2}{\sqrt{n}}\right)^{k-2}\left|P(f - Pf)^k\right| \\
&\leq \frac{\lambda x}{2nPf^2}\sum_{k=2}^{\infty}\frac{1}{(2\|f\|_\infty)^{k-2}k!}\left|P(f - Pf)^k\right|.
\end{aligned}$$

Note that

$$\begin{aligned}\left|P(f-Pf)^k\right| &\leq \int_E (f-Pf)^2 |f-Pf|^{k-2} dP \\ &\leq (2\|f\|_\infty)^{k-2} \left(\int_E (f-Pf)^2 dP\right) \leq (2\|f\|_\infty)^{k-2} Pf^2\end{aligned}$$

where we use the fact that $|f-Pf| \leq 2\|f\|_\infty$ on E. Therefore,

$$\sum_{k=2}^{\infty} \frac{1}{k!} \left(\frac{\lambda}{\sqrt{n}}\right)^k P(f-Pf)^k \leq \frac{\lambda x}{2n} \sum_{k=2}^{\infty} \frac{1}{k!} = \frac{\lambda x}{2n}(e-2) \leq \frac{\lambda x}{2n}.$$

It follows that

$$\mathbb{P}(G_n(f) > x) \leq e^{-\lambda x} \left(1 + \frac{\lambda x}{2n}\right)^n,$$

and since the binomial theorem implies

$$\begin{aligned}\left(1+\frac{a}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{a}{n}\right)^k \\ &= \sum_{k=0}^{n} \frac{n(n-1)\times\cdots\times(n-k+1)}{n^k} \left(\frac{a^k}{k!}\right) \leq \sum_{k=0}^{n} \frac{a^k}{k!} = e^a\end{aligned}$$

for any $a > 0$, we have

$$\mathbb{P}(G_n(f) > x) \leq e^{-\lambda x} e^{\frac{1}{2}\lambda x} = e^{-\frac{1}{2}\lambda x} = e^{-\frac{1}{4}\frac{x^2}{Pf^2 + \frac{1}{\sqrt{n}} x\|f\|_\infty}}.$$

This holds for f replaced with $-f$ as well, and since $G_n(f) = -G_n(f)$, it follows that

$$\mathbb{P}(-G_n(f) > x) = \mathbb{P}(G_n(-f) > x) \leq e^{-\frac{1}{2}\lambda x} = e^{-\frac{1}{4}\frac{x^2}{Pf^2 + \frac{1}{\sqrt{n}} x\|f\|_\infty}}.$$

The desired result follows from the observation that

$$\mathbb{P}(|G_n(f)| > x) \leq \mathbb{P}(G_n(f) > x) + \mathbb{P}(-G_n(f) > x).$$

□

Lemma 10.3.3. *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, $(E, \mathcal{E})$ a measurable space and P a probability measure on $(E, \mathcal{E})$. Let $\{X_n\}_{n \in N_+}$ be an iid sequence of random variables taking values in E with distribution P.*

There exists a $C > 0$ such that, for any finite class $\mathcal{F}$ of bounded functions in $L^2(P)$ and $n \in N_+$, we have

$$\mathbb{E}^* \|G_n\|_{\mathcal{F},\infty} \leq C \left\{ \left(\max_{f \in \mathcal{F}} \frac{\|f\|_\infty}{\sqrt{n}}\right) \log(1 + |\mathcal{F}|) + \left(\max_{f \in \mathcal{F}} \|f\|_{P,2}\right) \sqrt{\log(1 + |\mathcal{F}|)} \right\}.$$

Proof. To prove the theorem we start by choosing any $n \in N_+$ and $f \in \mathcal{F}$ such that $Pf^2 > 0$. Defining $a = 24\frac{\|f\|_\infty}{\sqrt{n}}$ and $b = 24Pf^2$, we first establish that, for any $x > 0$, if $x \geq \frac{b}{a}$, then $-\frac{1}{4}\frac{x^2}{\int_E f^2 dP + \frac{1}{\sqrt{n}}x\|f\|_\infty} \leq \frac{-3x}{a}$, while if $x \leq \frac{b}{a}$, then $-\frac{1}{4}\frac{x^2}{\int_E f^2 dP + \frac{1}{\sqrt{n}}x\|f\|_\infty} \leq \frac{-3x^2}{b}$. To see this, note that $\frac{b}{a} = \frac{Pf^2}{\|f\|_\infty/\sqrt{n}}$, so that $x \geq \frac{b}{a}$ implies $\frac{1}{\sqrt{n}}x\|f\|_\infty \geq Pf^2$ and therefore

$$-\frac{1}{4}\frac{x^2}{Pf^2 + \frac{1}{\sqrt{n}}x\|f\|_\infty} \leq -\frac{1}{4}\frac{x^2}{2\frac{1}{\sqrt{n}}x\|f\|_\infty} = -\frac{x}{8\|f\|_\infty/\sqrt{n}} = -\frac{3x}{a}.$$

Meanwhile, if $x \leq \frac{b}{a}$, then $\frac{1}{\sqrt{n}}x\|f\|_\infty \leq Pf^2$ and we have

$$-\frac{1}{4}\frac{x^2}{Pf^2 + \frac{1}{\sqrt{n}}x\|f\|_\infty} \leq -\frac{1}{4}\frac{x^2}{2Pf^2} = -\frac{x^2}{8Pf^2} = -\frac{3x^2}{b}.$$

Now choose any $x > 0$. We can see that $|G_n(f)| \cdot I_{\{|G_n(f)|>b/a\}} > x$ implies that $|G_n(f)| > x$ and $|G_n(f)| > b/a$, so that

$$\mathbb{P}\left(|G_n(f)| \cdot I_{\{|G_n(f)|>b/a\}} > x\right) \leq \mathbb{P}\left(|G_n(f)| > x \vee \frac{b}{a}\right).$$

By Bernstein's inequality and the fact that $x \vee \frac{b}{a} \geq \frac{b}{a}$, it follows that

$$\mathbb{P}\left(|G_n(f)| \cdot I_{\{|G_n(f)|>b/a\}} > x\right) \leq 2\exp\left(-\frac{3}{a}\left(x \vee \frac{b}{a}\right)\right) \leq 2\exp\left(-\frac{3x}{a}\right).$$

