

Closed-Form Augmentation for Locally Robust Bayesian Inference

Seung Hyun Kim

Department of Economics, Princeton University
Princeton, NJ 08544, United States

Email: seunghyun@princeton.edu

Abstract

Bayesian estimation of structural models often conditions on calibrated or plug-in values because joint estimation may be computationally infeasible. Treating these values as fixed ignores calibration uncertainty and can invalidate posterior inference under local misspecification. This paper proposes an augmentation principle that robustifies posterior inference and yields closed-form augmented likelihoods for exponential-family models and time-series models admitting a Gaussian Wold representation. The augmentation introduces auxiliary parameters that recover the nuisance score directions lost when plug-in values are treated as fixed, and in doing so provides a computationally tractable alternative to full-information joint estimation. A Bernstein–von Mises theorem establishes first-order robustness and efficiency of the resulting posterior. We apply the method to a dynamic discrete choice model of firm entry/exit and the Krusell–Smith heterogeneous-agent DSGE model. Monte Carlo experiments show that augmentation restores nominal posterior coverage under local misspecification while retaining the computational advantages of plug-in estimation.

Keywords: Bayesian inference; calibration; local misspecification; structural estimation; model augmentation

JEL Classification: C11, C13, C32, C33.

1 Introduction

The likelihoods of structural models are often highly non-linear functions of the model parameters. The resulting computational problems for estimation have been well-documented in the literature.

For dynamic discrete choice (DDC) models, works such as Imai et al. (2009) are dedicated to overcoming the computational burden of repeatedly solving dynamic programming blocks during estimation. Dynamic stochastic general equilibrium (DSGE) models in macroeconomics face a similar problem. In heterogeneous agent (HA) models, it is well known that the infinite-dimensional wealth distribution becomes an additional state variable that makes evaluating the likelihood multiple times in finite time extremely difficult (e.g. Winberry (2018), Auclert et al. (2021)).

Faced with a computational roadblock, many have turned to two-step estimation. Here, first-stage estimators, also referred to as plug-in estimates or calibrations, are computed for the computationally intensive parts of the model and used to form population moment conditions that identify the structural parameters of interest. Examples in microeconometrics include the Hotz and Miller (1993) CCP estimator for DDC models and Bajari et al. (2007)’s two-step estimator for dynamic games. This two-step program is prevalent when it comes to the estimation of DSGE models, where it is customary to calibrate a subset of the parameters before estimating the rest using likelihood-based methods (Smets and Wouters (2007); Acharya et al. (2023); Melosi et al. (2022)).

Under a two-step estimation procedure, misspecification in the plug-in values could invalidate inference based on the second-stage estimators of the parameters of interest—and turning back to full-information joint estimation may not be the solution in light of its computational complexity. This paper develops closed-form likelihood augmentations for parametric structural models that protect inference on parameters of interest from local misspecification in calibrated nuisance parameters. The structural model is enlarged along nuisance-score directions so that ordinary likelihood-based inference is automatically locally robust and efficient. For two broad classes of models for structural estimation, linearizing suitably chosen components of the likelihood delivers the desired augmentation.

Consider a structural model with parameters (β, η) , where η collects the block of nuisance parameters and we are interested in conducting inference about β . Suppose the nuisance parameters are fixed at some calibration or first-stage estimate η^* , and that inference on β proceeds as if η were known to be equal to η^* . If η^* were locally misspecified for η , then the usual point estimates of β would be asymptotically biased and the actual uncertainty associated with the estimation of β would be underestimated. This bias arises because the role of the η -score is ignored when estimating β .

To restore valid inference for β , we construct an augmented model that nests the original model with η fixed at η^* . The augmented model introduces an auxiliary parameter δ , and the model is designed so that the score with respect to (β, δ) exactly replicates the full score of the original model, which restores the missing η -score. δ enters into the likelihood of the augmented model in a much simpler manner than η does in the full model, while preserving the same first-order model geometry locally around the calibration η^* . Because the asymptotic properties of maximum likelihood and Bayes estimators are governed by this first-order geometry, the estimator of β

obtained from the augmented model inherits the same local robustness and efficiency properties of the full-information joint estimator of β .

The standard approach to robustifying estimators of β against local misspecification or first-stage estimation error is Neyman orthogonalization, which involves constructing an estimator for β using its efficient score instead of the plug-in score (e.g. Chernozhukov, Chetverikov, et al. (2018)). Mathematically, our proposed model augmentation is equivalent to Neyman orthogonalization in the sense that we engineer the augmentation so that the maximum likelihood estimator for β obtained from the augmented model is asymptotically linear in its efficient score. The important distinction is that model augmentation constructs a well-defined likelihood under which ordinary joint estimation of (β, δ) automatically orthogonalizes the estimator of β , while Neyman orthogonalization only modifies the estimating equation used to produce an estimator of β .

The idea of robustifying estimation by constructing a likelihood that incorporates least favorable score directions also has precedent. van der Laan and Rubin (2006), for instance, construct parametric submodels through an initial density estimate whose scores replicate the efficient influence direction, and estimate the associated auxiliary parameters by maximum likelihood. In the context of semiparametric models, Müller and Norets (2026) observe that Bayesian inference can be locally robustified by augmenting a baseline model with auxiliary parameters that restore missing score directions.

While the augmentation of parametric structural models proposed in this paper is based on a similar insight, the closed-form construction is an important distinction. As noted in Müller and Norets (2026), in general it is difficult to find well-defined statistical models that satisfy the desired score conditions, since they are formulated in terms of the derivative instead of the level of the likelihood and produce an unknown constant of integration that depends on the parameters (β, δ) . Their proposed solution makes use of auxiliary latent variables and acceptance sampling to avoid calculating this unknown parameter-dependent constant of integration. This procedure may be difficult to apply for large scale structural models where evaluating the baseline likelihood is already computationally expensive, as in DSGE models where even a single Gaussian likelihood evaluation may require large Toeplitz-matrix operations. A related construction problem arises in van der Laan and Rubin (2006): while a generic exponential tilting submodel can reproduce a desired score direction, its normalizing constant need not be available in closed form, so tractable augmentations are constructed on a problem-specific basis.

By comparison, the proposed linearization rules for exponential family and Gaussian structural vector moving average (SVMA) models deliver well-defined likelihoods in closed form. Because the augmented models themselves lie in the same parametric class as the original models under the proposed linearization rule, we can make use of standard off-the-shelf samplers when sampling from the augmented model. In addition to allowing us to sidestep the computational overhead from

Müller and Norets (2026)’s auxiliary variable and acceptance sampling process, the closed-form constructions also preserve the principal computational advantage of plug-in estimation: namely, that expensive equilibrium objects and their local derivatives can be computed once prior to posterior sampling, rather than being recomputed whenever the nuisance parameters change.

There are two further distinctions from the existing augmentation literature. First, in augmentations of semiparametric models, δ indexes the degree to which the baseline model must move in the least favorable score direction to match the data and generally does not correspond to some finite-dimensional nuisance parameter. In a full parametric augmentation, by contrast, δ is locally interpretable as $\eta - \eta^*$, permitting inference on and updating of the calibrated nuisance parameters themselves. Second, the constructions in Müller and Norets (2026) and van der Laan and Rubin (2006) target the least favorable score directions associated with a particular parameter or functional of interest. Restoration of the entire finite-dimensional nuisance score supports locally robust inference for any smooth $H(\beta, \eta)$, including functions of the calibrated nuisance parameters themselves.

The close connection between δ and η is particularly useful in the context of updating calibrations for DSGE models. Often, steady-state parameters are calibrated using estimates obtained from micro-level data even though the aggregate macroeconomic data used to estimate the model also contain information about these parameters¹. Directly combining these two sources of information is difficult because macroeconomic theory typically specifies a likelihood for the aggregate variables but not a joint data generating process for the aggregate and micro-level data². Full augmentation provides an intermediate approach. Micro-level information can be incorporated through η^* and the prior distribution for δ , while posterior inference on δ allows the aggregate data to revise the calibration. Because this updating takes place through the auxiliary parameter, the structural model need not be repeatedly solved at different values of η during posterior sampling.

For cases where the nuisance parameter is high-dimensional and we are only interested in conducting inference on β , we also develop a parsimonious augmentation that introduces only as many auxiliary parameters as parameters of interest in a manner that restores only the η -score directions relevant for the efficient estimation of β . These relevant nuisance score directions are precisely the least favorable score directions in the semiparametric setting of Müller and Norets (2026). The parsimonious augmentation is therefore the finite-dimensional parametric analogue of the least favorable direction construction in Müller and Norets (2026), while full augmentation is specific to the finite-dimensional nuisance setting and additionally permits inference on and updating of the nuisance parameters themselves. Consequently, parsimonious augmentation has

¹Liu, Plagborg-Møller, and Tan (2026), for instance, attempt to leverage the information in aggregate macro data to form estimates for the discount rate and intertemporal elasticity of substitution in a two-asset HANK model.

²Liu and Plagborg-Møller (2023) address this difficulty by resorting to a numerical construction of the joint likelihood of macro and micro data.

a computational advantage over full augmentation but also a drawback in that the relationship between δ and η is not as simple as in the fully augmented model.

The proposed augmentation procedure is first applied to the class of DDC models. A model of firm entry/exit demonstrates that inference under the augmented model is valid when the baseline model fails, and that model augmentation possesses clear computational gains compared to full-information joint estimation of the model parameters. The computational gains in particular are substantial: even with a modestly fine discretization of the continuous state variable, sampling from the augmented model costs only slightly more than the baseline, whereas joint sampling of (β, η) is hundreds or thousands of times more expensive. Next, we turn to the Krusell and Smith (1998) model to show how the augmented model helps restore valid inference for parameters of interest such as impulse responses under local misspecification of the capital share α , while also letting us systematically update the capital share calibration using the information in aggregate output data. That the proposed augmentation allows us to study the sensitivity of local calibration error to model estimates is one of its greatest strengths. The potential problems arising from fixed calibrations have long been recognized by macroeconomists, but the computational costs of estimating HA models have resulted in only a few studies (e.g. Jørgensen (2023)) seeking to formally investigate the effects fixing calibrations has on the conclusions drawn from structural macroeconomic models.

A secondary technical contribution establishes a Bernstein–von Mises (BvM) theorem under conditions convenient for stationary Gaussian models. Standard parametric BvM results typically impose a uniform local asymptotic normality (LAN) condition that is less straightforward to verify for dependent Gaussian processes, whose scores are quadratic forms rather than sums of iid observations. We replace uniform LAN with an L^1 condition that is readily verified from primitive conditions for both stationary Gaussian models and models with iid data.

Outline of Paper. The paper is organized as follows. Section 2 presents a simple motivating example for the procedure before deriving the general augmentation principle. In section 3, we provide a closed-form augmentation for exponential family models and apply it to DDC models. Section 4 does the same for Gaussian SVMA models and includes an application to a heterogeneous agent DSGE model. Finally, section 5 concludes.

2 The Augmentation Principle for Parametric Models

2.1 A Simple Motivating Example

Consider a Gaussian non-linear regression model

$$y_i = x_i\beta + z_i g(\eta) + \epsilon_i, \quad \epsilon_i \mid x_i, z_i \sim \mathcal{N}[0, 1],$$

where the DGP under (β, η) is denoted by $P_{\beta, \eta}$. Here the function $g(\cdot)$ is a smooth non-linear function that represents a model-implied object that may be costly to evaluate as η varies. Although this computational burden is suppressed in the simple example, in the structural applications considered below evaluating the analogue of $g(\eta)$ can require repeatedly solving an equilibrium model. Let x_i, z_i be mean zero regressors. Denote by Q_{xx} , Q_{zx} and Q_{zz} the variance of x_i , the covariance of z_i, x_i and the variance of z_i .

The plug-in approach when conducting likelihood-based inference on the parameter of interest β is to fix the nuisance parameter η at some calibration or first-stage estimate η^* . Having fixed η at η^* , the maximum likelihood estimator $\hat{\beta}_n$ of β is exactly the OLS estimator from regressing $y_i - z_i g(\eta^*)$ on x_i . Under P_{β_0, η^*} , $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges weakly to a normal distribution centered at 0 with variance Q_{xx}^{-1} . However, under local misspecification, so that the true value of η is local to η^* in the sense that $\eta_0 = \eta^* + \frac{h}{\sqrt{n}}$ for some $h \neq 0$, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xRightarrow{P_{\beta_0, \eta_0}} \mathcal{N}\left[\frac{Q_{xz}}{Q_{xx}} g'(\eta^*) h, \frac{1}{Q_{xx}}\right],$$

so that the plug-in estimator of β is asymptotically biased.

The first-best choice of an estimator of β that is asymptotically efficient and locally robust to misspecification in η^* is the joint maximum likelihood estimator, where the likelihood is maximized with respect to both β and η . The efficiency bound for β is $(Q_{xx} - Q_{zx}^2/Q_{zz})^{-1} > Q_{xx}^{-1}$, so that the plug-in maximum likelihood estimator for β is erroneously precise for the wrong target if η^* is locally misspecified.

Since joint estimation of β and η is computationally expensive, we turn to the augmented model

$$y_i - g(\eta^*) z_i = x_i \beta + g'(\eta^*) z_i \delta + \epsilon_i, \quad \epsilon_i \mid x_i, z_i \sim \mathcal{N}[0, 1].$$

The augmented model nests the original model with η fixed at η^* , as we can see by putting δ equal to 0. Joint estimation of (β, δ) in the augmented model is computationally feasible where joint estimation of (β, η) in the original model was not because we have replaced the complicated non-linear manner in which the original likelihood depends on η with a simpler linear relationship

for δ in the augmented likelihood. The augmented model replaces repeated evaluations of $g(\eta)$ away from η^* with the local objects $g(\eta^*)$ and $g'(\eta^*)$, which can be computed before estimation.

Now we can see that the maximum likelihood estimator of (β, δ) in the augmented model is exactly the OLS estimator from regressing $y_i - g(\eta^*)z_i$ on $R_i = (x_i, w_i)$ where $w_i = g'(\eta^*)z_i$. Letting the estimators be denoted $\hat{\beta}_n^{\mathcal{A}}, \hat{\delta}_n^{\mathcal{A}}$, under local misspecification in η of the form $\eta_0 = \eta^* + \frac{h}{\sqrt{n}}$, we can see that

$$\sqrt{n} \begin{bmatrix} \hat{\beta}_n^{\mathcal{A}} - \beta_0 \\ \hat{\delta}_n^{\mathcal{A}} \end{bmatrix} = \begin{pmatrix} 0 \\ h \end{pmatrix} + Q_R^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n R_i \epsilon_i + o_p(1)$$

where Q_R is the covariance matrix of R_i . Restricting our attention to $\hat{\beta}_n^{\mathcal{A}}$ reveals that

$$\sqrt{n}(\hat{\beta}_n^{\mathcal{A}} - \beta_0) \xrightarrow{P_{\beta_0, \eta_0}} \mathcal{N} \left[0, \frac{1}{Q_{xx} - Q_{zx}^2/Q_{zz}} \right],$$

so the MLE $\hat{\beta}_n^{\mathcal{A}}$ of β from the augmented model is locally robust to misspecification in η^* and attains its full-information efficiency bound. Similarly, if we define $\hat{\eta}_n = \hat{\delta}_n^{\mathcal{A}} + \eta^*$, then

$$\sqrt{n}(\hat{\eta}_n - \eta_0) \xrightarrow{P_{\beta_0, \eta_0}} \mathcal{N} \left[0, \frac{1}{g'(\eta^*)^2(Q_{zz} - Q_{xz}^2/Q_{xx})} \right],$$

so that $\hat{\eta}_n$ serves as an efficient estimator of η . As such, model augmentation not only restores valid inference for β under local misspecification, but also delivers an estimator for the nuisance parameter η .