Similarly, since $|G_n(f)| \cdot I_{\{|G_n(f)|\leq b/a\}} > x$ implies $x < |G_n(f)| \leq \frac{b}{a}$, Bernstein's inequality tells us that

$$\begin{aligned}\mathbb{P}\left(|G_n(f)| \cdot I_{\{|G_n(f)|\leq b/a\}} > x\right) &\leq \mathbb{P}\left(|G_n(f)| > x\right) \cdot I_{\{x\leq b/a\}} \\ &\leq 2\exp\left(-\frac{3x^2}{b}\right) \cdot I_{\{x\leq b/a\}} \leq 2\exp\left(-\frac{3x^2}{b}\right).\end{aligned}$$

For any $p \geq 1$, define $\psi_p : \mathbb{R}_+ \to \mathbb{R}_+$ as $\psi_p(x) = \exp(x^p) - 1$. Then, $\psi_p(\cdot)$ is strictly convex on $\mathbb{R}_+$ and admits an inverse function $\psi_p^{-1}(y) = (\log(1+y))^{\frac{1}{p}}$ for any $y \geq 0$. For any non-negative random variable X,

$$\begin{aligned}\mathbb{E}\left[\psi_1(X)\right] &= \int_\Omega \int_0^{X(\omega)} e^z dz \, d\mathbb{P}(\omega) \\ &= \int_\Omega \int_0^\infty e^z \cdot I_{[0,X(\omega)]}(z) dz \, d\mathbb{P}(\omega) \\ &= \int_0^\infty \mathbb{P}\left(X > z\right) e^z dz. \qquad \text{(Fubini's Theorem)}\end{aligned}$$

As such, we have

$$\mathbb{E}\left[\psi_1\left(\frac{1}{a}|G_n(f)|\cdot I_{\{|G_n(f)|>b/a\}}\right)\right] = \int_0^\infty \mathbb{P}\left(|G_n(f)|\cdot I_{\{|G_n(f)|>b/a\}} > xa\right) e^x dx$$
$$\leq \int_0^\infty e^{-2x} dx = \frac{1}{2} \leq 1.$$

Likewise, since

$$\begin{aligned}\mathbb{E}\left[\psi_2(X)\right] &= \int_\Omega \int_0^{X(\omega)} 2ze^{z^2}\, dz\, d\mathbb{P}(\omega) \\ &= \int_\Omega \int_0^\infty 2ze^{z^2} I_{[0,X(\omega)]}(z)\, dz\, d\mathbb{P}(\omega) \\ &= \int_0^\infty \mathbb{P}\left(X > z\right) 2ze^{z^2} dz\end{aligned}$$

for any non-negative random variable X by Fubini's theorem,

$$\mathbb{E}\left[\psi_2\left(\frac{1}{\sqrt{b}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq b/a\}}\right)\right] = \int_0^\infty \mathbb{P}\left(|G_n(f)|\cdot I_{\{|G_n(f)|\leq b/a\}} > x\sqrt{b}\right) 2xe^{x^2} dx$$
$$\leq \int_0^\infty 2xe^{-2x^2} dx = \frac{1}{2} \leq 1.$$

So far, we have established that

$$\mathbb{E}\left[\psi_1\left(\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right] \leq 1$$

$$\mathbb{E}\left[\psi_2\left(\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right] \leq 1$$

for any $f \in \mathcal{F}$ such that $Pf^2 > 0$. Then, by the convexity of $\psi_1(\cdot)$ and Jensen's inequality,

$$\begin{aligned}\psi_1\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\left\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\right\}}\right]\right) &\leq \mathbb{E}\left[\psi_1\left(\max_{f\in\mathcal{F}}\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\left\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\right\}}\right)\right] \\ &\qquad \text{(Jensen's inequality)} \\ &\leq \mathbb{E}\left[\max_{f\in\mathcal{F}}\psi_1\left(\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right] \\ &\leq \sum_{f\in\mathcal{F}}\mathbb{E}\left[\psi_1\left(\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\left\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\right\}}\right)\right] \\ &\qquad (\psi_1 \text{ is non-negative}) \\ &\leq |\mathcal{F}|.\end{aligned}$$

Taking the inverse of ψ_1 across both sides yields

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\leq\log(1+|\mathcal{F}|),$$

and since

$$\left(\max_{f\in\mathcal{F}}\frac{24\|f\|_\infty}{\sqrt{n}}\right)^{-1}\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\leq\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{\sqrt{n}}{24\|f\|_\infty}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right],$$

we have

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\leq 24\left(\max_{f\in\mathcal{F}}\frac{\|f\|_\infty}{\sqrt{n}}\right)\log(1+|\mathcal{F}|).$$

By a similar line of reasoning,

$$\begin{aligned}\psi_2\left(\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\right)&\leq\mathbb{E}\left[\psi_2\left(\max_{f\in\mathcal{F}}\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right]\\&\leq\mathbb{E}\left[\max_{f\in\mathcal{F}}\psi_2\left(\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right]\\&\leq\sum_{f\in\mathcal{F}}\mathbb{E}\left[\psi_2\left(\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right)\right]\\&\leq\mathcal{F},\end{aligned}$$

so that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}\frac{1}{\sqrt{24Pf^2}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\leq\sqrt{\log(1+|\mathcal{F}|)}.$$

As such,

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]\leq\sqrt{24}\left(\max_{f\in\mathcal{F}}\|f\|_{P,2}\right)\sqrt{\log(1+|\mathcal{F}|)},$$

where $\|f\|_{P,2}=\sqrt{Pf^2}$.

Finally, the relation

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\right]\leq\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\cdot I_{\{|G_n(f)|>\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]+\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\cdot I_{\{|G_n(f)|\leq\frac{Pf^2}{\|f\|_\infty/\sqrt{n}}\}}\right]$$

implies that

$$\mathbb{E}\left[\max_{f\in\mathcal{F}}|G_n(f)|\right]\leq 24\left\{\left(\max_{f\in\mathcal{F}}\frac{\|f\|_\infty}{\sqrt{n}}\right)\log(1+|\mathcal{F}|)+\left(\max_{f\in\mathcal{F}}\|f\|_{P,2}\right)\sqrt{\log(1+|\mathcal{F}|)}\right\},$$

which is the desired result with $C=24$, which holds across any $n\in N_+$ and collection $\mathcal{F}$.

□

We are now ready to use the chaining argument alongside maximal inequalities for finite classes of functions. Given a class $\mathcal{F}$ of real-valued functions on E, we say that the function $F : E \to \mathbb{R}_+$ is an envelope function of $\mathcal{F}$ if $|f| \leq F$ for any $f \in \mathcal{F}$. Measurability of F is generally not required. Given a random variable X taking values in E with distribution P, we denote by $P^*(F)$ the outer expectation $\mathbb{E}^*[F(X)]$.