The local robustness of $\hat{\beta}_n^{\mathcal{A}}$ to misspecification in η^* arises precisely because the augmented model is engineered to restore the score directions that are missing from plug-in estimation. By the Gaussianity of the errors, the per-observation score of (β, η) in the original model is given as

$$\dot{\ell}(\beta, \eta) = \begin{pmatrix} x_i \\ g'(\eta)z_i \end{pmatrix} \{y_i - x_i\beta - g(\eta)z_i\},$$

but the plug-in estimator is based on only the first entry $x_i(y_i - x_i\beta - g(\eta)z_i)$. In the augmented model the per-observation score of (β, δ) at $(\beta_0, 0)$ is exactly equal to $\dot{\ell}(\beta_0, \eta^*)$, so that the missing η -score is restored when conducting inference on β . The augmentation procedure can therefore be described as replacing a complicated non-linear relationship in the original model with a well-chosen linearization that preserves the score of the original model, so that the robustness properties of joint estimation can be recreated without having to pay all of the computational costs. Importantly, the linearization preserves the class of the original statistical model: both the original and augmented specifications are Gaussian regressions. Hence the augmented score corresponds

immediately to a properly normalized likelihood. This illustrates how restricting attention to model classes with known normalizing functions makes the abstract score-matching principle operational.

2.2 Model Setup and Notation

Now we extend the ideas of the simple example to a more general setting of parametric structural models. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space and $\{X_t\}_{t \in \mathbb{Z}}$ a doubly infinite stochastic process taking values in $\mathbb{R}^n$. Denote $\mathbf{X}_T = (X'_1, \dots, X'_T)'$ for each $T \in \mathbb{N}_+$. We let $\ell_T(\theta \mid x)$ be the T -sample log-likelihood of $\mathbf{X}_T = x$ evaluated at the parameter vector $\theta \in \Theta$.

Suppose θ can be decomposed into three parts. First, we have the k_β -dimensional parameter of interest β . Next, we let γ be the k_γ -dimensional parameter that is normally estimated using likelihood-based methods but which is not the immediate object of our attention; γ is effectively an estimable nuisance parameter. Finally, there is an k_η -dimensional nuisance parameter η that is computationally expensive to estimate. Due to this computational cost, we wish to avoid estimating η jointly with (β, γ) using likelihood-based methods. Instead, some value η^* is plugged into η . In a heterogeneous-agent New Keynesian (HANK) model, for instance, β may collect the Taylor rule parameters and the slope of the Phillips curve; γ the parameters governing the persistence and magnitude of the exogenous shocks; and η preference parameters that determine the steady state of the economy. Below, we use the following notation:

$$\theta = (\beta, \gamma, \eta), \quad \nu = (\beta, \gamma), \quad \zeta = (\gamma, \eta). \quad (2.1)$$

We are interested in the estimation of some function $H(\theta)$ of the parameters. In particular, we may be interested in the case $H(\theta) = \beta$, where $H(\cdot)$ is a function that selects only the parameters of interest. Errors in η^* of order $T^{-1/2}$ are our focus because errors of a smaller order are asymptotically negligible, while errors of a higher order can be detected with probability approaching 1 on the basis of the likelihood.

Here we adopt a local misspecification framework for η , that is, η^* is treated as a fixed value that may or may not be equal to the true value η_0 of the nuisance parameter. In this case, we want to robustify the model against first order calibration error of the form $\eta_0 = \eta^* + \frac{h}{\sqrt{T}}$, where $h \in \mathbb{R}^{k_\eta}$ is some unknown direction in which η^* is locally misspecified. As such, the natural theoretical framework is Le Cam asymptotic theory. An alternative is to take a first-stage estimation approach, where η^* is viewed as an estimator of the true value η_0 that is generally sample-size dependent and often random. It turns out that the augmented model developed under the local misspecification framework also furnishes local robustness against first-stage estimation error—the theoretical minutiae are discussed at the end in section 2.7.

Let $\theta_0 = (\beta_0, \gamma_0, \eta_0)$ be the true value of the parameter θ . We denote by $\dot{\ell}_{T,\theta}(d \mid \mathbf{X}_T)$ the

T -sample score function evaluated at the parameter value $d \in \Theta$. Let $\mathbb{I}_*$ denote the positive definite Fisher information matrix at the baseline parameter values $\theta^* = (\nu_0, \eta^*)$. The score function and the information matrix can be partitioned as

$$\dot{\ell}_{T,\theta}(d \mid \mathbf{X}_T) = \begin{pmatrix} \dot{\ell}_{T,\beta}(d \mid \mathbf{X}_T) \\ \dot{\ell}_{T,\gamma}(d \mid \mathbf{X}_T) \\ \dot{\ell}_{T,\eta}(d \mid \mathbf{X}_T) \end{pmatrix}, \quad \mathbb{I}_* = \begin{pmatrix} \mathbb{I}_{*,\beta\beta} & \mathbb{I}_{*,\beta\gamma} & \mathbb{I}_{*,\beta\eta} \\ \mathbb{I}_{*,\gamma\beta} & \mathbb{I}_{*,\gamma\gamma} & \mathbb{I}_{*,\gamma\eta} \\ \mathbb{I}_{*,\eta\beta} & \mathbb{I}_{*,\eta\gamma} & \mathbb{I}_{*,\eta\eta} \end{pmatrix}$$

conformably along the dimensions of β , γ and η . In accordance with the notation defined in 2.1, we let $\dot{\ell}_{T,\nu}(d \mid \mathbf{X}_T)$ be the vector collecting the β - and γ -scores, and let $\dot{\ell}_{T,\zeta}(d \mid \mathbf{X}_T)$ be defined in a similar manner. The blocks of the information matrix are denoted the same way.

Suppose that the model admits a local asymptotic normal (LAN) expansion at θ^* and that the T -sample score is asymptotically normal at the parametric rate:

$$\frac{1}{\sqrt{T}} \dot{\ell}_{T,\theta}(\theta^* \mid \mathbf{X}_T) \xrightarrow{P_{T,\theta^*}} \mathcal{N}[O_{k \times 1}, \mathbb{I}_*]. \quad (2.2)$$

$o_{P_{T,\theta^*}}(1)$ and $\xrightarrow{P_{T,\theta^*}}$ indicate convergence in probability to 0 and weak convergence under the distributions of $\mathbf{X}_T$ at the baseline parameter θ^* .

2.3 Bias under Local Misspecification and the Efficient Influence Function

Suppose that we are interested in conducting robust and efficient inference on some k_H -dimensional function $H(\theta)$ of the model parameters. Let the naive maximum likelihood estimator of ν obtained from plugging in η^* in place of η be denoted by $\hat{\nu}_T^{nv}$, and by the invariance of maximum likelihood estimation $H(\hat{\nu}_T^{nv}, \eta^*)$ is the maximum likelihood estimator of $H(\theta)$ in this model. Assuming that $H(\cdot)$ is sufficiently smooth to apply the delta method, a first order expansion shows us that

$$\sqrt{T} \{H(\hat{\nu}_T^{nv}, \eta^*) - H(\theta^*)\} = \dot{H}_{*,\nu} \mathbb{I}_{*,\nu\nu}^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\nu}(\theta^* \mid \mathbf{X}_T) + o_{P_{T,\theta^*}}(1)$$

where $\dot{H}_{*,\nu}$ is the derivative of $H(\cdot)$ with respect to ν at θ^* . Suppose that η^* is locally misspecified, so that $\eta_0 = \eta^* + \frac{h}{\sqrt{T}}$ for some $h \in \mathbb{R}^{k_\eta}$. Then, an application of Le Cam's third lemma shows us that $H(\hat{\nu}_T^{nv}, \eta^*)$ is asymptotically biased under local misspecification with bias equal to $(\dot{H}_{*,\nu} \mathbb{I}_{*,\nu\nu}^{-1} \mathbb{I}_{*,\nu\eta} - \dot{H}_{*,\eta}) h$. This bias is non-zero unless $\dot{H}_{*,\nu} \mathbb{I}_{*,\nu\nu}^{-1} \mathbb{I}_{*,\nu\eta} = \dot{H}_{*,\eta}$, which is generally untrue.

The asymptotic bias in the maximum likelihood estimator of $H(\theta)$ arises due to the absence of the η -score in the first order expansion of $\sqrt{T} \{H(\hat{\nu}_T^{nv}, \eta^*) - H(\theta^*)\}$. A locally robust and efficient estimator of $H(\theta)$ is then one whose first order expansion that restores the missing η -score

direction—it involves the derivative $\dot{H}_*$ of $H(\cdot)$ with respect to all the model parameters at the possibly misspecified baseline θ^* , alongside the full θ -score. Consider an estimator $\hat{H}_T$ of $H(\theta)$ that can be expressed as

$$\sqrt{T} \{ \hat{H}_T - H(\theta^*) \} = \underbrace{\dot{H}_* \mathbb{I}_*^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\theta}(\theta^* \mid \mathbf{X}_T)}_{\psi_{T,H}(\theta^* \mid \mathbf{X}_T)} + o_{P_{T,\theta^*}}(1).$$

Another application of Le Cam’s third lemma shows us that $\hat{H}_T$ is asymptotically unbiased but has a larger asymptotic variance compared to the naive estimator $H(\hat{\nu}_T^{nv}, \eta^*)$. In the literature $\psi_{T,H}(\theta^* \mid \mathbf{X}_T)$ is referred to as the efficient influence function and $\dot{H}_* \mathbb{I}_*^{-1} \dot{H}_*$ as the efficiency bound for $H(\theta)$.

2.4 Recovering the Efficient Estimator via Model Augmentation

The classic frequentist approach to constructing the efficient estimator of $H(\theta)$ is to form the one-step estimator

$$\hat{H}_T = H(\hat{\nu}_T^{nv}, \eta^*) + \frac{1}{\sqrt{T}} \psi_{T,H}(\hat{\nu}_T^{nv}, \eta^* \mid \mathbf{X}_T).$$

Bickel et al. (1993) show that, under additional differentiability and regularity conditions, $\hat{H}_T$ is indeed asymptotically linear in the efficient influence function $\psi_{T,H}(\theta^* \mid \mathbf{X}_T)$ and thus robust to local misspecification in η .

However, the one-step estimation approach may be awkward to implement in a Bayesian context. Leaving aside joint estimation of $\theta = (\nu, \eta)$, whose computational infeasibility is the motivation behind this entire exercise, one possible approach to ensure that the posterior targets the efficient score instead of the ordinary score is to construct a pseudo-likelihood for $H(\theta)$ whose score is exactly the efficient score (see Greco et al. (2008), Chernozhukov and Hong (2003) for Bayesian inference under pseudo-likelihoods). In general pseudo-likelihoods need not be proper likelihoods, and thus the resulting quasi-posterior does not satisfy the likelihood principle.

An attractive alternative is to embed the baseline model into a larger model and sample from the posterior implied by the larger model (e.g. Andrade and O’Hagan (2011) embeds the baseline model into a larger class of models with fat tails to better deal with outliers). Unlike pseudo-likelihood procedures based on general estimating criteria, the embedding approach preserves likelihood-based Bayesian inference. Because the likelihood of the larger model is the likelihood of a well-defined statistical model that nests the baseline model, posterior inference satisfies the likelihood principle. The augmentation proposed in this paper is a natural extension of this embedding idea.

Given a plug-in value η^* for η , we refer to the baseline model as the model where η is fixed at η^* and potential local misspecification is ignored. Consider augmenting this model with a k_δ -dimensional auxiliary parameter δ , so that $\mathbf{X}_T$ now depends on the parameters (ν, δ) with log-likelihood denoted by $\ell_T^{\mathcal{A}}(\nu, \delta \mid \mathbf{X}_T)$. Suppose

$$\ell_T^{\mathcal{A}}(\nu, \delta_0 \mid \mathbf{X}_T) = \ell_T(\nu, \eta^* \mid \mathbf{X}_T) \quad \forall \nu \quad (2.3)$$

$$\dot{\ell}_{T,\nu,\delta}^{\mathcal{A}}(\nu_0, \delta_0 \mid \mathbf{X}_T) := \begin{pmatrix} \frac{\partial \ell_T^{\mathcal{A}}(\nu_0, \delta_0 \mid \mathbf{X}_T)}{\partial \nu} \\ \frac{\partial \ell_T^{\mathcal{A}}(\nu_0, \delta_0 \mid \mathbf{X}_T)}{\partial \delta} \end{pmatrix} = \begin{pmatrix} \dot{\ell}_{T,\nu}(\nu_0, \eta^* \mid \mathbf{X}_T) \\ \mathbf{B}_0' \dot{\ell}_{T,\zeta}(\nu_0, \eta^* \mid \mathbf{X}_T) \end{pmatrix} \quad (2.4)$$

where $\delta_0 = O_{k_\delta \times 1}$ and $\mathbf{B}_0$ is some $k_\zeta \times k_\delta$ matrix to be specified below. The first equation ensures that the baseline model is a special case of the new model when δ is equal to δ_0 , so that the augmented model is a true augmentation of the original model. The limiting Fisher information of the augmented model at (ν_0, δ_0) is a function of the Fisher information $\mathbb{I}_*$ at θ^* :

$$\mathbb{I}_*^{\mathcal{A}} = \begin{pmatrix} \mathbb{I}_{*,\nu\nu} & \mathbb{I}_{*,\nu\zeta} \mathbf{B}_0 \\ \mathbf{B}_0' \mathbb{I}_{*,\zeta\nu} & \mathbf{B}_0' \mathbb{I}_{*,\zeta\zeta} \mathbf{B}_0 \end{pmatrix}. \quad (2.5)$$

Condition (2.4) imposes restrictions only on the score of the augmented model. Suppose η enters into the likelihood of the full model in a complicated non-linear manner that precludes estimation in finite time. Then, we can construct an augmented model that sidesteps the computational problem by formulating the augmented likelihood as a much simpler function of δ but where the score structure of the full model is nonetheless preserved.

Below we consider two different specifications for $\mathbf{B}_0$ and how they are related to our problem.

2.4.1 Full Augmentation

Suppose first that $k_\delta = k_\eta$ and $\mathbf{B}_0' = \begin{pmatrix} O_{k_\eta \times k_\gamma} & I_{k_\eta} \end{pmatrix}$. Under this choice of $\mathbf{B}_0$, the second equation specifies the score function of the augmented model as exactly that of the original model, and in particular $\mathbb{I}_*^{\mathcal{A}} = \mathbb{I}_*$. In other words, the augmented model restores the entire η -score, and hence the augmentation under this choice of $\mathbf{B}_0$ is called a full augmentation. Under full augmentation the maximum likelihood estimators of (ν, δ) can be used to form an estimator of $H(\theta)$ that is asymptotically linear in the efficient influence function, and therefore locally robust and asymptotically efficient.

Proposition 1. *Let $\hat{\nu}_T^{\mathcal{A}}$ and $\hat{\delta}_T^{\mathcal{A}}$ be the maximum likelihood estimators of ν, δ in a fully augmented*

model that satisfies conditions (2.3) and (2.4). Then, defining

$$H^{\mathcal{A}}(\nu, \delta) = H(\nu, \eta^*) + \frac{\partial H(\nu, \eta^*)}{\partial \eta'} \delta, \quad (2.6)$$

for any (ν, δ) , we have

$$\sqrt{T} \{ \hat{H}^{\mathcal{A}}(\hat{\nu}_T^{\mathcal{A}}, \hat{\delta}_T^{\mathcal{A}}) - H(\theta^*) \} = \psi_{T,H}(\theta^* | \mathbf{X}_T) + o_{P_{T,\theta^*}}(1).$$

Proof. Under the usual regularity conditions for maximum likelihood estimators, the T -sample MLEs of (ν, δ) in the augmented model will be asymptotically linear in the augmented T -sample score, which under full augmentation is simply the T -sample score of the original model evaluated at θ^* . Then, applying the delta method to the mapping $H(\cdot, \eta^*)$ and the continuous mapping theorem to the mapping $\theta \mapsto \frac{\partial H(\theta)}{\partial \eta'}$ yields the desired expansion. $\square$

The above proposition suggests the following two-step procedure for constructing a posterior sample of $H(\theta)$:

- 1) Generate draws $\{\nu^{(i)}, \delta^{(i)}\}_{1 \leq i \leq N}$ of (ν, δ) from the augmented model.
- 2) Generate a posterior sample of $H(\theta)$ as $\left\{ H(\nu^{(i)}, \eta^*) + \frac{\partial H(\nu^{(i)}, \eta^*)}{\partial \eta'} \delta^{(i)} \right\}_{1 \leq i \leq N}$.

In light of the asymptotic equivalence between maximum likelihood and Bayesian estimation, this procedure will generate posterior draws that possess the same efficiency properties as the maximum likelihood estimator in the limit.

One of the advantages of full augmentation is that the auxiliary parameter δ has a clear interpretation. Suppose that $H(\theta) = \eta$, so that the nuisance parameters themselves are the objects of interest. Then, following the procedure laid out in section 2.4.1 produces a posterior sample of $H(\theta)$ that is simply

$$\{\eta^* + \delta^{(i)}\}_{1 \leq i \leq N}.$$

Often, the calibration η^* is pre-estimated from a sample independent of the data under which the model likelihood is evaluated. Under our assumption that $\mathbb{I}_*$ is non-singular, the second stage sample contains non-negligible information about η that is discarded if η is treated as fixed in the second stage. In this context, the augmentation procedure can be understood as a formal way to incorporate the information about η contained in the second stage sample without having to estimate η jointly with $\nu = (\beta, \gamma)$, and δ as the specific direction in which the baseline calibration is updated.

2.4.2 Parsimonious Augmentation

Now suppose that $H(\theta) = \beta$ and that $k_\delta = k_\beta$ and $\mathbf{B}_0 = \mathbb{I}_{*,\zeta}^{-1} \mathbb{I}_{*,\zeta} \beta$, the coefficient from regressing the β -score onto the scores for γ and η . From standard efficiency theory for parametric models we know that the efficient score function for $H(\theta)$ is given as

$$\tilde{\ell}_{T,\beta}(\theta^* \mid \mathbf{X}_T) = \dot{\ell}_{T,\beta}(\theta^* \mid \mathbf{X}_T) - \mathbf{B}_0' \dot{\ell}_{T,\zeta}(\theta^* \mid \mathbf{X}_T).$$

From the construction in Müller and Norets (2026), it now follows almost immediately that, because

$$\dot{\ell}_{T,\beta}(\theta^* \mid \mathbf{X}_T) - \tilde{\ell}_{T,\beta}(\theta^* \mid \mathbf{X}_T) = \mathbf{B}_0' \dot{\ell}_{T,\zeta}(\theta^* \mid \mathbf{X}_T),$$

we will be able to obtain a posterior sample of β that is efficient and robust to local misspecification in η in large samples by estimating the augmented model satisfying equations (2.3) and (2.4).

In this alternative augmentation, the dimension of δ is k_β instead of k_η , suggesting that we can arrive at a much more parsimonious augmentation by using this construction if $k_\beta < k_\eta$. The parsimonious augmentation is valuable, for example, in DDC models where the nuisance parameters governing the dynamics of the observed state are often high-dimensional and not the immediate object of interest. The directional matrix $\mathbf{B}_0$ is exactly the link that allows us to efficiently estimate β using only k_β augmentations, by summarizing the information in the score for η that is relevant for the estimation of β . Mathematically, $\mathbf{B}_0$ is a device that selects the direction in the linear span of $\dot{\ell}_{T,\zeta}(\theta^* \mid \mathbf{X}_T)$ that is least favorable to the estimation of β , that is, induces the highest loss of information that comes with not knowing the value of η .

2.5 Asymptotic Results for Augmented Posterior

Consider the general augmented model with a k_δ -dimensional auxiliary parameter, where the log-likelihood of $\mathbf{X}_T$ satisfies the conditions (2.3) and (2.4) for some $\mathbf{B}_0 \in \mathbb{R}^{k_\zeta \times k_\delta}$. We formally establish that the posterior distribution of $\sqrt{T} \{H^\mathcal{A}(\nu, \delta) - H(\theta_0)\}$ given $\mathbf{X}_T$ converges in total variation distance to a normal distribution centered at $\dot{H}_*(\mathbb{I}_*^\mathcal{A})^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\nu,\delta}^\mathcal{A}(\nu_0, \delta_0 \mid \mathbf{X}_T)$ with variance equal to $\dot{H}_*(\mathbb{I}_*^\mathcal{A})^{-1} \dot{H}_*'$, where $H^\mathcal{A}(\nu, \delta)$ is defined as

$$H^\mathcal{A}(\nu, \delta) = H(\nu, \eta^*) + \frac{\partial H(\nu, \eta^*)}{\partial \zeta'} \mathbf{B}_0 \delta.$$

Note the distinction with the definition in equation (2.6) due to the theorem needing to accommodate general forms of $\mathbf{B}_0$. We define $R_T^{\mathcal{A}}(h)$ as the LAN residual

$$R_T^{\mathcal{A}}(h) = \ell_T^{\mathcal{A}}\left(\vartheta_0 + \frac{h}{\sqrt{T}} \mid \mathbf{X}_T\right) - \ell_T^{\mathcal{A}}(\vartheta_0 \mid \mathbf{X}_T) - h' \frac{1}{\sqrt{T}} \dot{\ell}_{T,\vartheta}^{\mathcal{A}}(\vartheta_0 \mid \mathbf{X}_T) + \frac{1}{2} h' \mathbb{I}_*^{\mathcal{A}} h$$

for any $h \in \mathbb{R}^{k_\nu + k_\delta}$, where $\vartheta_0 = (\nu_0, \delta_0)$. We let $q_T(\cdot)$ be the possibly sample-size dependent Lebesgue density of the prior on (ν, δ) and $\Pi_{T,H}^h[x, \cdot]$ the posterior distribution of $\sqrt{T} \{H^{\mathcal{A}}(\nu, \delta) - H(\theta_0)\}$ given $\mathbf{X}_T = x$. Finally, for any $h \in \mathbb{R}^{k_\eta}$ let $\Phi_{T,H}^h[x, \cdot]$ be the normal distribution centered at

$$\dot{H}_* (\mathbb{I}_*^{\mathcal{A}})^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\vartheta}^{\mathcal{A}}(\vartheta_0 \mid x) - \dot{H}_{*,\eta} h$$

and variance $\dot{H}_* (\mathbb{I}_*^{\mathcal{A}})^{-1} \dot{H}_*'$.

Assumption 1 (BvM Regularity). *The augmented model satisfies the following conditions:*

- (i) *The parameter space for $\vartheta = (\nu, \delta)$ is open and convex. The prior density q_T is strictly positive at ϑ_0 for sufficiently large T , and for any compact set $K \subset \mathbb{R}^{k_\nu + k_\delta}$,*

$$\sup_{h \in K} \left| \log \frac{q_T(\vartheta_0 + h/\sqrt{T})}{q_T(\vartheta_0)} \right| = o(1) \quad \text{as } T \rightarrow \infty.$$

- (ii) *For every compact set $K \subset \mathbb{R}^{k_\nu + k_\delta}$,*

$$\sup_{h \in K} \mathbb{E}_{\vartheta^*} [|R_T^{\mathcal{A}}(h)|] = o(1), \quad \sup_{h \in K} \left| \ell_T^{\mathcal{A}}\left(\vartheta_0 + \frac{h}{\sqrt{T}} \mid \mathbf{X}_T\right) - \ell_T^{\mathcal{A}}(\vartheta_0 \mid \mathbf{X}_T) \right| = O_{P_{T,\vartheta^*}}(1).$$

- (iii) *For every sequence $M_T \rightarrow \infty$,*

$$\Pi_T \left(\sqrt{T} |\vartheta - \vartheta_0| > M_T \mid \mathbf{X}_T \right) \xrightarrow{L^1} 0.$$

- (iv) *The mapping $H^{\mathcal{A}}(\cdot)$ is continuously differentiable in a neighborhood of ϑ_0 , with derivative $\dot{H}_*^{\mathcal{A}}$ of full row rank.*

Assumption (i) requires the prior to be asymptotically flat on neighborhoods of radius $T^{-1/2}$ around the true parameter. This permits sample-size-dependent priors, provided their concentration is sufficiently slow that they do not affect the local Gaussian approximation of the posterior. The assumption is automatically satisfied by any fixed prior density q that is continuous and strictly positive at ϑ_0 .

Assumption ii) represents the technical innovation for this version of the BvM theorem. A key assumption required for the parametric BvM theorem in Kleijn and van der Vaart (2012) to hold is a uniform LAN expansion, which posits that the LAN residual must converge in probability to 0 uniformly on a compact set. This uniform expansion follows easily from the pointwise LAN expansion in the iid case from the theory on Glivenko-Cantelli classes, because the score associated with a LAN expansion of the n -sample log-likelihood can be expressed as the sum of n iid random vectors that arise from the same score function evaluated at different sample observations. In the stationary Gaussian case the score associated with the T -sample log-likelihood is a vector of demeaned quadratic forms instead of a partial sum of iid random vectors, so that the usual empirical process theory does not automatically yield a uniform version of the LAN expansion from the pointwise version. Therefore, we replace the uniform LAN assumption with a condition that requires only uniform control of the LAN expansion residuals in L^1 , which turns out to follow much more readily from familiar primitive conditions in either the iid or stationary Gaussian case.

For models with iid data assumption ii) follows from the same assumptions as in lemma 2.1 of Kleijn and van der Vaart (2012), which furnish sufficient conditions for uniform LAN to hold. Meanwhile, in the stationary Gaussian case assumption ii) follows from the quadratic mean differentiability (QMD) of the spectral density (found in, for instance, van der Vaart (2000)) that holds uniformly over compact sets. However, uniform QMD of the spectral density is not enough for uniform LAN, which requires higher-order Lipschitz continuity assumptions in addition to QMD.

Given our regularity assumptions, the desired BvM theorem follows.

Theorem 1. *Suppose the regularity conditions in assumption 1 hold. Whenever $\eta_0 = \eta^* + \frac{h}{\sqrt{T}}$ for some $h \in \mathbb{R}^{k_\eta}$ and $P_{T,\theta_0} \triangleleft P_{T,\theta^*}$, we have*

$$d_{TV} \left(\Pi_{T,H} [\mathbf{X}_T, \cdot], \Phi_{T,H}^h [\mathbf{X}_T, \cdot] \right) = o_{P_{T,\theta_0}}(1).$$

where $d_{TV}(\cdot)$ is total variation distance.

Proof Sketch. We proceed in two steps. First, we establish a BvM theorem for the augmented parameter $\vartheta = (\nu, \delta)$ under the baseline sequence of distributions $\{P_{T,\theta^*}\}_{T \in \mathbb{N}_+}$. Let $\Pi_T[\mathbf{X}_T, \cdot]$ denote the posterior distribution of

$$\sqrt{T}(\vartheta - \vartheta_0),$$

and let $\Phi_T[\mathbf{X}_T, \cdot]$ be the normal distribution with mean

$$(\mathbb{I}_*^{\mathcal{A}})^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\vartheta}^{\mathcal{A}}(\vartheta_0 \mid \mathbf{X}_T)$$

and variance $(\mathbb{I}_*^{\mathcal{A}})^{-1}$. Following the truncation argument in Kleijn and van der Vaart (2012), we first

compare the posterior and Gaussian approximation on an arbitrary compact set K , and then use the posterior consistency assumption (assumption iii)) to extend this result to the entire euclidean space. The key departure from the standard argument is that we do not assume uniform convergence of the LAN residual. Instead, using the identity

$$(1 - x)^+ \leq |\log x|, \quad x > 0,$$

the total variation distance between the distributions truncated to K can be bounded in terms of the uniform L^1 norm of the LAN residual, which converges to zero by assumption ii). Hence,

$$d_{TV}(\Pi_T[\mathbf{X}_T, \cdot], \Phi_T[\mathbf{X}_T, \cdot]) = o_{P_{T, \theta^*}}(1).$$

Second, we transfer this result to the posterior distribution of

$$\sqrt{T} \{H^{\mathcal{A}}(\vartheta) - H(\theta_0)\}.$$

Assumption iv) gives the differentiability needed for a delta-method result in total variation, so the BvM theorem for ϑ implies the corresponding Gaussian approximation for $H^{\mathcal{A}}(\vartheta)$ under P_{T, θ^*} . Finally, $P_{T, \theta_0} \triangleleft P_{T, \theta^*}$ (which does not follow from the LAN of the augmented model, since mutual contiguity of the DGPs follows from LAN of the original model) combined with the translation from $H(\theta^*)$ to $H(\theta_0)$ yields the stated approximation under P_{T, θ_0} . Further details are given in the online appendix. $\square$

2.6 Prior Selection for δ

Consider an augmented model with directional matrix $\mathbf{B}_0$, and let $\hat{\delta}_T^{\mathcal{A}}$ be the maximum likelihood estimator of δ from this model. Let $\hat{\zeta}_T$ and $\hat{\eta}_T$ denote the maximum likelihood estimators of γ and η from the full model. From equation (2.4) we notice that, to first order,

$$\sqrt{T}\delta_T^{\mathcal{A}} = (\mathbf{B}_0' \mathbb{I}_{*, \zeta \zeta} \mathbf{B}_0)^{-1} \mathbf{B}_0' \mathbb{I}_{*, \zeta \zeta} \times \sqrt{T} \begin{bmatrix} \hat{\gamma}_T - \gamma_0 \\ \hat{\eta}_T - \eta_0 \end{bmatrix},$$

which suggests that δ can be heuristically understood as an approximation, locally at θ^* , to the following linear transformation of the deviations of ζ :

$$\underbrace{(\mathbf{B}_0' \mathbb{I}_{*, \zeta \zeta} \mathbf{B}_0)^{-1} \mathbf{B}_0' \mathbb{I}_{*, \zeta \zeta}}_{\mathbf{C}_0^*} \begin{bmatrix} \gamma - \gamma_0 \\ \eta - \eta^* \end{bmatrix}.$$

Based on this relationship, prior information on the parameters $\zeta = (\gamma, \eta)$ can be utilized to construct the prior on δ . For instance, if the prior on $\zeta - (\gamma_0, \eta^*)$ is $\mathcal{N} [O_{k_\zeta \times 1}, \Sigma^*]$, then the implied prior on δ would be $\mathcal{N} [O_{k_\delta \times 1}, \mathbf{C}_0^* \Sigma^* \mathbf{C}_0^{*'}]$.

In the asymptotic theory presented above, we have implicitly assumed that the augmented model is locally identified at θ^* , or in other words that the information matrix $\mathbb{I}_*^{\mathcal{A}}$ is non-singular. Because the augmentation is local around θ^* , the augmented model may not be globally identified across the entire augmented parameter space—but the non-singularity of $\mathbb{I}_*^{\mathcal{A}}$ suggests that restricting the prior of δ to a sufficiently small neighborhood around δ_0 suggests the posterior consistency assumption required by theorem 1.

In practice, the neighborhood should be chosen so that the linear approximation underlying the augmented model remains accurate throughout the support of the prior, which motivates the use of prior information on ζ to select this neighborhood. Suppose that Σ_ζ^* is the prior covariance matrix for the deviation of ζ from (ζ_0, η^*) , so that $\mathbf{C}_0^* \Sigma_\zeta^* \mathbf{C}_0^{*'}$ is the implied prior covariance matrix for δ . δ is then restricted to lie in the set of points in $\mathbb{R}^{k_\delta}$ whose Mahalanobis distance from the origin with respect to $\mathbf{C}_0^* \Sigma_\zeta^* \mathbf{C}_0^{*'}$ is bounded above by the square of some cutoff point $c > 0$. If the prior for δ is chosen to be a truncated normal prior with prior covariance $\mathbf{C}_0^* \Sigma_\zeta^* \mathbf{C}_0^{*'}$, then choosing c^2 to be the q th quantile of the chi-squared distribution with k_δ degrees of freedom ensures that we retain a fraction q of the original Gaussian prior mass.