Lemma 10.3.4. *Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, $(E, \mathcal{E})$ a measurable space, P a probability measure on $(E, \mathcal{E})$. Let $\{X_n\}_{n \in N_+}$ be an iid sequence of random variables taking values in E with distribution P.*

Choose any $\delta > 0$ and class $\mathcal{F}$ of $\mathcal{E}$-measruable real valued functions with envelope function $F \in \mathcal{E}_+$ such that $\|f\|_{P,2} < \delta$ for any $f \in \mathcal{F}$. Suppose that $N_{[]}(\epsilon, \mathcal{F}, L^2(P)) < +\infty$ for any $\epsilon > 0$. Then, there exists a constant $C > 0$ such that

$$\mathbb{E}^* \|G_n\|_{\mathcal{F},\infty} \leq C \left\{ J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta + \sqrt{n} P\left(F \cdot I_{\{F > a(\delta)\sqrt{n}\}}\right) \right\}$$

for any $n \in N_+$, where $a(\delta) = \frac{\delta}{\sqrt{\log N_{[]}(\delta, \mathcal{F}, L^2(P)) \vee 1}} > 0$. The constant $C > 0$ is a universal constant that does not depend on any parameters.

Proof. By the assumption that $Pf^2 < \delta^2$ for any $f \in \mathcal{F}$, we can see that $N_{[]}(\delta, \mathcal{F}, L^2(P)) \in [1, +\infty)$, so that $a(\delta)$ is a well-defined finite number. Fix $n \in N_+$ and define two classes $\mathcal{G}_<$ and $\mathcal{G}_>$ as

$$\mathcal{G}_< = \{f \cdot I_{\{F \leq a(\delta)\sqrt{n}\}} \mid f \in \mathcal{F}\}, \quad \mathcal{G}_> = \{f \cdot I_{\{F > a(\delta)\sqrt{n}\}} \mid f \in \mathcal{F}\}.$$

For any $f \in \mathcal{F}$, the additivity of the mapping $G_n(\cdot)$ implies that

$$G_n(f) = G_n\left(f \cdot I_{\{F \leq a(\delta)\sqrt{n}\}}\right) + G_n\left(f \cdot I_{\{F > a(\delta)\sqrt{n}\}}\right),$$

so that

$$\sup_{f \in \mathcal{F}} |G_n(f)| \leq \sup_{f \in \mathcal{G}_<} |G_n(f)| + \sup_{f \in \mathcal{G}_>} |G_n(f)|.$$

Note how, for any $f \in \mathcal{F}$,

$$\left|G_n\left(f \cdot I_{\{F > a(\delta)\sqrt{n}\}}\right)\right| \leq \frac{1}{\sqrt{n}} \sum_{i=1}^n F(X_i) I_{\{F(X_i) > a(\delta)\sqrt{n}\}} + \sqrt{n} \int_E \left(F I_{\{F > a(\delta)\sqrt{n}\}}\right) dP,$$

so that

$$\sup_{f \in \mathcal{G}_>} |G_n(f)| \leq \frac{1}{\sqrt{n}} \sum_{i=1}^n F(X_i) I_{\{F(X_i) > a(\delta)\sqrt{n}\}} + \sqrt{n} P\left(F \cdot I_{\{F > a(\delta)\sqrt{n}\}}\right)$$

and we have

$$\mathbb{E}^*\|G_n\|_{\mathcal{G}_>,\infty} \leq 2\sqrt{n}P\left(F \cdot I_{\{F>a(\delta)\sqrt{n}\}}\right).$$

Therefore,

$$\mathbb{E}^*\|G_n\|_{\mathcal{F},\infty} \leq \mathbb{E}^*\|G_n\|_{\mathcal{G}_<,\infty} + 2\sqrt{n}P\left(FI_{\{F>a(\delta)\sqrt{n}\}}\right).$$

It remains to bound the first term with a multiple of $J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta$.

First, we choose an integer q_0 that belongs to the interval $[-\log_2(\delta)-3,\ -\log_2(\delta)-2]$; this interval is of length 1 and thus must always contain a (possibly negative) integer. Then, $4\delta \leq 2^{-q_0} \leq 8\delta$ for this choice of q_0.

The initial task is to construct a sequence of nested partitions of $\mathcal{G}_<$ satisfying certain conditions. For any $f \in \mathcal{F}$, if $[\ell, u]$ is an ϵ-bracket containing f, then

$$\left[\ell \cdot I_{\{F \leq a(\delta)\sqrt{n}\}},\ u \cdot I_{\{F \leq a(\delta)\sqrt{n}\}}\right]$$

is an ϵ-bracket containing $f \cdot I_{\{F \leq a(\delta)\sqrt{n}\}}$, so that $N_{[]}(\epsilon, \mathcal{G}_<, L^2(P)) \leq N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ for any $\epsilon > 0$. Thus, for any integer $q \geq q_0$ we can find a finite collection of ϵ-brackets $[\ell_1^q, u_1^q], \cdots, [\ell_{k_q}^q, u_{k_q}^q]$ that covers $\mathcal{G}_<$, where $k_q = N_{[]}(2^{-q}, \mathcal{F}, L^2(P))$. We can then define the partition $\{B_1^q, \cdots, B_{k_q}^q\}$ of $\mathcal{G}_<$ as

$$B_1^q = [\ell_1^q, u_1^q], \quad B_j^q = [\ell_j^q, u_j^q] \setminus \left(\bigcup_{i=1}^{j-1} [\ell_i^q, u_i^q]\right) \quad \forall 2 \leq j \leq k_q.$$

Defining $\mathbf{D}_j^q = (u_j^q - \ell_j^q) \wedge 2F$ for any $1 \leq j \leq k_q$, it follows that

$$\sup_{f,g \in B_j^q} |f - g| \leq \mathbf{D}_j^q, \quad \left\|\mathbf{D}_j^q\right\|_{P,2} \leq 2^{-q}$$

for each j. Furthermore, since the mapping $\epsilon \mapsto N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ is decreasing on $(0, +\infty)$ and $k_q = N_{[]}(2^{-q}, \mathcal{F}, L^2(P))$, it follows that

$$\sum_{q \geq q_0} 2^{-q-1} \sqrt{\log(k_q) \vee 1} \leq \int_0^{8\delta} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon,$$

where we used the fact that $2^{-q_0} \leq 8\delta$. We can now see that

$$\sum_{q \geq q_0} 2^{-q} \sqrt{\log(k_q) \vee 1} \leq 2 \int_0^{8\delta} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon \leq 16 \int_0^{\delta} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon,$$

where for the last inequality we used the fact that the mapping $\epsilon \mapsto \log N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ is decreasing.