2.7 Robustness to First-Stage Estimation Error

So far we have treated η^* as a fixed calibration and modeled the truth η_0 as a quantity that is local to η^* . Often it is more natural to treat the truth η_0 as a fixed but unknown value and η^* as an estimator of η_0 . Given a sample of size T , we write η_T^* to make the role of η^* as a first-stage estimator explicit. In contrast to the local misspecification framework, where score expansions were performed around the baseline calibration η^* , in the first-stage framework these same expansions are around the truth η_0 . This shift encapsulates the core difference between these two frameworks. Below we briefly run through the theory in a similar manner to the development above, and furnish sufficient conditions under which the augmented model delivers estimates that are robust to first-stage estimation error.

Suppose η_T^* is sample size-dependent but non-random³, and assume $\sqrt{T}(\eta_T^* - \eta_0) \rightarrow h$ as $T \rightarrow \infty$. Letting $H(\theta)$ be the object of interest, its plug-in maximum likelihood estimator is $H(\hat{\nu}_T^{nv}, \eta_T^*)$, where $\hat{\nu}_T^{nv}$ is the plug-in MLE for ν . Denote by $\dot{H}_0$ the derivative of $H(\cdot)$ at the truth $\theta_0 = (\nu_0, \eta_0)$; and likewise, we denote $\mathbb{I}_0$ as the Fisher information evaluated at the truth θ_0 instead of the baseline calibration. The same expansion as in section 2.3 reveals that the first-order asymptotic expansion of $H(\hat{\nu}_T^{nv}, \eta_T^*)$ involves only the ν -score evaluated at (ν_0, η_T^*) . A second order expansion

³We can deal with the case of random η_T^* by sample-splitting in the vein of Chernozhukov, Escanciano, et al. (2022), as mentioned in the introduction.

and the information matrix equality then reveals that $H(\hat{\nu}_T^{nv}, \eta_T^*)$ is asymptotically biased and the bias is equal to $(\dot{H}_{0,\eta} - \dot{H}_{0,\nu} \mathbb{I}_{0,\nu\nu}^{-1} \mathbb{I}_{0,\nu\eta}) h$.

On the other hand, if we use the η -score in addition to the ν -score when constructing an estimator of $H(\theta)$, so that $\sqrt{T} \{\hat{H}_T - H(\nu_0, \eta_T^*)\}$ is asymptotically equivalent to $\dot{H}_0 \mathbb{I}_0^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\theta}(\nu_0, \eta_T^* | \mathbf{X}_T)$, then the same first-order expansion of the score yields

$$\sqrt{T} \{\hat{H}_T - H(\theta_0)\} = \dot{H}_0 \mathbb{I}_0^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\theta}(\nu_0, \eta_0 | \mathbf{X}_T) + o_{P_{T,\theta_0}}(1),$$

which will be asymptotically normal with mean 0 in the limit. In other words, $\dot{H}_0 \mathbb{I}_0^{-1} \frac{1}{\sqrt{T}} \dot{\ell}_{T,\theta}(\nu_0, \eta_T^* | \mathbf{X}_T)$ remains the efficient influence function of $H(\theta)$ under the first-stage estimation framework.

Suppose we construct the full augmentation so that the T -sample log-likelihood of the augmented model satisfies

$$\begin{aligned} \ell_T^{\mathcal{A}}(\nu, \delta_0 | \mathbf{X}_T) &= \ell_T(\nu, \eta_T^* | \mathbf{X}_T) \quad \forall \nu \\ \dot{\ell}_{T,\theta}^{\mathcal{A}}(\nu_0, \delta_0 | \mathbf{X}_T) &= \dot{\ell}_{T,\theta}(\nu_0, \eta_T^* | \mathbf{X}_T) \end{aligned}$$

These conditions are identical to those in equations (2.3) and (2.4); thus, the same augmentation principle leads to local robustification in both asymptotic frameworks. The score matching condition again ensures that the MLE of the $\vartheta = (\nu, \delta)$ will be asymptotically linear in the original score $\dot{\ell}_{T,\theta}$. The continuity of the original Fisher information with respect to θ and the consistency of η_T^* for η_0 then ensures that the limiting Fisher information of the augmented model at (ν_0, δ_0) is exactly the Fisher information $\mathbb{I}_0$ of the original model evaluated at the truth. As such, the MLE of (ν, δ) obtained from the augmented model will be first order equivalent to $\mathbb{I}_0^{-1} \dot{\ell}_{T,\theta}(\nu_0, \eta_T^* | \mathbf{X}_T)$ and likelihood-based inference on $H(\theta)$ based on the augmented model will be efficient and asymptotically unbiased.

The BvM theorem holds for this version of the augmented model under almost identical assumptions as in theorem 1. The only major change is that the differentiability assumption for the function $H(\cdot)$ is formulated at the truth θ_0 instead of the baseline calibration.

2.8 Closed-Form Augmentation for Two Model Classes

So far, we have demonstrated that any augmented model satisfying conditions (2.3) and (2.4) generates posterior draws that enables locally robust and efficient inference of parameters of interest. In practice, it is difficult to find a well-defined statistical model that satisfies these conditions, since the score equation (2.4) requires us to solve a differential equation where instead of a boundary condition we must impose the condition that the augmented density integrates to 1. We subsequently introduce

two important classes of models, from micro and macroeconometrics respectively, for which there exist closed-form likelihood augmentations. Section 3 concerns closed-form augmentations for models belonging to the exponential family, followed by an application to dynamic discrete choice models with logit choice probabilities. Section 4 pertains to the augmentation of Gaussian SVMA models, and an application to the Krusell-Smith model.

3 Exponential Family and Dynamic Discrete Choice Models

One important class of models for which there exists a closed-form augmentation are models with iid data distributed according to a member of the exponential family of distributions. A multitude of structural models that belong to this class naturally lend themselves to augmentation, including dynamic discrete choice models (Rust (1987)); dynamic firm entry games (Aguirregabiria, Mira, and Roman (2007)); and spatial discrete choice models (Brock and Durlauf (2001)).

In all these models the choice probabilities often follow a multinomial logit distribution, which belongs to the exponential family. Furthermore, joint estimation of all the structural parameters involve the costly computation of equilibrium objects such as conditional choice probabilities in dynamic discrete choice models and continuation values in dynamic games. Plug-in estimators are already used in these models to avoid recomputing the computationally expensive part of the model for different values of the structural parameters—the model augmentation procedure proposed above robustifies inference with respect to local misspecification in the plug-in values.

3.1 Augmentation for Exponential Families

Formally, we consider models where $\{X_i\}_{i \in \mathbb{N}}$ is iid and the common density across X_i is given as

$$p(x; \theta) = h(x) \exp(\alpha(\theta)' S(x) - A(\alpha(\theta))) \quad (3.1)$$

for any x in the support E of X_i , positive valued measurable function $h(\cdot)$, measurable m -dimensional functions $\alpha(\cdot)$, $S(\cdot)$ and measurable scalar-valued function $A(\cdot)$. In particular, the log-partition function $A(\cdot)$ is chosen to ensure that the density integrates to 1 for all values of the natural parameter. The n -sample score function is given as

$$\dot{\ell}_{n,\theta}(\theta \mid \mathbf{X}_n) = \sum_{i=1}^n (\partial_\theta \alpha(\theta))' \{S(X_i) - \partial_\alpha A(\alpha(\theta))\},$$

where ∂_θ is the differentiation operator with respect to θ , and ∂_α is defined similarly.

Let θ be partitioned into (β, γ, η) as above, and let η^* be a plug-in value for η . Consider an

augmented model with k_δ -dimensional auxiliary parameter δ and directional matrix $\mathbf{B}_0 \in \mathbb{R}^{k_\zeta \times k_\delta}$. Then, conditions (2.3) and (2.4) will be satisfied if

$$p^{\mathcal{A}}(x; \nu, \delta_0) = p(x; \nu, \eta^*) \quad \forall \nu \quad (3.2)$$

$$\frac{\partial \log p^{\mathcal{A}}(x; \nu_0, \delta_0)}{\partial \vartheta'} = \begin{pmatrix} \partial_\nu \alpha(\nu_0, \eta^*)' \\ \mathbf{B}_0' (\partial_\zeta \alpha(\nu_0, \eta^*))' \end{pmatrix} \{S(x) - \partial_\alpha A(\alpha(\nu_0, \eta^*))\}, \quad (3.3)$$

where ϑ is shorthand for (ν, δ) .

It is now very simple to verify that an augmented model that also belongs to the exponential family with the following per-observation density

$$p^{\mathcal{A}}(x; \nu, \delta) = h(x) \exp\left(\alpha^{\mathcal{A}}(\nu, \delta)' S(x) - A\left(\alpha^{\mathcal{A}}(\nu, \delta)\right)\right)$$

with augmented natural parameter

$$\alpha^{\mathcal{A}}(\nu, \delta) = \alpha(\nu, \eta^*) + D_\zeta \alpha(\nu, \eta^*) \mathbf{B}_0 \delta \quad (3.4)$$

satisfies conditions (3.2) and (3.3). The key feature of this construction is that we linearize the natural parameter $\alpha(\theta)$, rather than the log-likelihood itself. By retaining the original log-partition function $A(\cdot)$, the augmented density remains a proper density for every admissible value of (ν, δ) . Moreover, because the augmented natural parameter coincides with the original natural parameter at $\delta_0 = O_{k \times 1}$, and its derivative with respect to δ reproduces the relevant nuisance-score directions, the augmented likelihood satisfies the score-matching conditions (3.2) and (3.3). As such, the construction delivers a well-defined likelihood that preserves the local score structure of the original model. Estimating (ν, δ) in the augmented model is also computationally cheap due to the fact that δ enters into the natural parameter in a linear manner instead of through the non-linear mapping $\eta \mapsto \alpha(\nu, \eta)$.

3.2 Dynamic Discrete Choice Models with Logit Choice Probabilities

Here we apply the practical augmentation strategy developed above to general DDC models of the kind studied in Rust (1987), Hotz and Miller (1993) and Aguirregabiria and Mira (2002), among others. Following Hotz and Miller (1993), a popular approach to estimating DDC models is to obtain a first stage estimator of the conditional choice probabilities (CCP) alongside the transition probability parameters and then obtain the model-implied probabilities given these plug-ins to formulate moment conditions for the structural parameters. The resulting structural parameter estimates are exposed to first-stage estimation error—there is a well-developed literature that

addresses this problem for frequentist inference, including but not limited to Kasahara and Shimotsu (2008), Pesendorfer and Schmidt-Dengler (2008), and more recently, Chernozhukov, Escanciano, et al. (2022).

Below we present a simple version of the DDC model and illustrate how it fits into the exponential family framework established above. We note how the Hotz and Miller (1993) type models studied in this section are Neyman orthogonal to the CCPs by design, so that estimation of the parameters of interest are not contaminated by first-order estimation error in the plug-in CCPs. The model remains vulnerable to first-order estimation error in the transition probability parameters, however, and we construct the augmented version of the model that to robustify the model against local transition probability misspecification.

3.2.1 Model Setup

We consider a model where a single agent faces a discrete time infinite-horizon problem where the choice set $\mathcal{D}$ consists of J distinct options, labeled 1 to J . The state vector s can be decomposed as $s = (x, \epsilon)$. x is the endogenous state that is Markovian, and so depends only on the value of x and the agent's choice in the immediate past. ϵ is the exogenous choice variable that is iid given the endogenous state, and whose distribution at each period depends only on the contemporaneous endogenous state. Let $f(x' | x, d)$ be the probability of observing x' tomorrow given x, d today. $g(x, \cdot)$ denotes the distribution of the exogenous state given today's endogenous state x . $u(s, d)$ is the flow utility of the agent given a pair (s, d) of the state and control, and here we assume that $u(s, d)$ is additively separable in the exogenous state, so that $u(s, d) = \tilde{u}(x, d) + \epsilon(d)$, where $\epsilon(d)$ is the d th coordinate of the random vector ϵ .

Letting $d^*(x, \epsilon)$ be the optimal decision function given the state variables (x, ϵ) , $P_d(x) = \mathbb{P}(d^*(x, \epsilon) = d | x)$ is the conditional choice probability (CCP) of choosing d given the endogenous state x . Collecting these CCPs into the finite-dimensional vector P , Aguirregabiria and Mira (2002) demonstrates that there exists a mapping $\mathbb{P}(d_{it} = d | x_{it} = x) = \Psi(x, d, P)$ between the model-implied probability $\mathbb{P}(d_{it} = d | x_{it} = x)$ and a given set of candidate CCPs P . The Hotz and Miller (1993) CCP estimator can then be understood as a two-step process where we first procure a first-stage estimator $\hat{P}$ of the CCPs and then estimate the structural parameters by minimizing the distance $|\mathbb{P}(d_{it} = d | x_{it}) - \Psi(x_{it}, d; \hat{P})|$ between the empirical and model-implied choice probabilities.

3.2.2 A Locally Enlarged Model

Let θ_u denote the parameters that govern the flow utility $u(\cdot)$, and θ_f the parameters determining the evolution of the endogenous state. If the CCPs P are treated as a parameter representing the candidate optimal policy, the mapping $\Psi(x, d; \theta_u, \theta_f, P)$ specifies a likelihood for the observations

$\{d_{it}, x_{it}\}_{1 \leq t \leq T_i, i \in \mathbb{N}_+}$ of agent choices and endogenous states up to some finite time, as a parametric model with parameters θ_u, θ_f and P . A model in which the CCP vector P is treated as an input to the mapping $\Psi(P)$ does not, in general, impose the equilibrium fixed-point restriction $P = \Psi(P)$. Instead, it can be viewed as a locally enlarged statistical model indexed by P . The original structural DDC model is nested within this enlarged model at the equilibrium CCP vector P_0 for which $P_0 = \Psi(\theta_u, \theta_f, P_0)$, while off-equilibrium values of P need not satisfy the fixed-point restriction themselves.

It is convenient to work with the enlarged model instead of a model that imposes the fixed-point restriction for each combination of (θ_u, θ_f) for two reasons. First, the model-implied probabilities depend on the flow utility parameters θ_u in a simple manner in the enlarged model, as seen in equation (3.5) for the specific case of type-I generalized extreme value distributed exogenous states. This greatly facilitates two-step estimation of θ_u given a first-stage estimate of θ_f and P based on maximizing the model-implied probabilities, a procedure that results in exactly the Hotz and Miller (1993) estimator for θ_u , and in the same vein leads to an augmented model that is computationally tractable.

Moreover, the enlarged model admits a zero-Jacobian property that renders likelihood-based two-step estimators of θ_u locally insensitive to first-stage estimation error in the CCPs P , so that we need only robustify the model against estimation error in the state transition parameters θ_f . As such, despite the enlarged model admitting the CCPs P as an additional set of parameters and therefore being much more high-dimensional compared to the original structural model that imposes the equilibrium fixed-point relation, when it comes to conducting locally robust inference on θ_u the enlarged model can essentially be treated as a parametric model of (θ_u, θ_f) with P fixed at its first-stage estimate. Therefore, the enlarged model is easier to estimate and augment compared to the original structural model—the computational simplicity is at the cost of efficiency, as noted in Aguirregabiria and Mira (2002).

3.2.3 Likelihood of the Enlarged Model

We adopt two conventions that are commonly found in DDC specifications: that $u(x, d; \theta_u)$ is multiplicatively separable in the parameter θ_u , so that $u(x, d; \theta_u) = h(x, d)'g(\theta_u)$ for some function $g(\cdot)$; and that ϵ is a J -dimensional random vector whose coordinates follow independent standard type-I generalized extreme value distributions. Under these assumptions, the model-implied probabilities are given as

$$\mathbb{P}(d_{it} = d \mid x_{it}; \theta_u, \theta_f, P) = \frac{\exp(M_0(x_{it}, d; P, \theta_f) + M_1(x_{it}, d; P, \theta_f)'g(\theta_u))}{1 + \sum_{d'=2}^J \exp(M_0(x_{it}, d'; P, \theta_f) + M_1(x_{it}, d'; P, \theta_f)'g(\theta_u))}, \quad (3.5)$$

for some functions $M_0(\cdot)$ and $M_1(\cdot)$, which are defined in the online appendix.