This property also implies that there exists some $0 < \eta$ for which

$$\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \geq 1 \quad \text{if } \epsilon \leq \eta$$

$$\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \leq 1 \quad \text{if } \epsilon \geq \eta.$$

In the case that $\delta \leq \eta$,

$$\int_0^{\delta} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon = \int_0^{\delta} \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))} d\epsilon = J_{[]}(\delta, \mathcal{F}, L^2(P)).$$

On the other hand, if $\delta > \eta$, then

$$\int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon \leq \int_0^\eta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))} d\epsilon + (\delta - \eta)$$

$$\leq J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta.$$

Therefore, in general we have

$$\int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon \leq J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta.$$

To construct partitions that are nested, for each $q \geq q_0$ let $N_q = \prod_{j=q_0}^{q} k_j$ and define $\{A_1^q, \cdots, A_{N_q}^q\}$ as the collection

$$\Big\{ \bigcap_{j=q_0}^{q} B_{i_j}^j \mid 1 \leq i_j \leq k_j \quad \forall q_0 \leq j \leq q \Big\}.$$

By construction $\{A_1^{q+1}, \cdots, A_{N_{q+1}}^{q+1}\}$ is nested within $\{A_1^q, \cdots, A_{N_q}^q\}$.Choose any $1 \leq j \leq N_q$ and let

$$A_j^q = \bigcap_{\ell=q_0}^{q} B_{i_\ell}^\ell$$

for indices $1 \leq i_\ell \leq k_\ell$. Then, we define $\Delta_j^q = \min_{q_0 \leq \ell \leq q} \mathbf{D}_{i_\ell}^\ell$, so that

$$\sup_{f,g \in A_j^q} |f - g| \leq \Delta_j^q \quad \text{and} \quad \left\| \Delta_j^q \right\|_{P,2} \leq \left\| \mathbf{D}_{i_q}^q \right\|_{P,2} \leq 2^{-q}.$$

By design, if $1 \leq j \leq N_q$ and $1 \leq j' \leq N_{q+1}$ satisfy and $A_{j'}^{q+1} \subset A_{j'}^q$, we have $\Delta_{j'}^{q+1} \leq \Delta_j^q$.

Finally, since $\sqrt{\log(N_q) \vee 1} \leq \sum_{p=q_0}^{q} \sqrt{\log(k_p) \vee 1}$ for each $q \geq q_0$, we have

$$\begin{aligned}
\sum_{q \geq q_0} 2^{-q} \sqrt{\log(N_q) \vee 1} &\leq \sum_{q \geq q_0} 2^{-q} \sum_{p=q_0}^{q} \sqrt{\log(k_p) \vee 1} \\
&= \sum_{q \geq q_0} 2^{-q+1} \sqrt{\log(k_q) \vee 1} \\
&\leq 32 \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1} d\epsilon.
\end{aligned}$$

We have established the existence of nested partitions $\{A_1^q, \cdots, A_{N_q}^q\}$ of $\mathcal{G}_<$ and corresponding non-negative random variables $\{\Delta_1^q, \cdots, \Delta_{N_q}^q\}$ such that:

- $\sup_{f,g \in A_j^q} |f - g| \leq \Delta_j^q \leq 2F$ for any $1 \leq j \leq N_q$
- $\left\| \Delta_j^q \right\|_{P,2} \leq 2^{-q}$ for any $1 \leq j \leq N_q$
- For any $1 \leq j \leq N_{q+1}$, if $A_j^{q+1} \subset A_{j'}^q$ for some $1 \leq j' \leq N_q$, then $\Delta_j^{q+1} \leq \Delta_{j'}^q$.
- The following inequality holds:

$$\sum_{q \geq q_0} 2^{-q} \sqrt{\log(N_q) \vee 1} \leq 32 \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \vee 1}\, d\epsilon \leq 32 \left(J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta \right).$$

For each $q \geq q_0$ and $1 \leq j \leq N_q$, choose some element $f_j^q \in A_j^q$ (if $A_j^q = \emptyset$, then put f_j^q equal to any aribtrary function in $\mathcal{G}_<$). Then, for any $f \in \mathcal{G}_<$, we define

$$\pi_q(f) = f_j^q, \quad \Delta_q(f) = \Delta_j^q \quad \text{if } f \in A_j^q.$$

Thus $\{\pi_q(f) \mid f \in \mathcal{G}_<\}$ and $\{\Delta_q(f) \mid f \in \mathcal{G}_<\}$ are both finite collections of functions that contain at most N_q elements. Now define

$$a_q = \frac{2^{-q}}{\sqrt{\log(N_{q+1}) \vee 1}}$$

$$[A_q(f)](x) = I_{\{\Delta_{q_0}(f) \leq a_{q_0}\sqrt{n}, \cdots, \Delta_q(f) \leq a_q\sqrt{n}\}}(x)$$

$$[B_q(f)](x) = I_{\{\Delta_{q_0}(f) \leq a_{q_0}\sqrt{n}, \cdots, \Delta_{q-1}(f) \leq a_{q-1}\sqrt{n}, \Delta_q(f) > a_q\sqrt{n}\}}(x)$$

for any $q \geq q_0$, $f \in \mathcal{G}_<$ and $x \in E$. The nestedness of the partitions implies that, for any $q \geq q_0$, $A_q(f) = A_q(g)$ and $B_q(f) = B_q(g)$ for any $f, g \in \mathcal{G}_<$ contained in the same q-level partition set A_j^q.

Because $\sum_{q \geq q_0} 2^{-q}\sqrt{\log(N_q) \vee 1} < +\infty$, it follows from the nth term test that $2^{-q}\sqrt{\log(N_q) \vee 1} \to 0$ as $q \to \infty$. This, together with the fact that $\log(N_q) \vee 1 \to +\infty$ as $q \to \infty$, implies that $a_q \to 0$ as $q \to \infty$. By definition,

$$a_{q_0} = \frac{2^{-q_0}}{\sqrt{(\log(k_{q_0+1}) + \log(k_{q_0})) \vee 1}}.$$

Since $k_{q_0+1} = N_{[]}(2^{-q_0-1}, \mathcal{F}, L^2(P)) \geq N_{[]}(2^{-q_0}, \mathcal{F}, L^2(P)) = k_{q_0}$, and similarly, $k_{q_0+1} \leq N_{[]}(\delta, \mathcal{F}, L^2(P))$ because $2^{-q_0-1} \geq 2\delta > \delta$, it follows that

$$a_{q_0} \geq \frac{4\delta}{\sqrt{2\left(\log N_{[]}(\delta, \mathcal{F}, L^2(P)) \vee 1\right)}} \geq 2\frac{\delta}{\sqrt{\log N_{[]}(\delta, \mathcal{F}, L^2(P)) \vee 1}} = 2a(\delta).$$

Since $|f| \leq a(\delta)\sqrt{n}$ for any $f \in \mathcal{G}_<$, it follows that $[A_{q_0}(f)](x) = 1$ for any $f \in \mathcal{G}_<$ and $x \in E$. Morever, for any $f \in \mathcal{G}_<$ and $x \in E$, either $[B_{q_1}(f)](x) = 1$ for exactly one $q_1(x) > q_0$, or $[B_q(f)](x) = 0$ for any $q \geq q_0$. For brevity, we say that there exists a possibly infinite $q_1(x) > q_0$ such that $[B_{q_1}(f)](x) = 1$ and $[B_q(f)](x) = 0$ for any $q \neq q_1(x)$. Then, $[A_q(f)](x) = 1$ for any $q < q_1(x)$ and $[A_q(f)](x) = 0$ for any $q \geq q_1(x)$.

For any $f \in \mathcal{G}_<$ and $x \in E$, we can consider two cases. First, suppose that $q_1(x) < +\infty$. In this case,

$$f(x) - [\pi_{q_0}(f)](x) = f(x) - [\pi_{q_1(x)}(f)](x) + \sum_{q=q_0+1}^{q_1(x)} ([\pi_q(f)](x) - [\pi_{q-1}(f)](x))$$

$$= \sum_{q=q_0+1}^{\infty} (f(x) - [\pi_q(f)](x))\,[B_q(f)](x) + \sum_{q=q_0+1}^{\infty} ([\pi_q(f)](x) - [\pi_{q-1}(f)](x))\,[A_{q-1}(f)](x).$$

Meanwhile, if $q_1(x) = +\infty$, then for any $q \geq q_0$,

$$|f(x) - [\pi_q(f)](x)| \leq [\Delta_q(f)](x) \leq a_q\sqrt{n},$$

and taking $q \to \infty$ on both sides shows us that $[\pi_q(f)](x) \to f(x)$ as $q \to \infty$. It follows that

$$\begin{aligned} f(x) - [\pi_{q_0}(f)](x) &= \sum_{q=q_0+1}^{\infty} ([\pi_q(f)](x) - [\pi_{q-1}(f)](x)) \\ &= \sum_{q=q_0+1}^{\infty} (f(x) - [\pi_q(f)](x))\,[B_q(f)](x) + \sum_{q=q_0+1}^{\infty} ([\pi_q(f)](x) - [\pi_{q-1}(f)](x))\,[A_{q-1}(f)](x). \end{aligned}$$

Therefore, in any case,

$$f - \pi_{q_0}(f) = \sum_{q=q_0+1}^{\infty} (f - \pi_q(f))\,B_q(f) + \sum_{q=q_0+1}^{\infty} (\pi_q(f) - \pi_{q-1}(f))\,A_{q-1}(f)$$

pointwise on E.

Now applying the linear operator $G_n(\cdot)$ on both sides yields

$$\begin{aligned} G_n(f) - G_n[\pi_{q_0}(f)] &= G_n(f - \pi_{q_0}(f)) \\ &= \sum_{q=q_0+1}^{\infty} G_n[(f - \pi_q(f))\,B_q(f)] + \sum_{q=q_0+1}^{\infty} G_n[(\pi_q(f) - \pi_{q-1}(f))\,A_{q-1}(f)]. \end{aligned}$$

It follows then from countable subadditivity that the outer expectation of $\|G_n\|_{\mathcal{G}_<}$ is bounded above by

$$\begin{aligned} \mathbb{E}^*\|G_n\|_{\mathcal{G}_<,\infty} \leq\ & \mathbb{E}^*\|G_n(\pi_{q_0}(\cdot))\|_{\mathcal{G}_<,\infty} \\ & + \mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n[(f - \pi_q(f))\,B_q(f)]\right|\right] \\ & + \mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n[(\pi_q(f) - \pi_{q-1}(f))\,A_{q-1}(f)]\right|\right]. \end{aligned}$$

The proof will be complete if we can show that each of the three terms on the right hand side are bounded above by some multiple of the entropy integral $J_{[]}(\delta, \mathcal{F}, L^2(P))$.

By construction, $\Delta_q(f)B_q(f) \le \Delta_{q-1}(f)B_q(f) \le a_{q-1}\sqrt{n}$ and $\|\Delta_q(f)B_q(f)\|_{P,2} \le 2^{-q}$ for any $q > q_0$. For any $f \in \mathcal{G}_<$ and a non-negative $\mathcal{E}$-measurable function g such that $|f| \le g$,

$$\begin{aligned}
|G_n(f)| &\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}|f(X_i)| + \sqrt{n}P|f| \\
&\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(X_i) + \sqrt{n}Pg \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(X_i) - Pg) + 2\sqrt{n}Pg = G_n(g) + 2\sqrt{n}Pg.
\end{aligned}$$

As such, since $|f - \pi_q(f)|B_q(f) \le \Delta_q(f)B_q(f) \le a_{q-1}\sqrt{n}$ for any $q > q_0$, we can see that

$$|G_n[(f - \pi_q(f))B_q(f)]| \le G_n(\Delta_q(f)B_q(f)) + 2\sqrt{n}P(\Delta_q(f)B_q(f))$$

for any $q > q_0$, which implies

$$\begin{aligned}
\left|\sum_{q=q_0+1}^{\infty} G_n[(f - \pi_q(f))B_q(f)]\right| &\le \sum_{q=q_0+1}^{\infty} G_n(\Delta_q(f)B_q(f)) + 2\sqrt{n}\sum_{q=q_0+1}^{\infty} P(\Delta_q(f)B_q(f)) \\
&\le \sum_{q=q_0+1}^{\infty} \|G_n(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty} + 2\sqrt{n}\sum_{q=q_0+1}^{\infty} \|P(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\sum_{q=q_0+1}^{\infty} G_n[(f - \pi_q(f))B_q(f)]\right|\right] \\
&\qquad \le \sum_{q=q_0+1}^{\infty} \mathbb{E}^*\|G_n(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty} + 2\sqrt{n}\sum_{q=q_0+1}^{\infty} \|P(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty}
\end{aligned}$$

by the countable subadditivity of outer measures. For each $q > q_0$, because $\Delta_q(f)B_q(f) = \Delta_q(g)B_q(g)$ for any f, g contained in the same level q-set A_j^q, $\|G_n(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty}$ is the maximum over exactly N_q functions bounded above by $a_{q-1}\sqrt{n}$. As such, the preceding lemma implies that

$$\mathbb{E}^*\|G_n(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty} \le 24\left\{a_{q-1}\log(1 + N_q) + 2^{-q}\sqrt{\log(1 + N_q)}\right\}.$$