From equation (3.5) we can see that the model-implied probabilities depend on θ_u linearly through the function $g(\cdot)$, while the nuisance parameters P, θ_f enter into the probabilities non-linearly through the intercept and slope functions $M_0(\cdot)$ and $M_1(\cdot)$. The computational cost of jointly estimating (θ_u, P, θ_f) can be traced back to how recomputing $M_0(x, d; P, \theta_f)$ and $M_1(x, d; P, \theta_f)$ for different values of (P, θ_f) requires the repeated solution of a complex dynamic programming problem.

Let $\{d_{it}, x_{it}\}_{1 \leq t \leq T, i \in \mathbb{N}_+}$ be an iid sample of time series observations (d_{it}, x_{it}) of agents' choices and endogenous states. Then, the log likelihood can be decomposed into the sum of a choice probability component and a state-transition component:

$$\ell(\theta, P; \{(d_{it}, x_{it})\}_{1 \leq t \leq T}) = \sum_{t=1}^T \log \mathbb{P}(d_{it} \mid x_{it}; \theta, P) + \sum_{t=1}^T \log f(x_{it} \mid x_{i,t-1}, d_{i,t-1}; \theta_f)$$

Under the extreme value distribution assumption for the exogenous states, the choice probabilities, as we saw above, belong to the exponential family. Suppose in addition that the state transition probability $f(x' \mid x, d)$ also belongs to the exponential family. Then, the per-unit log-likelihood can be written as

$$\begin{aligned} \ell(\theta, P; \{(d_{it}, x_{it})\}_{1 \leq t \leq T}) &= \sum_{t=1}^T \{ \alpha_d(x_{it}; \theta, P)' S_d(d_{it}) - A_d(\alpha(x_{it}; \theta, P)) \} \\ &\quad + \sum_{t=1}^T \{ f(x_{it}) + \alpha_f(x_{i,t-1}, d_{i,t-1}; \theta_f)' S_f(x_{it}) - A_f(\alpha_f(x_{i,t-1}, d_{i,t-1}; \theta_f)) \} \end{aligned}$$

where

$$\begin{aligned} \alpha_d(x; \theta, P) &= \begin{pmatrix} 0 \\ M_0(x, 2; P, \theta_f) \\ \vdots \\ M_0(x, J; P, \theta_f) \end{pmatrix} + \begin{pmatrix} O_{1 \times m} \\ M_1(x, 2; P, \theta_f)' \\ \vdots \\ M_1(x, J; P, \theta_f)' \end{pmatrix} g(\theta_u), \quad S_d(d) = \begin{pmatrix} I_{\{d=1\}} \\ \vdots \\ I_{\{d=J\}} \end{pmatrix}, \\ A_d(\alpha) &= \log \left(1 + \sum_{i=2}^J \exp(\alpha_i) \right) \quad \forall \alpha = (\alpha_1, \dots, \alpha_J) \in \mathbb{R}^J, \end{aligned}$$

and $f(\cdot)$, $\alpha_f(\cdot)$ and $A_f(\cdot)$ are the base measure, natural parameter and log-partition function of the state transition distribution.

3.2.4 Natural Robustness to First-Stage CCP Estimation Error

A well-known property of the model-implied probability mapping $\Psi(\theta, P)$ in Aguirregabiria and Mira (2002) is that its derivative with respect to P vanishes at the equilibrium CCP P_0 . Because the fixed-point condition $P = \Psi(\theta, P)$ holds at equilibrium, this zero-Jacobian property also implies $D_P \alpha_d(x; \theta_0, P_0) = 0$, so that the Fisher information blocks involving P vanish at the truth. Hence the score for θ_u is already Neyman orthogonal with respect to P , and first-stage CCP estimation error satisfying $\sqrt{n}(P_n^* - P_0) = O(1)$ has no first-order effect on estimation of θ_u .

This robustness does not generally extend to the transition parameters θ_f . Although θ_f is often estimated in a first stage using only the observed state transitions $\{x_{it}\}$, the conditional choice probabilities $\mathbb{P}(d_{it} = d \mid x_{it} = x; \theta, P)$ also depend on θ_f and therefore contain additional information about these parameters. Consequently, plugging in first-stage estimates of θ_f may affect inference on θ_u at first order.

Collecting the nuisance parameters as $\eta = (P, \theta_f)$, it therefore suffices to restore only the relevant components of the θ_f -score in order to obtain inference on θ_u that is robust to first-order estimation error in η .

3.2.5 Augmentation of the Likelihood

The log-likelihood of the full model corresponds directly to a conditional multinomial logit model, which is part of the exponential family. Let the baseline model be the full model with η fixed at some (non-random) first-stage estimate η^* , so that it is a parametric model with the sole parameter block being θ_u . The augmented model includes a k_{θ_u} -dimensional auxiliary parameter δ and its log likelihood is specified as

$$\begin{aligned} \ell^{\mathcal{A}}(\theta_u, \delta; \{d_{it}, x_{it}\}_{1 \leq t \leq T}) &= \sum_{t=1}^T \left\{ \alpha_d^{\mathcal{A}}(x_{it}; \theta_u, \delta)' S_d(d_{it}) - A \left(\alpha_d^{\mathcal{A}}(x_{it}; \theta_u, \delta) \right) \right\} \\ &\quad + \sum_{t=1}^T \left\{ f(x_{it}) + \alpha_f^{\mathcal{A}}(x_{i,t-1}, d_{i,t-1}; \delta)' S_f(x_{it}) - A_f(\alpha_f^{\mathcal{A}}(x_{i,t-1}, d_{i,t-1}; \delta)) \right\} \end{aligned}$$

where, as discussed above, we only restore the relevant directions of the θ_f -score:

$$\begin{aligned} \alpha_d^{\mathcal{A}}(x; \theta_u, \delta) &= \alpha_d(x; \theta_u, \eta^*) + D_{\theta_f} \alpha_d(x; \theta_u, \eta^*) \mathbf{B}_0 \delta \\ &= \alpha_d(x; \theta_u, \eta^*) + \begin{pmatrix} O_{1 \times k_{\eta}} \\ D_{\theta_f} M_0(x, 2; \eta^*) + g(\beta)' \left(D_{\theta_f} M_1(x, 2; \eta^*) \right) \\ \vdots \\ D_{\theta_f} M_0(x, J; \eta^*) + g(\beta)' \left(D_{\theta_f} M_1(x, J; \eta^*) \right) \end{pmatrix} \mathbf{B}_0 \delta \end{aligned}$$

$$\alpha_f^{\mathcal{A}}(x, d; \delta) = \alpha_f(x, d; \eta^*) + D_{\theta_f} \alpha_f(x, d; \eta^*) \mathbf{B}_0 \delta$$

and $\mathbf{B}_0$ is the population regression coefficient from regressing the θ_u -score onto the θ_f -score. The requisite derivatives can be obtained via two-sided numerical differentiation in general; in the present case, we can procure closed-form expressions for the derivatives.

Below, we apply this augmentation principle to a model of firm entry and exit, which is intended to provide insight into how model augmentation restores valid inference and illustrate the computational tractability of the augmented model in a setting with continuous state variables.

3.3 A Model of Firm Entry and Exit

The model in this section is based on a similar model in Imai et al. (2009). The fact that the state variable is continuous means that, in general, the state space must be discretized before solving the dynamic programming problem—for instance when using Gauss-Hermite quadrature to numerically compute value function expectations. The results below demonstrate once more how model augmentation robustifies inference about flow utility parameters against local bias and error in first stage estimates of the state transition parameters. In addition, it is shown how, even with a single continuous state variable, full information joint estimation of (θ_u, θ_f) in the enlarged model with CCPs given as inputs rapidly becomes infeasible as the state space grid becomes finer. This suggests that, for DDC models with multiple continuous state variables, the original model faces a real computational problem that the augmented model helps resolve.

The economy is populated by firms that choose in each period whether to enter the market, if they are potential entrants, or exit the market, if they are incumbents. We denote by $\mathcal{D} = \{0, 1\}$ the choice set of the firms, where $d = 0$ represents market inactivity and $d = 1$ activity. Their choices are guided by an unobserved state ϵ that represent idiosyncratic shocks to their profits and two observed states: capital stock k and their status s as an incumbent or a potential entrant. As above we assume that the unobserved states ϵ are iid and its components follow independent type-I generalized extreme value distributions. Collecting the observed states into the vector $x = (k, s) \in \mathcal{X}$, the firm's flow utility is given as $u(x, d) + \epsilon(d)$, where

$$u(x, d) = (\alpha k s - c(1 - s))d$$

for flow utility parameters α, c . α is the marginal utility from remaining in the market and accumulating capital, while c is the fixed cost of competition incurred by an entrant at the time of entry into the market. For each firm i , we know that $s_{i,t+1} = d_{it}$ mechanically, that is, the choice of whether to enter or exit today will determine their incumbency status tomorrow. The log capital of an incumbent firm i that remains in the market is assumed to evolve according to the AR(1)

process $\log k_{i,t+1} = b_0 + b_k \log k_{it} + \sigma_u u_{i,t+1}$, where $u_{i,t+1}$ is a standard normal shock that is iid across both firms and time; the time $t + 1$ log capital of a firm that decides to enter at time t is given as $\log k_{i,t+1} = b_e + \sigma_u u_{i,t+1}$. The capital stock of a firm that is inactive at the market at time t is normalized to 0. The parameters of the model are collected as $\theta = \{\theta_u, P, \theta_f\}$ where the flow utility parameters are $\theta_u = \{\alpha, c\}$, the state transition parameters are $\theta_f = \{b_0, b_1, b_e, \sigma_u^2\}$ and the input CCPs are denoted by P .

Letting the discretized capital state space consist of M points $\{k_1, \dots, k_M\}$, the entire observed state space can be assumed to consist of $M + 1$ points $\{inactive\} \cup \{k_1, \dots, k_M\}$ because we know that firms hold non-zero capital stock only when they are active. Solving the dynamic programming problem for each value of θ then involves factorizing a $(M + 1) \times (M + 1)$ matrix each time.

3.3.1 Model Likelihoods

As in the general case, the likelihood of the model can be decomposed into two parts: a term representing the choice probabilities conditional on the observed states and another representing the state transition probabilities. Given a sample $\{d_{it}, k_{it}\}_{0 \leq t \leq T, i \in \mathbb{N}_+}$ of entry choices and capital stock for independent firms, the log-likelihood for firm i can be expressed as

$$\begin{aligned} \ell(\theta, P; \{d_{it}, k_{it}\}_{0 \leq t \leq T}) = & \underbrace{\sum_{t=1}^T \{\alpha_d(x_{it}; \theta, P) d_{it} - \log(1 + \exp(\alpha_d(x_{it}; \theta, P)))\}}_{\ell_{choice}(\theta, P)} \\ & + \underbrace{\sum_{t=1}^T d_{i,t-1} \{f(k_{it}) + \alpha_f(x_{i,t-1}; \theta_f)' S_f(k_{it}) - A_f(\alpha_f(x_{i,t-1}; \theta_f))\}}_{\ell_{state}(\theta_f)} \quad (3.6) \end{aligned}$$

where $f(\cdot)$, $S_f(k)$ and $A_f(\cdot)$ are the base measure, sufficient statistic and log-partition function for the lognormal distribution. The expressions for the natural parameters $\alpha_d(x; \theta, P)$ and $\alpha_f(x; \theta_f)$ of the choice probability and state transition are provided in the appendix.

Let $\eta^* = \{P^*, \theta_f^*\}$ be the first-stage estimates of the input CCPs and the state transition parameters. We assume that η^* is formed using an auxiliary panel of choices and observed states that is independent of the panel used to form the model likelihood⁴. The independence of the two panels separates the uncertainty represented by the first-stage estimates from the sampling information contained in the enlarged likelihood. For simplicity both panels contain the same number of individual units and time periods.

⁴This is equivalent to sample-splitting in the vein of Chernozhukov, Chetverikov, et al. (2018) in the case that we start from a single unified panel of choices and states using which we must form both the first-stage estimates and the model likelihood.

Given η^* the flow utility parameters θ_u are estimated in the baseline model by ignoring the potential estimation error in P^*, θ_f^* . Because the state transition part of the log-likelihood does not depend on θ_u , the baseline model makes use only of $\ell_{choice}(\theta_f)$. The augmented model is again formed by linearizing the model around the first stage estimates with respect to θ_f . Letting δ be the auxiliary parameters introduced in the augmented model, the augmented log-likelihood for firm i simply replaces the natural parameters $\alpha_d(\cdot)$ and $\alpha_f(\cdot)$ in equation (3.6) by

$$\begin{aligned}\alpha_d^{\mathcal{A}}(x; \theta_u, \delta) &= \alpha_d(x; \theta_u, \eta^*) + \frac{\partial \alpha_d(x; \theta_u, \eta^*)}{\partial \theta_f'} \mathbf{B}_0 \delta \\ \alpha_f^{\mathcal{A}}(x; \delta) &= \alpha_f(x; \theta_f^*) + \frac{\partial \alpha_f(x; \theta_f^*)}{\partial \theta_f'} \mathbf{B}_0 \delta,\end{aligned}$$

where $\mathbf{B}_0$ is the 4×2 directional matrix constructed as in section 2.4.2 to ensure parsimonious model augmentation.

3.3.2 Prior Selection and Posterior Sampler

The first stage estimate θ_f^* is usually formed by maximizing the state transition likelihood $\ell_{state}(\theta_f)$ over the parameter space for θ_f . Therefore, a natural candidate for the prior covariance matrix for θ_f and by extension δ is $\frac{1}{n} \hat{V}_n$, where $\hat{V}_n$ is a consistent estimator of the asymptotic variance of the MLE of θ_f computed from the state transition likelihood $\ell_{state}(\theta_f)$. However, if we use $\frac{1}{n} \hat{V}_n$ directly as the prior covariance, we would be imposing a shrinkage prior on δ that shrinks at the parametric rate and therefore lead to a failure of the BvM theorem.