Since $\log(1 + m) \le \log(2m) = \log(2) + \log(m) \le 2\log(m)$ for any $m \in N_+$, it follows now that

$$\mathbb{E}^*\|G_n(\Delta_q(\cdot)B_q(\cdot))\|_{\mathcal{G}_<,\infty} \le 48\left\{a_{q-1}(\log(N_q) \vee 1) + 2^{-q}\sqrt{\log(N_q) \vee 1}\right\}.$$

On the other hand, since $\Delta_q(f) > a_q\sqrt{n}$ if $B_q(f) = 1$, it follows that

$$\Delta_q(f)B_q(f) \le \frac{1}{a_q\sqrt{n}}\Delta_q(f)^2 B_q(f)$$

and we have

$$2\sqrt{n}P(\Delta_q(f)B_q(f)) \le \frac{2}{a_q}P\left(\Delta_q(f)^2 B_q(f)\right) \le \frac{2}{a_q}2^{-2q}$$

for any $f \in \mathcal{G}_<$. Thus,

$$2\sqrt{n}\left[\sup_{f\in\mathcal{G}_<} P(\Delta_q(f)B_q(f))\right] \le \frac{2}{a_q}2^{-2q}$$

and we have

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(f-\pi_q(f)\right)B_q(f)\right]\right|\right] \le 48\sum_{q=q_0+1}^{\infty}\left\{a_{q-1}(\log(N_q)\vee 1) + 2^{-q}\sqrt{\log(N_q)\vee 1} + \frac{2}{a_q}2^{-2q}\right\}.$$

By the definition of a_q,

$$a_{q-1}(\log(N_q)\vee 1) = 2^{-q+1}\sqrt{\log(N_q)\vee 1},$$

$$\frac{2}{a_q} = 2^{q+1}\sqrt{\log(N_{q+1})\vee 1},$$

so that

$$\begin{aligned}\sum_{q=q_0+1}^{\infty}&\left\{a_{q-1}(\log(N_q)\vee 1) + 2^{-q}\sqrt{\log(N_q)\vee 1} + \frac{2}{a_q}2^{-2q}\right\}\\ &\le 4\left(\sum_{q=q_0+1}^{\infty}2^{-q}\sqrt{\log(N_q)\vee 1}\right) + 4\left(\sum_{q=q_0+1}^{\infty}2^{-(q+1)}\sqrt{\log(N_{q+1})\vee 1}\right)\\ &\le 8\left(\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log(N_q)\vee 1}\right).\end{aligned}$$

We have obtained out first bound:

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(f-\pi_q(f)\right)B_q(f)\right]\right|\right] \le 384\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log(N_q)\vee 1}.$$

The next step is to bound

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(\pi_q(f)-\pi_{q-1}(f)\right)A_{q-1}(f)\right]\right|\right].$$

For any $q > q_0$ and $f \in \mathcal{G}_<$, since $\pi_q(f)$ and $\pi_{q-1}(f)$ belong to the same $(q-1)$-level partition of $\mathcal{G}_<$,

$$|\pi_q(f)-\pi_{q-1}(f)|A_{q-1}(f) \le \Delta_{q-1}(f)A_{q-1}(f),$$

so that, by countable subadditivity,

$$\mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(\pi_q(f)-\pi_{q-1}(f)\right)A_{q-1}(f)\right]\right|\right] \le \sum_{q=q_0+1}^{\infty}\mathbb{E}^*\|G_n\left[\Delta_{q-1}(\cdot)A_{q-1}(\cdot)\right]\|_{\mathcal{G}_<,\infty}.$$

Since $\Delta_q(f)A_q(f) \le a_q\sqrt{n}$ for any $q \ge q_0$ by definition, and $\|G_n\left[\Delta_{q-1}(\cdot)A_{q-1}(\cdot)\right]\|_{\mathcal{G}_<}$ is a supremum taken over a finite collection of N_{q-1} functions in $\mathcal{G}_<$, we apply the preceding lemma to see that

$$\mathbb{E}^*\|G_n\left[\Delta_{q-1}(\cdot)A_{q-1}(\cdot)\right]\|_{\mathcal{G}_<,\infty} \le 48\left\{a_{q-1}(\log(N_{q-1})\vee 1) + 2^{-q+1}\sqrt{\log(N_{q-1})\vee 1}\right\}$$

for each $q > q_0$. Using the fact that

$$a_q(\log(N_q)\vee 1) \le a_q(\log(N_{q+1})\vee 1) = 2\times 2^{-(q+1)}\sqrt{\log(N_{q+1})\vee 1},$$

we can see that

$$\begin{aligned}\mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(\pi_q(f)-\pi_{q-1}(f)\right)A_{q-1}(f)\right]\right|\right] &\le 48\sum_{q=q_0}^{\infty}\left\{2\times 2^{-(q+1)}\sqrt{\log(N_{q+1})\vee 1} + 2^{-q}\sqrt{\log(N_q)\vee 1}\right\}\\ &\le 384\sum_{q=q_0}^{\infty}2^{-q}\sqrt{\log(N_q)\vee 1}.\end{aligned}$$