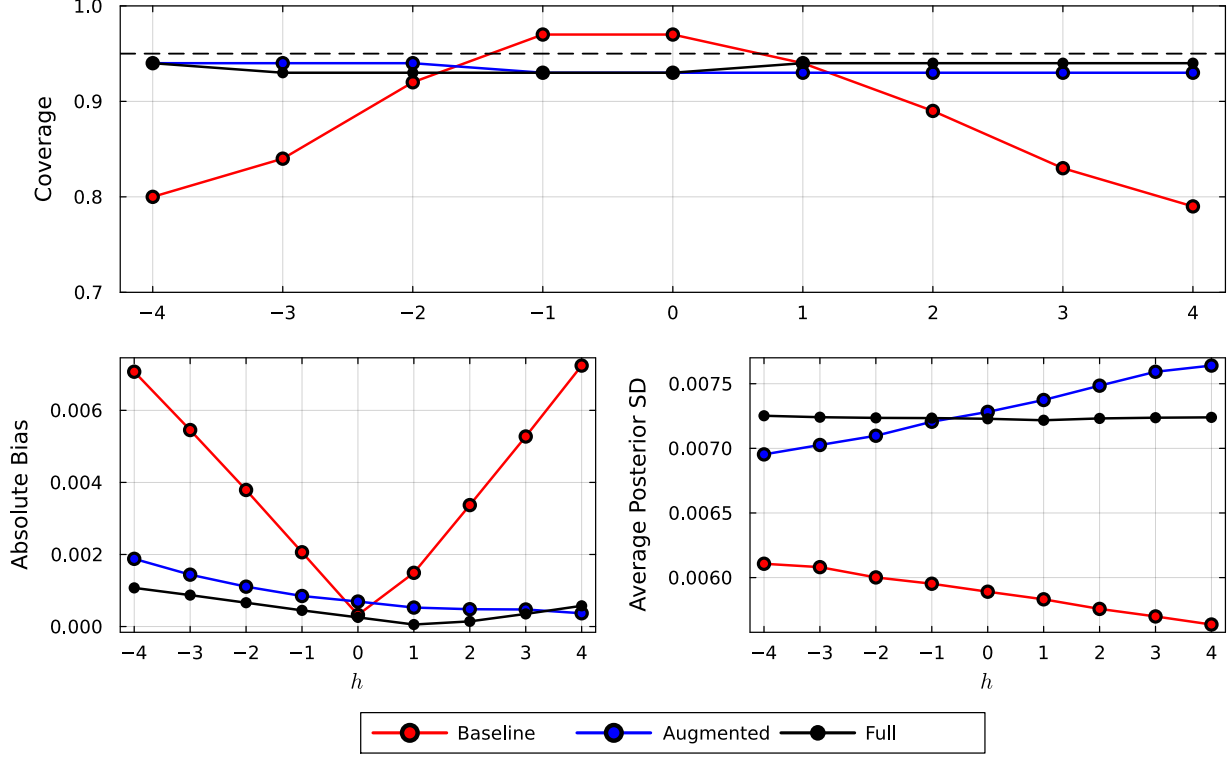
A simple remedy is to use an inflated version of $\frac{1}{n} \hat{V}_n$ as the prior variance, where the inflation is chosen so that the prior covariance on δ continues to shrink to 0 but at a rate slower than $n^{-1/2}$. This allows the prior to continue to reflect the sampling uncertainty in θ_f^* while remaining negligible in the limit. The general variance inflation strategy is to choose a sequence of scales $\{a_n\}_n$ such that $a_n \rightarrow +\infty$ and $a_n n \rightarrow 0$ as $n \rightarrow \infty$ and set the prior distribution on δ as

$$\delta \sim \mathcal{N} \left[0, \frac{a_n}{n} \hat{V}_n \right]$$

subject to the Mahalanobis truncation rule described in section 2.6. In this case, $\frac{a_n}{n} \hat{V}_n = \frac{a_n}{n} \times O(1)$ and the prior variance shrinks to 0 as $n \rightarrow \infty$, but at the same time $a_n \times \hat{V}_n \rightarrow +\infty$, so that information accumulates much faster in the likelihood compared to the prior and the effect of the prior ultimately disappears in the limit. In the simulation exercise below we choose $a_n = \log(n)$.

For the posterior sampler, the non-linear nature of the likelihood function motivates the use Hamiltonian Monte Carlo (HMC) implemented using the No U-Turn Sampler (NUTS). The HMC sampler requires as inputs only the model likelihood and priors, and is readily implementable in

Figure 1: **Monte Carlo comparison of the baseline and augmented models under first-stage estimation error.** We report the empirical coverage of nominal 95% credible intervals, the mean absolute bias of the posterior mean of α , and the average posterior standard deviation across 500 Monte Carlo iterations and various choices of h in the local perturbation $\theta_f^* = \theta_f + h \frac{\mathbf{d}}{\sqrt{n}}$.



statistical software such as R and Julia.

3.3.3 Simulation Results

We generate data based on the values $\alpha = 0.18$, $c = 0.8$, $b_0 = 0$, $b_1 = 0.78$, $b_e = 0.5$ and $\sigma_u^2 = 0.42$, alongside equilibrium CCPs implied by these parameters. A panel of choice and state variables are generated with $T = 10$ and $n = 1,000$. We conduct inference on $\theta_u = (\alpha, c)$ under the first stage estimate $\theta_f^* = \theta_f + h \frac{\mathbf{d}}{\sqrt{n}}$ across the baseline and augmented models, as well as the model where (θ_u, θ_f) are estimated jointly under a fixed set of input CCPs. The vector $\mathbf{d}$ is the direction of the first stage estimation error, and for illustrative purposes it is chosen as the vector with unit norm that maximizes the theoretical asymptotic bias of θ_u under the baseline model. h is a scalar that governs the magnitude of the misspecification.

We use an independently simulated panel to estimate the asymptotic covariance matrix of the first-stage estimator, which is then used to construct the prior covariance of δ according to the

Table 1: **Computational Cost of Alternative Estimation Methods.** Median computation times for likelihood and gradient evaluations across alternative estimation methods and grid sizes.

Grid size M	Baseline	Parsimonious Aug.	Joint	Joint / Pars. Aug.
<i>Panel A: Likelihood evaluation time (ms)</i>				
25	0.120	0.218	0.312	1.43
50	0.120	0.219	0.482	2.20
100	0.120	0.219	1.057	4.84
300	0.120	0.219	6.769	30.88
500	0.123	0.226	23.369	103.33
<i>Panel B: Gradient evaluation time (ms)</i>				
25	0.173	0.320	0.534	1.67
50	0.173	0.319	1.220	3.83
100	0.174	0.323	4.826	14.94
300	0.173	0.326	84.877	260.29
500	0.173	0.320	399.584	1247.72

Notes: The table reports median computation times, in milliseconds, for a single likelihood evaluation (Panel A) and a single gradient evaluation (Panel B), across alternative state-grid sizes M . The simulated dataset consists of $N = 1,000$ firms observed over $T = 10$ periods. The baseline model fixes the nuisance parameters, the parsimoniously augmented model estimates a low-dimensional augmentation, and the joint model estimates both structural and nuisance parameters. Gradients are computed using forward-mode automatic differentiation (`ForwardDiff.jl`). Execution times are measured using `BenchmarkTools.jl` and exclude model-specific precomputation. The final column reports the ratio of joint-model to parsimoniously augmented-model computation times. All benchmarks were performed using Julia 1.11.6 on an Apple M2 processor with 8 GB of RAM.

variance inflation rule described above. To isolate the effect of first-stage estimation error, rather than recomputing the first-stage estimator in each Monte Carlo replication, we introduce controlled local perturbations of the transition parameters along the prespecified direction $\mathbf{d}$.

In figure 1 we compare the true coverage of nominal 95% credible intervals for α , as well as the associated average absolute bias $|\hat{\alpha} - 0.18|$ and posterior standard deviation for the three models under values of h ranging from -4 to 4. For this Monte Carlo exercise we use a common capital grid with $M = 25$ points. Actual Monte Carlo coverage rates of the posterior credible intervals remain close to the nominal rate of 95% across different values of h when α is estimated from the augmented model or jointly estimated with the state transition parameters. However, the baseline model suffers from undercoverage for large $|h|$, which can be attributed to high asymptotic bias and low posterior standard deviations for such values of h . The augmented model corrects for such undercoverage by shifting the posterior distribution in two dimensions: it reduces asymptotic bias and does not underestimate the true uncertainty associated with α , represented by higher posterior standard deviations across the board compared to the baseline model.

In table 1 we tabulate the time needed, across the three models, for a single evaluation and

automatic differentiation of the log-likelihood, under different resolutions of the state space grid governed by the number of grid points M . The computing times for the baseline and augmented models are largely invariant to M , which is unsurprising since coefficients that are pre-computed once for each given M are fed into the log-likelihoods of these models as fixed inputs. While the joint estimation model is almost as fast as the other two models for smaller values of M , the computational cost quickly balloons once as M grows. At $M = 100$, differentiating the full-model likelihood is already approximately fifteen times as costly as differentiating the augmented likelihood. This gap grows rapidly with grid resolution, reaching more than 250 times at $M = 300$ and more than 1,000 times at $M = 500$. HMC routines require multiple gradient evaluations at each sampling iteration, so it is prohibitively costly to jointly estimate (θ_u, θ_f) even for a state space grid that is only moderately large. In contrast, the augmented model remains computationally tractable regardless of the resolution of the state space grid.

4 Gaussian SVMA and Heterogeneous-Agent DSGE Models

In the context of macroeconometrics, for the class of Gaussian SVMA models it is possible to analytically find an augmented model—also belonging to the class of Gaussian SVMA models—that satisfies conditions (2.3) and (2.4). Gaussianity and the SVMA representation are both indispensable for the tractability of this class. The Gaussianity of the primitive innovations means that the score function is linear in the spectral density, so that the score function can be formulated in terms of the spectral density. Meanwhile, the spectral density of SVMA models are quadratic in the corresponding MA lag polynomials, which suggests that a linear augmentation of the individual MA coefficients should be enough to achieve the desired augmentation of the spectral density.

Many models of interest in macroeconomics can be expressed as Gaussian SVMA models. In almost all instances in the literature where VAR or VMA models are estimated via Bayesian methods, the innovations are assumed to be Gaussian (e.g. Giannone et al. (2015), Plagborg-Møller (2019)). Exogenous variables in DSGE models are often assumed to follow simple ARMA processes with Gaussian errors, and it has been known as far back as Hansen and Sargent (1981) that linearized DSGE models admits an SVMA representation in addition to a state-space formulation. Concerns about the validity of VAR models under non-invertibility (Chahrour and Jurado (2022)) also makes the focus on VMA, instead of VAR, models more natural.

Below we derive the closed-form augmentation for Gaussian SVMA models that robustifies inference against local misspecification in a subset of the model parameters. Afterward, we apply the result to estimate the Krusell and Smith (1998) model under possible misspecification in the capital share parameter α .

4.1 Augmentation of Gaussian SVMA Models

Suppose that $\{X_t\}_{t \in \mathbb{Z}}$ is generated as

$$X_t = \mu(\theta) + \sum_{j=0}^{\infty} \Psi_j(\theta) \epsilon_{t-j}, \quad \epsilon_t \stackrel{iid}{\sim} \mathcal{N} [O_{n_\epsilon \times 1}, I_{n_\epsilon}]. \quad (4.1)$$

with $n \leq n_\epsilon$. Under absolute summability assumptions, the spectral density is given as

$$f(\omega; \theta) = \frac{1}{2\pi} \Psi(e^{i\omega} | \theta) \Psi(e^{i\omega} | \theta)^*$$

where $\Psi(z | \theta) = \sum_{j=0}^{\infty} \Psi_j(\theta) z^j$ for any $z \in \mathbb{C}$ and the superscript $*$ indicates conjugate transposition. Each $\mathbf{X}_T$ follows a normal distribution with mean depending on $\mu(\theta)$ and covariance matrix that is a Toeplitz matrix generated by the spectral density $f(\cdot; \theta)$.

Let η^* be a plug-in value for η . The Gaussianity of the DGP implies that the T -sample score of the baseline model, where η is considered fixed at η^* , is linear in the derivative of the spectral density with respect to the parameters and the long run mean function $\mu(\cdot)$. These are also the only derivatives that appear in the T -sample score, so that we need only impose appropriate conditions on the derivatives $\frac{\partial f(x; \theta^*)}{\partial \theta_i}$ and $\frac{\partial \mu(\theta)}{\partial \theta_i}$ for $1 \leq i \leq k$ in order for condition (2.4) to be satisfied.

As in the exponential family case, conjecture that the augmented model is also given as a Gaussian SVMA model:

$$X_t = \mu^{\mathcal{A}}(\nu, \delta) + \sum_{j=0}^{\infty} \Psi_j^{\mathcal{A}}(\nu, \delta) \epsilon_{t-j}, \quad \epsilon_t \stackrel{iid}{\sim} \mathcal{N} [O_{n_\epsilon \times 1}, I_{n_\epsilon}].$$

Letting $f^{\mathcal{A}}(\cdot; \nu, \delta)$ be the spectral density of the augmented model, note how condition (2.3) requires that

$$\mu^{\mathcal{A}}(\nu, \delta_0) = \mu(\nu, \eta^*) \quad \text{and} \quad f^{\mathcal{A}}(\cdot | \nu, \delta_0) = f(\cdot | \nu, \eta^*)$$

for any ν . Meanwhile, (2.4) can be formulated in terms of the mean function and spectral density as

$$\frac{\partial \mu^{\mathcal{A}}(\nu_0, \delta_0)}{\partial \nu'} = \frac{\partial \mu(\theta^*)}{\partial \nu'}, \quad \frac{\partial \mu^{\mathcal{A}}(\nu_0, \delta_0)}{\partial \delta'} = \frac{\partial \mu(\theta^*)}{\partial \zeta'} \mathbf{B}_0 \quad (4.2)$$

$$\frac{\partial f^{\mathcal{A}}(\omega | \nu_0, \delta_0)}{\partial \nu_j} = \frac{\partial f^{\mathcal{A}}(\omega | \theta^*)}{\partial \nu_j}, \quad \frac{\partial f^{\mathcal{A}}(\omega | \nu_0, \delta_0)}{\partial \delta_i} = \sum_{\ell=1}^{k_\zeta} \mathbf{B}_{0, \ell i} \frac{\partial f(\omega | \theta^*)}{\partial \zeta_\ell} \quad (4.3)$$

for any $1 \leq j \leq k_\nu$ and $1 \leq i \leq k_\delta$, where $\mathbf{B}_{0, ij}$ is the (i, j) th element of the $k_\zeta \times k_\delta$ matrix $\mathbf{B}_0$. The

quadratic nature of the spectral density for SVMA models then means that the desired augmentation can be implemented via a linear expansion of the lag polynomial $\Psi(z \mid \nu, \eta^*)$ around η^* . The lag polynomial, in turn, is linear in the MA coefficients, so that linearizing each MA coefficient around η^* delivers an augmented model that satisfies condition (4.3).

Proposition 2 (Augmentation for Gaussian SVMA Models). *Suppose $\{X_t\}_{t \in \mathbb{Z}}$ is generated according to the model in (4.1), and that η^* is the plug-in value for η . Consider the augmented model with k_δ -dimensional auxiliary parameter δ and directional matrix $\mathbf{B}_0 \in \mathbb{R}^{k_\zeta \times k_\delta}$ specified as*

$$X_t = \mu^{\mathcal{A}}(\nu, \delta) + \sum_{j=0}^{\infty} \Psi_j^{\mathcal{A}}(\nu, \delta) \epsilon_{t-j}, \quad \epsilon_t \stackrel{iid}{\sim} \mathcal{N}[O_{n_\epsilon \times 1}, I_{n_\epsilon}] \quad (4.4)$$

with

$$\mu^{\mathcal{A}}(\nu, \delta) = \mu(\nu, \eta^*) + \frac{\partial \mu(\nu, \eta^*)}{\partial \zeta'} \mathbf{B}_0 \delta \quad (4.5)$$

$$\Psi_j^{\mathcal{A}}(\nu, \delta) = \Psi_j(\nu, \eta^*) + \sum_{r=1}^{k_\delta} \left(\sum_{\ell=1}^{k_\zeta} \mathbf{B}_{0,\ell r} \frac{\partial \Psi_j(\nu, \eta^*)}{\partial \zeta_\ell} \right) \delta_r \quad \forall j \geq 0. \quad (4.6)$$

where $\mathbf{B}_{0,\ell r}$ is the (ℓ, r) element of the matrix $\mathbf{B}_0$. Then, the augmented model satisfies conditions (2.3) and (2.4).

Proof Sketch. Condition (2.3) is immediate from the construction by putting $\delta = O_{k_\delta \times 1}$. Condition (2.4) follows from plugging in the augmented impulse responses $\Psi_j^{\mathcal{A}}(\nu, \delta)$ into the formula for the spectral density of Gaussian SVMA models, and then differentiating with respect to each parameter in (ν, δ) . Further details can be found in the online appendix. □

4.2 Application to the Krusell-Smith Model

Here we we apply the augmentation principle to a HA version of the real business cycle (RBC) model in the vein of Krusell and Smith (1998). HA models of this type naturally lend themselves to model augmentation due to the computational difficulty of estimating them via full-information methods.