Now notice how $|\pi_{q_0}(f)| \leq a(\delta)\sqrt{n} \leq a_{q_0}\sqrt{n}$ and $\|\pi_{q_0}(f)\|_{P,2} \leq \delta$ for any $f \in \mathcal{G}_<$ (this is the only place where we use this assumption). As such,

$$\mathbb{E}^*\|G_n\left[\pi_{q_0}(\cdot)\right]\|_{\mathcal{G}_<,\infty} \leq 48\left\{a_{q_0}(\log(N_{q_0}) \vee 1) + \delta\sqrt{\log(N_{q_0}) \vee 1}\right\}$$

by a final application of the preceding lemma. Since

$$a_{q_0} = \frac{2^{-q_0}}{\sqrt{\log(N_{q_0+1} \vee 1)}} \leq \frac{2^{-q_0}}{\sqrt{\log(N_{q_0}) \vee 1}} \leq \frac{8\delta}{\sqrt{\log(N_{q_0}) \vee 1}},$$

it follows that

$$\begin{aligned}\mathbb{E}^*\|G_n\left[\pi_{q_0}(\cdot)\right]\|_{\mathcal{G}_<,\infty} \leq 432\delta\sqrt{\log(N_{q_0}) \vee 1} &\leq 108 \times 2^{-q_0}\sqrt{\log(N_{q_0}) \vee 1} \\ &\leq 384\sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\log(N_q) \vee 1}.\end{aligned}$$

Putting everything together, we can see that

$$\begin{aligned}\mathbb{E}^*\|G_n\|_{\mathcal{G}_<,\infty} &= \mathbb{E}^*\|G_n\left[\pi_{q_0}(\cdot)\right]\|_{\mathcal{G}_<,\infty} \\ &\quad + \mathbb{E}^*\left[\sup_{f\in\mathcal{F}}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(f - \pi_q(f)\right)B_q(f)\right]\right|\right] \\ &\quad + \mathbb{E}^*\left[\sup_{f\in\mathcal{G}_<}\left|\sum_{q=q_0+1}^{\infty} G_n\left[\left(\pi_q(f) - \pi_{q-1}(f)\right)A_{q-1}(f)\right]\right|\right] \\ &\leq (384 \times 3)\sum_{q=q_0}^{\infty} 2^{-q}\sqrt{\log(N_q) \vee 1} \\ &\leq (384 \times 3 \times 32)\left(J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta\right).\end{aligned}$$

□

10.3.3 Donsker's Theorem

Now we can proceed to prove Donsker's theorem. It relies on an application of theorem 10.2.1 where the preceding maximal inequality will be used to establish condition ii).

Theorem 10.3.2 (Donsker's Theorem)**.** *Let $(\Omega,\mathcal{H},\mathbb{P})$ be the underlying probability space and $\{X_n\}_{n\in N_+}$ a random sample drawn from a distribution P on the measurable space $(E,\mathcal{E})$. Suppose $\mathcal{F}$ is a collection of $\mathcal{E}$-measurable real-valued functions with $J_{[]}(1,\mathcal{F},L^2(P))<+\infty$. Then, $\mathcal{F}$ is P-Donsker.*

Proof. We start by establishing a couple of preliminary facts. Since $J_{[]}(1,\mathcal{F},L^2(P))<+\infty$, it must be the case that $N_{[]}(\epsilon,\mathcal{F},L^2(P))<+\infty$ for any $\epsilon<+\infty$. To see why, note that, if $N_{[]}(\epsilon,\mathcal{F},L^2(P))=+\infty$ for some $\epsilon>0$, then $N_{[]}(\eta,\mathcal{F},L^2(P))=+\infty$ for any $0<\eta<\epsilon$ as well, so that

$$J_{[]}(1,\mathcal{F},L^2(P))\geq\int_0^\epsilon\sqrt{\log N_{[]}(\eta,\mathcal{F},L^2(P))}d\eta=+\infty,$$

a contradiction. By implication, $\mathcal{F}\subset L^2(P)$ and $G_n:\Omega\to\ell^\infty(\mathcal{F})$.

Thus, to prove the theorem it suffices to show that:

- There exists a random function $G:\Omega\to\ell^\infty(\mathcal{F})$ such that, for any $(f_1,\cdots,f_k)\in\mathcal{F}$

$$(G_n(f_1),\cdots,G_n(f_k))\xrightarrow{d}(G(f_1),\cdots,G(f_k)).$$

- For any $m\in N_+$, there exists a partition $\{\mathcal{F}_i\}_{1\leq i\leq k}$ of $\mathcal{F}$ such that

$$\limsup_{n\to\infty}\mathbb{P}^*\left[\sup_{1\leq i\leq k}\sup_{f,g\in\mathcal{F}_i}|G_n(f)-G_n(g)|>2^{-m}\right]<2^{-m}.$$

Then, by theorem 10.2.1, we will be able to claim that there exists some semimetric ρ on $\mathcal{F}$ and some random function $\bar{G}:\Omega\to UC_b(\mathcal{F},\rho)$ such that $G_n\xrightarrow{d}\bar{G}$, which would prove that $\mathcal{F}$ is P-Donsker. Condition i) follows easily for the Lindberg-Levy CLT for iid square integrable random variables, where the target random function G is the random function corresponding to the Gaussian process on $\mathcal{F}$. It remains to show that condition ii) holds.

Define the collection $\mathcal{G}$ of $\mathcal{E}$-measurable real valued functions as

$$\mathcal{G} = \{f - g \mid f, g \in \mathcal{F}\}.$$

We can immediately see that $\mathcal{G} \subset L^2(P)$ because $L^2(P)$ is a vector space. Choose any $\epsilon > 0$ and let $[\ell_1, u_1], \cdots, [\ell_k, u_k]$ be the minimal collection of $\frac{\epsilon}{2}$-brackets in $L^2(P)$ that covers $\mathcal{F}$. For any $f - g \in \mathcal{G}$, suppose that $f \in [\ell_i, u_i]$ and $g \in [\ell_j, u_j]$ for some $1 \leq i, j \leq k$. Then,

$$\ell_i - u_j \leq f - g \leq u_i - \ell_j$$

and

$$\|(u_i - \ell_j) - (\ell_i - u_j)\|_{P,2} \leq \|u_i - \ell_i\|_{P,2} + \|u_j - \ell_j\|_{P,2} \leq \epsilon.$$

We have shown that the collection $\{[\ell_i - u_j,\ u_i - \ell_j] \mid 1 \leq i, j \leq k\}$ is a collection of ϵ-brackets in $L^2(P)$ that covers $\mathcal{G}$; since this collection consists of $\binom{k}{2}$ elements, we can see that

$$N_{[]}(\epsilon, \mathcal{G}, L^2(P)) \leq \binom{N_{[]}(\epsilon/2, \mathcal{F}, L^2(P))}{2} \leq N_{[]}(\epsilon/2, \mathcal{F}, L^2(P))^2.$$

In other words, the ϵ-bracketing number of $\mathcal{G}$ is bounded above by the square of the $\frac{\epsilon}{2}$-bracketing number of $\mathcal{F}$. By implication, for any $\delta \in (0, 1)$,

$$\begin{aligned} J_{[]}(\delta, \mathcal{G}, L^2(P)) &= \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{G}, L^2(P))} d\epsilon \\ &\leq \int_0^\delta \sqrt{2 \log N_{[]}(\epsilon/2, \mathcal{F}, L^2(P))} d\epsilon \\ &= 2 \int_0^{\frac{\delta}{2}} \sqrt{2 \log N_{[]}(\eta, \mathcal{F}, L^2(P))} d\eta \leq 2\sqrt{2} \times J_{[]}(\delta, \mathcal{F}, L^2(P)) < +\infty. \end{aligned}$$

Choose any $\delta \in (0, 1)$. Since $N_{[]}(\delta, \mathcal{F}, L^2(P)) < +\infty$, we can choose a finite collection $[\ell_1^\delta, u_1^\delta], \cdots, [\ell_{k_\delta}^\delta, u_{k_\delta}^\delta]$ of δ-brackets in $L^2(P)$ that covers $\mathcal{F}$. Now we construct the partition $\{\mathcal{F}_i^\delta\}_{1 \leq i \leq k_\delta}$ of $\mathcal{F}$ as

$$\begin{aligned} \mathcal{F}_1^\delta &= [\ell_1^\delta, u_1^\delta] \\ \mathcal{F}_j^\delta &= [\ell_j^\delta, u_j^\delta] \setminus \left(\bigcup_{i=1}^{j-1} [\ell_i^\delta, u_i^\delta] \right) \quad \forall 2 \leq j \leq k_\delta. \end{aligned}$$

It follows then that each $\mathcal{F}_j^\delta$ has L^2-diameter smaller than δ, so that, if we define

$$\mathcal{G}^\delta = \{f - g \mid f, g \in \mathcal{F}_j^\delta,\ 1 \leq j \leq k_\delta\},$$

each $f - g \in \mathcal{G}^\delta$ has L^2-norm bounded above by δ. $\mathcal{G}^\delta$ is therefore a collection of $\mathcal{E}$-measurable

real valued functions with envelope function

$$F^\delta = \max_{1\le i\le k_\delta}\left(\left|u_i^\delta\right| \vee \left|\ell_i^\delta\right|\right) \in L^2(P)$$

such that $\|h\|_{P,2} < \delta$ for any $h \in \mathcal{G}^\delta$. Furthermore, since $\mathcal{G}^\delta \subset \mathcal{G}$, it follows that

$$N_{[]}(\epsilon, \mathcal{G}^\delta, L^2(P)) \le N_{[]}(\epsilon, \mathcal{G}, L^2(P)) \le N_{[]}(\epsilon/2, \mathcal{F}, L^2(P))^2 < +\infty$$

for any $\epsilon > 0$. We can now apply the preceding theorem to see that there exists some $a(\delta) > 0$ such that

$$\mathbb{E}^*\|G_n\|_{\mathcal{G}^\delta,\infty} \le C\left\{J_{[]}(\delta, \mathcal{G}^\delta, L^2(P)) + \delta + \sqrt{n}\int_E \left(F^\delta \cdot I_{\{F^\delta > a(\delta)\sqrt{n}\}}\right) dP\right\}$$

for any $n \in N_+$, where $C > 0$ is a universal constant. Note how

$$F^\delta \cdot I_{\{F^\delta > a(\delta)\sqrt{n}\}} \le \frac{1}{\sqrt{n}a(\delta)}(F^\delta)^2 \cdot I_{\{F^\delta > a(\delta)\sqrt{n}\}},$$

so that

$$\sqrt{n}\int_E \left(F^\delta \cdot I_{\{F^\delta > a(\delta)\sqrt{n}\}}\right) dP \le \frac{1}{a(\delta)}\int_E \left[(F^\delta)^2 \cdot I_{\{F^\delta > a(\delta)\sqrt{n}\}}\right] dP.$$

Since F^δ is P-square integrable, the DCT now implies that the right hand side goes to 0 as $n \to \infty$. As such,

$$\limsup_{n\to\infty} \mathbb{E}^*\|G_n\|_{\mathcal{G}^\delta,\infty} \le C\left\{J_{[]}(\delta, \mathcal{G}^\delta, L^2(P)) + \delta\right\} \le C\left\{2\sqrt{2} \times J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta\right\}.$$

By the definition of $\mathcal{G}^\delta$ and Markov's inequality, we can see that, for any $m \in N_+$,

$$\mathbb{P}^*\left[\sup_{1\le i\le k_\delta}\sup_{f,g\in\mathcal{F}_i^\delta} |G_n(f-g)| > 2^{-m}\right] \le \frac{1}{2^{-m}}\mathbb{E}^*\left[\sup_{1\le i\le k_\delta}\sup_{f,g\in\mathcal{F}_i^\delta} |G_n(f-g)|\right] = \mathbb{E}^*\|G_n\|_{\mathcal{G}^\delta,\infty}$$

for any $n \in N_+$, so that

$$\limsup_{n\to\infty} \mathbb{P}^*\left[\sup_{1\le i\le k_\delta}\sup_{f,g\in\mathcal{F}_i^\delta} |G_n(f-g)| > 2^{-m}\right] \le \frac{C}{2^{-m}}\left\{2\sqrt{2} \times J_{[]}(\delta, \mathcal{F}, L^2(P)) + \delta\right\}.$$

By the DCT,

$$J_{[]}(\delta, \mathcal{F}, L^2(P)) = \int_0^1 \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L^2(P))} \cdot I_{[0,\delta]}(\epsilon) d\epsilon \to 0$$

as $\delta \to 0$, so that the right hand side of the above inequality goes to 0 as $\delta \to 0$. Choosing $\delta \in (0,1)$ small enough so that the right hand side is bounded above by 2^{-m} gives us a partition

$\{\mathcal{F}_i^\delta\}_{1\le i\le k_\delta}$ of $\mathcal{F}$ such that

$$\limsup_{n\to\infty} \mathbb{P}^* \left[\sup_{1\le i\le k_\delta} \sup_{f,g\in\mathcal{F}_i^\delta} |G_n(f-g)| > 2^{-m} \right] < 2^{-m},$$

and since we can find such a partition for any $m \in N_+$, condition ii) is satisfied.

□