In HA models the agents' steady state wealth distribution, along with steady state parameters such as the discount rate or relative risk aversion, determines the steady state values of aggregate macro variables. A simple example is the steady state real interest rate: in a representative agent model, the steady state value of r_t is given as $r_{ss} = \frac{1}{\beta_d} - 1$ for the discount rate β_d . In contrast, in the

Krusell and Smith (1998) model the same value is defined implicitly via the capital market clearing condition, where the interest rate determines capital supply both through the optimal saving policy and through the distribution of wealth across households.

The presence of the wealth distribution as an additional infinite-dimensional state variable destroys the simple mapping between the steady state parameters and the steady state values of aggregate macro variables, which now have to be recovered via costly bisection loops. Estimating these steady state parameters using maximum likelihood or Bayesian methods now requires one to recompute the steady state values for every new combination of parameters, rendering estimation extremely costly.

4.2.1 Model Setup and Steady State Computations

A continuum of households on the unit interval $[0, 1]$ have CRRA preferences, and choose consumption and supply fixed hours of labor, normalized to 1. These households own shares of capital from which they receive financial income. Households are subject to uninsurable idiosyncratic labor productivity shocks e_t as well as a borrowing constraint. Household i thus faces the problem

$$\max_{\{c_t^i\}_{t \geq 0}} \mathbb{E}_0 \left[\sum_{t=0}^{\infty} \beta_d^t \frac{c_t^{1-\sigma} - 1}{1-\sigma} \right] \quad \text{s.t.} \quad c_t^i + a_{t+1}^i = (1 + r_t)a_t^i + w_t e_t^i, \quad a_{t+1}^i \geq 0$$

where β_d is the discount rate, $\sigma > 0$ is the relative risk aversion coefficient. We assume that $\log(e_t^i) = m_e + \rho_e \log(e_{t-1}^i) + \sigma_e v_t^i$ for $v_t^i \sim \mathcal{N}[0, 1]$ that are iid across i and t . We denote the corresponding transition kernel for e_t^i by $P(\cdot, \cdot)$. Finally, w_t, r_t are the time t wage level and interest rate on capital respectively.

There is a representative firm using a Cobb-Douglas production function with capital share $\alpha \in (0, 1)$ to produce output using two factors, labor and capital. Production depends on TFP Z_t , so that the firm's problem at time t is maximizing profits $Z_t K_t^\alpha L_t^{1-\alpha} - w_t L_t - (r_t + \delta_d)K_t$ with respect to the factors K_t, L_t , where $\delta_d \in (0, 1)$ is the depreciation rate of capital. We assume that $\log(Z_t) = \rho_z \log(Z_{t-1}) + \sigma_z \epsilon_t$ for $\epsilon_t \sim \mathcal{N}[0, 1]$ that is iid across time. The market clearing conditions require that $K_t = \int_0^1 a_t^i di$ and $L_t = \int_0^1 e_t^i di$.

A recursive equilibrium for this model is standard and its definition can be found in, say, Winberry (2018). It is characterized by a collection of value, policy and pricing functions satisfying the household and firm optimality conditions, the market clearing conditions for capital and labor, and finally a specification for how the wealth distribution evolves alongside a condition that ensures household beliefs are consistent.

When the economy is at the steady state, so that $Z_t = 1$ for $t \geq 0$, the model reduces to that of Aiyagari (1994). Denoting the steady state prices as r_{ss} and w_{ss} , by the labor market

clearing condition and the firm's first order conditions the steady state wage w_{ss} is given as $w_{ss} = (1 - \alpha) \left(\frac{\alpha}{r_{ss} + \delta} \right)^{\frac{\alpha}{1-\alpha}}$. The steady state interest rate r_{ss} is then found as the value that makes capital demand equal to capital supply. Capital demand as a function of the interest rate is given by $K^d(r) = \left(\frac{\alpha}{r + \delta} \right)^{\frac{1}{1-\alpha}} L^*$, where $L^* = \frac{m_e}{1 - \rho_e^2}$ is aggregate effective labor. On the other hand, capital supply $K^s(r)$ necessitates solving for the steady-state savings policy function for a given interest rate r and then deriving the implied stationary wealth distribution to compute aggregate savings given r .

The first step requires iterating through either the value function or the policy function until convergence is reached; and the second step requires iterating over the transition kernel induced by the policy function found in the first step until convergence. These steps make the evaluation of $K^s(r)$ for even a single candidate interest rate r , and subsequently the search for the steady state interest rate r_{ss} , very costly. Here the nuisance parameter η consists only of the capital share α , since it is strongly identified via steady state output Y_{ss} . The risk aversion coefficient σ and the heterogeneity parameters m_e, ρ_e, σ_e^2 are not included in η because they are weakly identified or entirely unidentified by aggregate macro data. The discount rate β_d is identified via the steady state interest rate r_{ss} , but it is doubtful whether it is strongly identified in a model with output as the only macro variable used to construct the likelihood. Likewise, the depreciation rate δ_d is potentially identified through the steady state wage level or the capital-output ratio, which once again involves aggregate macro variables that are not included in the model.

4.2.2 The SVMA Representation and Jacobians

In this model, the parameters of interest are the parameters governing the evolution of TFP, collected into $\beta = \{\rho_z, \sigma_z^2\}$. Let $\mathbf{G}(\eta) = \frac{\partial \mathbf{Y}}{\partial \mathbf{Z}}(\eta)$ be the general equilibrium Jacobian for output whose $(t + 1, s + 1)$ th element is $\frac{\partial Y_t}{\partial Z_s}(\eta)$, or the impulse response of time t output to a deviation at time s of TFP from its steady state value, and define $d\mathbf{Z}(\beta) = \sigma_z \times (1, \rho_z, \rho_z^2, \dots)$. It is shown in Auclert et al. (2021) that the DGP for Y_t can be approximated via the MA process

$$Y_t \approx Y_{ss}(\eta) + \sum_{j=0}^{T_{MA}-1} \Psi_j(\beta, \eta) \epsilon_{t-j}$$

for large enough T_{MA} that need not equal the sample size T^5 . Here the j th impulse response coefficient is defined as $\Psi_j(\beta, \eta) = (\mathbf{G}(\eta) d\mathbf{Z}(\beta))_j$, or as the j th element of the vector $\mathbf{G}(\eta) d\mathbf{Z}(\beta)$, and $Y_{ss}(\eta)$ is the steady state value of output as a function of the steady state parameters η .

We can obtain $\mathbf{G}(\eta)$ from the sequence space Jacobian, which for the present model is the

⁵The choice of T_{MA} is based on macroeconomic considerations; see Auclert et al. (2021), where it is set equal to 300.

matrix $\frac{\partial \mathbf{K}}{\partial \mathbf{Z}}(\eta)$ whose $(t + 1, s + 1)$ th element is $\frac{\partial K_{t+1}}{\partial Z_s}(\eta)$, or the impulse response of time t saving to a deviation in time s TFP from its steady state value. Since time t output Y_t is given by $Y_t = Z_t K_t^\alpha (L^*)^{1-\alpha}$, it follows from the chain rule that

$$\mathbf{G}(\eta) = Y_{ss}(\eta) I_{T_{MA}} + (r_{ss}(\eta) + \delta_d) \frac{\partial \mathbf{K}}{\partial \mathbf{Z}}(\eta) I_{T_{MA}} (-1)$$

where $I_{T_{MA}}(-1)$ is the $T_{MA} \times T_{MA}$ matrix with its 1s on its first subdiagonal and zeros elsewhere. The only missing piece required to compute the impulse responses is the sequence space Jacobian $\frac{d\mathbf{K}}{d\mathbf{Z}}(\eta)$, for which we use the Fake news algorithm detailed in Auclert et al. (2021).

4.2.3 Augmentation of the Likelihood

Let the baseline calibration of η be denoted by η^* . Then, the general augmented model is constructed as

$$Y_t = Y_{ss}(\eta^*) + \frac{\partial Y_{ss}(\eta^*)}{\partial \eta} \delta + \sum_{j=0}^{T_{MA}-1} \Psi_j^{\mathcal{A}}(\beta, \delta) \epsilon_{t-j}$$

where the expressions for $\Psi_j^{\mathcal{A}}(\beta, \delta)$ are given in equation (4.6). Collecting the augmented MA coefficients into the vector $\Psi^{\mathcal{A}}(\beta, \delta)$, since $\Psi(\beta, \eta^*) = \mathbf{G}(\eta^*) d\mathbf{Z}(\beta)$, we can express

$$\Psi^{\mathcal{A}}(\beta, \delta) = \left\{ \mathbf{G}(\eta^*) + \frac{\partial \mathbf{G}(\eta^*)}{\partial \eta} \delta \right\} d\mathbf{Z}(\beta),$$

As such, the MA coefficients for the augmented model are also separable in δ and β . Computationally, this decomposition indicates that we need only compute the derivatives $\frac{\partial \mathbf{G}(\eta^*)}{\partial \eta}$ once and reuse them to construct $\Psi^{\mathcal{A}}(\beta, \delta)$ for different pairs (β, δ) . The same goes for the mean augmentation, where we reuse the derivative $\frac{\partial Y_{ss}(\eta^*)}{\partial \eta}$ that is computed once before sampling (β, δ) from the augmented model.

There is a computational issue related to the derivatives $\frac{\partial Y_{ss}(\eta^*)}{\partial \eta}$ and $\frac{\partial \mathbf{G}(\eta^*)}{\partial \eta}$. The GE Jacobian $\mathbf{G}(\eta^*)$ is constructed using the Fake news algorithm, which in turn involves steady state values of output and the interest rate that are obtained through expensive bisection loops. For the augmentation to remain practical, we therefore obtain each $\frac{\partial \mathbf{G}(\eta^*)}{\partial \eta}$ via two-sided numerical differentiation; the details are contained in the online appendix. $\frac{\partial Y_{ss}(\eta^*)}{\partial \eta}$ is obtained in a similar manner.

Let $H(\beta, \eta)$ be the (scalar) object of interest, and denote by $D_\eta H(\beta, \eta)$ the derivative of $H(\cdot)$ with respect to η . We construct a posterior sample for $H(\beta, \eta)$ as follows:

1. *Estimating the Augmented Model.* Draw a sample from the posterior of (β, δ) from the augmented model, denoted by $\{\beta^{(i)}, \delta^{(i)}\}_{1 \leq i \leq N}$.

2. *Forming the Posterior Sample of $H(\beta, \eta)$.* Construct the posterior sample of $H(\beta, \eta)$ as

$$\left\{ H(\beta^{(i)}, \eta^*) + \frac{\partial H(\beta^{(i)}, \eta^*)}{\partial \eta} \delta^{(i)} \right\}_{1 \leq i \leq N}.$$

4.2.4 Prior Selection and Posterior Sampler

For ρ_z we impose a beta prior with mean 0.5 and standard deviation 0.2, and for σ_z^2 an inverse gamma prior with mean 0.16 and standard deviation 0.1. Posterior results turn out to be relatively insensitive to these choices.

The prior variance of δ is informed by the prior information on the nuisance parameters η . Suppose the prior information on η suggests that the variability in the calibrated value η^* is summarized in the positive definite covariance matrix Σ_η^* . Then, in the model used to conduct inference on the shock parameters, we impose a normal prior on δ with mean zero and variance Σ_η^* , with the Mahalanobis distance restriction mentioned in section 2.6. The choice of prior variability Σ_η^* is in turn informed by the literature dealing with standard errors for calibrated parameters. A principled approach is to elicit Σ_η^* from steady state targets, as in Del Negro and Schorfheide (2008). We consider a simple version of this elicitation process. Let $\eta \mapsto X_{ss}(\eta)$ be the mapping from steady-state parameters to the steady-state values of endogenous variables, where $X_{ss}(\cdot)$ and η have the same dimension, and Σ_x the diagonal matrix specifying how much prior variation we are willing to allow for each steady-state value. If $X_{ss}(\cdot)$ has a non-singular derivative at η^* , then denoting the derivative by $DX_{ss}(\eta^*)$, the induced prior covariance for η is

$$\Sigma_\eta^* = (DX_{ss}(\eta^*))^{-1} \Sigma_x (DX_{ss}(\eta^*))^{-1'}.$$

For the present Krusell-Smith model, we choose to target the steady state value of output, that is, we elicit Σ_η^* from the steady state mapping $\alpha \mapsto Y_{ss}(\alpha)$. The prior variance for $\eta = \alpha$ is chosen so that one prior standard deviation of δ represents a 15% change in the steady state value of output.

For sampling we use the random-walk Metropolis-Hastings (RWMH) algorithm as described in Herbst and Schorfheide (2015). The proposal variance is computed as the inverse negative Hessian of the posterior density evaluated at the posterior mode, where we use automatic differentiation to obtain the Hessian up to machine precision. The proposal variance is multiplied by a scaling parameter $\tau > 0$ that we tune over the burn-in phase to achieve the desired acceptance rate.

4.3 Simulation Results for the Krusell-Smith Model

We conduct a simulation study to investigate the effect of local misspecification and evaluate how the augmented model performs under such forms of misspecification. The chosen baseline calibration

Table 2: **Monte Carlo study under local misspecification.** We report the coverage of nominal 95% credible intervals, the mean absolute error of the posterior mode, and the average posterior standard deviation for the shock parameters ρ_z and σ_z^2 .

h	T	Model	Coverage		Mean Absolute Error		Posterior SD	
			ρ_z	σ_z^2	ρ_z	σ_z^2	ρ_z	σ_z^2
$h = 0.3$	500	Baseline	0.96	0.22	0.0256	0.0192	0.0317	0.0075
		Augmented	0.95	0.96	0.0269	0.0128	0.0321	0.0175
	1000	Baseline	0.96	0.22	0.0171	0.0127	0.0227	0.0050
		Augmented	0.98	0.94	0.0179	0.0098	0.0229	0.0126
$h = 0.5$	500	Baseline	0.95	0.01	0.0255	0.0347	0.0313	0.0085
		Augmented	0.96	0.95	0.0268	0.0140	0.0323	0.0179
	1000	Baseline	0.95	0.00	0.0172	0.0227	0.0226	0.0055
		Augmented	0.97	0.94	0.0179	0.0102	0.0229	0.0128

Notes: Coverage is computed from nominal 95% posterior credible intervals over 100 Monte Carlo replications. Mean absolute error is the Monte Carlo average of $|\hat{\theta}_{\text{mode}} - \theta_0|$. Posterior standard deviations are averaged across Monte Carlo replications. The nuisance parameter is the capital share α , with local misspecification $\alpha_0 = \alpha^* + h/\sqrt{T}$ and $\alpha^* = 0.33$.

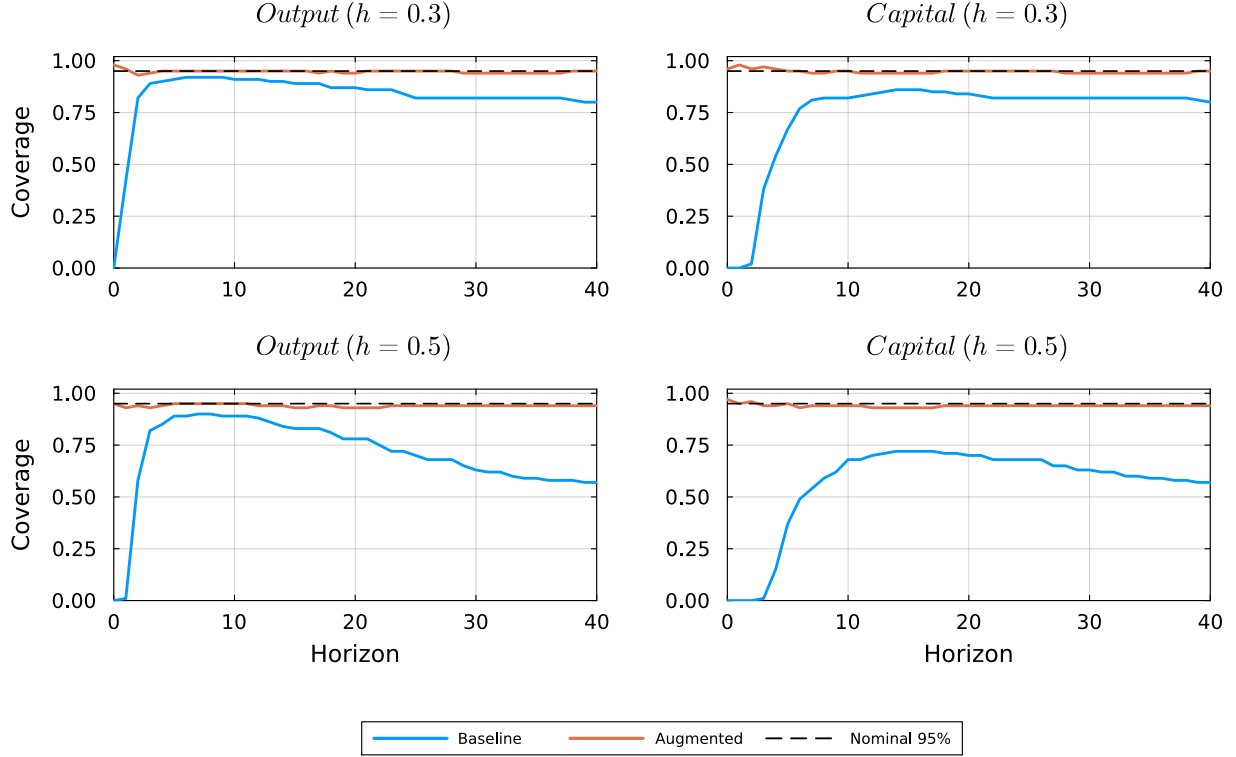
is $\alpha^* = 0.33$, $\delta_d = 0.025$, $\rho_e = 0.966$ and $\sigma_e^2 = 0.5^2(1 - \rho_e^2)$. β_d and m_e are chosen to match the steady state targets $Y_{ss} = 1$ and $r_{ss} = 0.01$ for output and the real interest rate respectively. For each sample size T , the artificial output gap data is generated under parameter values that are the same as the baseline calibration except for α , which is locally perturbed as $\alpha_0 = \alpha^* + \frac{h}{\sqrt{T}}$ for $h = 0.3$ or 0.5 .

4.3.1 Inference on Shock Parameters ρ_z, σ_z^2

The simulation results in table 2 tell a consistent story across sample sizes T and the magnitude of local misspecification. Throughout the simulations, the actual coverage rate of the credible interval around ρ_z is close to its nominal rate of 95% for both the baseline and augmented models, indicating that local misspecification in α does not significantly distort inference on shock persistence. However, local calibration error in α is enough to completely shift the posterior of σ_z^2 away from the truth in the baseline model, with the nominal 95% credible interval for σ_z^2 failing to cover the true value of σ_z^2 almost all the time when $h = 0.5$. The large bias incurred by the baseline model's posterior for σ_z^2 , coupled with a relatively small standard deviation, leads to a posterior that is overly confident for the wrong target.

Meanwhile, the actual coverage rate for σ_z^2 in the augmented model recovers close to the nominal 95% rate regardless of the magnitude of the local misspecification in α . The restoration of valid inference for σ_z^2 can be attributed to a combination of reduced posterior bias and considerably larger posterior standard deviations—mean absolute error falls by roughly 30% for $h = 0.3$ and

Figure 2: **Monte Carlo study for impulse response coverage.** Below we display the coverage of pointwise credible bands for the impulse response of output and capital to TFP shocks up to 40 horizons. The impulse responses are evaluated for the baseline and augmented models. The nuisance parameter is the capital share α , with local misspecification $\alpha_0 = \alpha^* + h/\sqrt{T}$ and $\alpha^* = 0.33$.



by almost 60% for $h = 0.5$, while the standard deviation nearly doubles when moving from the baseline to the augmented model. Model augmentation both recenters the posterior of the shock variance and allows it to better reflect the underlying calibration uncertainty.

4.3.2 Inference on Impulse Responses

Local misspecification in the steady state parameters also shifts the impulse responses of the endogenous macro variables to TFP shocks by altering their GE Jacobians. Under the same local misspecification setting for α as above, we compare the actual coverage of nominal 95% pointwise credible bands for the impulse response of output and capital to TFP shocks across the baseline and augmented models. When obtaining the posterior sample of the impulse response of output, we use the composition function $H(\beta, \eta) = \mathbf{G}(\eta)dZ(\beta)$; for the impulse response of capital, the composition function is $H(\beta, \eta) = \frac{\partial \mathbf{K}}{\partial \mathbf{Z}}(\eta)dZ(\beta)$, where we use the sequence space Jacobian $\frac{\partial \mathbf{K}}{\partial \mathbf{Z}}(\eta)$ in place of the GE Jacobian $\mathbf{G}(\eta)$.

Table 3: **Average 95% credible-band widths under local misspecification.** Average band widths are reported for the output and capital impulse responses under $T = 500$ and local misspecification $\alpha_0 = \alpha^* + 0.5/\sqrt{T}$. The final row reports the ratio of augmented to baseline average width.

	Output			Capital		
	1–4	5–12	13–40	1–4	5–12	13–40
Baseline	0.1681	0.1140	0.0260	0.2761	0.7773	0.5754
Augmented	0.2389	0.1275	0.0333	0.6879	1.1204	0.7954
Augmented / Baseline	1.42	1.12	1.28	2.49	1.44	1.38

Notes: Each entry is the average width of the nominal 95% pointwise credible band over the indicated horizon block. Horizon 0 is excluded. A ratio above one indicates that the augmented credible bands are wider on average.

Figure 2 shows us that impulse responses obtained from the baseline model undercovers relative to the nominal 95% rate for short and long horizons. For $h = 0.3$ the undercoverage is moderate, with the credible bands maintaining around 75% coverage for medium and longer horizons—but for local misspecification of a larger magnitude coverage drops to near 50% at the longest horizons and the baseline model fails to cover the true impulse response almost all the time for the shortest horizons. The augmented model, meanwhile, maintains close to nominal coverage throughout. Thus, in the presence of even local calibration error, the usual fixed calibration approach to inference can be invalidated not only for the shock variance parameter σ_z^2 but also for economically meaningful objects such as impulse responses. Since the choice of whether to calibrate the capital share as 0.33 or 0.36 often comes down to the vintage of the data, model augmentation is a way to protect subsequent inference against the possibility of making the incorrect choice.

Table 3 displays the average widths of the impulse response credible intervals for short (1 to 4 periods ahead), medium (5 to 12 periods ahead), and long (13 to 40 periods ahead) horizons, computed across Monte Carlo iterations. Across all horizons, the credible intervals produced by the augmented model are wider than their baseline model counterparts, but not by too much. Except for the shortest horizons in the case of the capital impulse responses, the credible bands reported by the augmented model are never more than 1.5 times wider than those of the baseline model. Model augmentation thus restores valid inference without incurring an exorbitantly large efficiency cost.

4.3.3 Updating the Capital Share Calibration

In section 2.4.1 we noted how, if we restore the full η -score instead of just the directions that are relevant for locally robust estimation of β , then it becomes possible to conduct inference on composite objects $H(\theta)$ that depend directly on η . This is especially useful if we are interested in quantifying the uncertainty in an object that depends only on η , since it would be treated as known

Table 4: **Updating the calibrated parameter.** The table compares the calibration-induced prior for the capital share α with the posterior implied by the augmented model under local misspecification with $h = 0.3$.

T	DGP Value	Prior 95% CI	Posterior Median	Posterior 95% CI	Post./Prior SD
500	0.3434	[0.2863, 0.3738]	0.3394	[0.3163, 0.3625]	0.53
1000	0.3395	[0.2863, 0.3738]	0.3372	[0.3198, 0.3552]	0.40

Notes: Posterior medians, credible interval bounds, and posterior standard deviations are averaged across 100 Monte Carlo replications. The final column reports the ratio of the average posterior standard deviation to the standard deviation of the calibration-induced prior. The baseline calibration is $\alpha^* = 0.33$, and the DGP satisfies $\alpha_0 = \alpha^* + 0.3/\sqrt{T}$.

in a baseline model that fixes η at η^* . In contrast, in the augmented model the posterior draws of the auxiliary parameter δ can be transformed in a way that yields a posterior sample of $H(\eta)$ that is robust to local calibration error in η^* .

The simplest examples of composite objects $H(\theta)$ that depend only on η are the calibrated parameters themselves. Given a posterior sample $\{\delta^{(i)}\}_{1 \leq i \leq N}$ of the auxiliary parameters, it was established as a special case in section 2.4.1 that a posterior sample of η that is robust to local misspecification in η^* can be obtained as

$$\left\{ \eta^* + \delta^{(i)} \right\}_{1 \leq i \leq N}.$$

95% posterior credible intervals for α constructed in this manner are given in table 4, where the data is generated under the same simulation design as in the parsimonious case. We can immediately see that, on average across the Monte Carlo iterations and sample sizes, the point estimators of α are very close to their DGP values. In addition to diagnosing any potential misspecification in the calibration, the results in table 4 serves as a means of determining how informative the aggregate macro variables used in the construction of the structural likelihood are for the steady-state parameters. For α the width of the posterior interval compared to that of the prior interval shrinks dramatically as the sample size increases, indicating that aggregate output is highly informative about α .

5 Conclusion

In this paper we devise, for two broad classes of parametric structural models, a practically implementable means of robustifying likelihood-based inference against local misspecification in plug-in values. Robustification proceeds by constructing a larger parametric model that nests the original model with fixed plug-in values and restores the elements of the score function that are

lost when those values are treated as fixed. For models with iid data in the exponential family and time-series models admitting a Gaussian SVMA representation, simple linearization rules deliver an augmented model with the same first-order score geometry as the original structural model while allowing the auxiliary parameter δ to enter in a much simpler manner than the original nuisance parameter η . The DDC and Krusell–Smith applications illustrate the resulting computational and inferential gains: augmentation avoids repeated solution of expensive structural blocks, restores robustness to local misspecification in first-stage or calibrated parameters, and allows the data to update those parameters without resorting to full-information joint estimation.

The augmentation principle in the paper is derived under the assumption of local identifiability at fixed plug-in values, defined mathematically as non-singularity of the Fisher information at that point. However, it is often the case that parameters in structural models are weakly identified or not identified at all. For instance, the discount factor in DDC models is generally unidentified, and even local identification of steady-state parameters in DSGE models is never guaranteed (e.g. Canova and Sala (2009), Iskrev (2010)). The BvM theorem generally breaks down in the presence of the failure of strong identification, and therefore is beyond the scope of the current augmentation procedure.

Nevertheless, once we do have local identification, the augmented method represents a middle ground between two extremes in the tradition of Bayesian estimation of structural models. Bayesian inference in structural models has traditionally forced practitioners to choose between two imperfect alternatives: fully joint estimation, which may be computationally infeasible or may disregard valuable external information, and fixed plug-in estimation, which ignores calibration uncertainty. Model augmentation offers a third alternative. By preserving the computational advantages of plug-in estimation while recovering the first-order score information lost through calibration, it provides a practical route to robust Bayesian inference in a broad class of structural models.

References

- Acharya, Sushant, William Chen, Marco Del Negro, Keshav Dogra, Aidan Gleich, Shlok Goyal, Ethan Matlin, Donggyu Lee, Reza Sarfaty, and Sikata Sengupta (2023). “Estimating HANK for central banks”. In: *FRB of New York Staff Report* 1071.
- Aguirregabiria, Victor and Pedro Mira (2002). “Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models”. In: *Econometrica* 70.4, p. 1519.
- Aguirregabiria, Victor, Pedro Mira, and Hernan Roman (2007). “An estimable dynamic model of entry, exit, and growth in oligopoly retail markets”. In: *American Economic Review* 97.2, pp. 449–454.

- Aiyagari, S Rao (1994). “Uninsured idiosyncratic risk and aggregate saving”. In: *The Quarterly Journal of Economics* 109.3, pp. 659–684.
- Andrade, Jose Ailton Alencar and Anthony O’Hagan (2011). “Bayesian robustness modelling of location and scale parameters”. In: *Scandinavian Journal of Statistics* 38.4, pp. 691–711.
- Auclert, Adrien, Bence Bardóczy, Matthew Rognlie, and Ludwig Straub (2021). “Using the sequence-space Jacobian to solve and estimate heterogeneous-agent models”. In: *Econometrica* 89.5, pp. 2375–2408.
- Bajari, Patrick, C Lanier Benkard, and Jonathan Levin (2007). “Estimating dynamic models of imperfect competition”. In: *Econometrica* 75.5, pp. 1331–1370.
- Bickel, Peter J, Chris AJ Klaassen, Ya’acov Ritov, and Jon A Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer.
- Brock, William A and Steven N Durlauf (2001). “Discrete choice with social interactions”. In: *The Review of Economic Studies* 68.2, pp. 235–260.
- Canova, Fabio and Luca Sala (2009). “Back to square one: Identification issues in DSGE models”. In: *Journal of Monetary Economics* 56.4, pp. 431–449.
- Chahrour, Ryan and Kyle Jurado (2022). “Recoverability and expectations-driven fluctuations”. In: *The Review of Economic Studies* 89.1, pp. 214–239.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). *Double/debiased machine learning for treatment and structural parameters*.
- Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins (2022). “Locally robust semiparametric estimation”. In: *Econometrica* 90.4, pp. 1501–1535.
- Chernozhukov, Victor and Han Hong (2003). “An MCMC approach to classical estimation”. In: *Journal of econometrics* 115.2, pp. 293–346.
- Del Negro, Marco and Frank Schorfheide (2008). “Forming priors for DSGE models (and how it affects the assessment of nominal rigidities)”. In: *Journal of Monetary Economics* 55.7, pp. 1191–1208.
- Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri (2015). “Prior selection for vector autoregressions”. In: *Review of Economics and Statistics* 97.2, pp. 436–451.
- Greco, Luca, Walter Racugno, and Laura Ventura (2008). “Robust likelihood functions in Bayesian inference”. In: *Journal of Statistical Planning and Inference* 138.5, pp. 1258–1270.
- Hansen, Lars Peter and Thomas J Sargent (1981). “Exact linear rational expectations models: Specification and estimation”. In: *Rational Expectations Econometrics*. CRC Press, pp. 45–75.
- Herbst, Edward P and Frank Schorfheide (2015). “Bayesian estimation of DSGE models”. In: *Bayesian Estimation of DSGE Models*. Princeton University Press.

- Hotz, V Joseph and Robert A Miller (1993). “Conditional choice probabilities and the estimation of dynamic models”. In: *The Review of Economic Studies* 60.3, pp. 497–529.
- Imai, Susumu, Neelam Jain, and Andrew Ching (2009). “Bayesian estimation of dynamic discrete choice models”. In: *Econometrica* 77.6, pp. 1865–1899.
- Iskrev, Nikolay (2010). “Local identification in DSGE models”. In: *Journal of Monetary Economics* 57.2, pp. 189–202.
- Jørgensen, Thomas H (2023). “Sensitivity to calibrated parameters”. In: *Review of Economics and Statistics* 105.2, pp. 474–481.
- Kasahara, Hiroyuki and Katsumi Shimotsu (2008). “Pseudo-likelihood estimation and bootstrap inference for structural discrete Markov decision models”. In: *Journal of Econometrics* 146.1, pp. 92–106.
- Kleijn, B. J. K. and A. W. van der Vaart (2012). “The Bernstein-Von-Mises theorem under misspecification”. In: *Electronic Journal of Statistics* 6, pp. 354–381.
- Krusell, Per and Anthony A Smith Jr (1998). “Income and wealth heterogeneity in the macroeconomy”. In: *Journal of political Economy* 106.5, pp. 867–896.
- Liu, Laura and Mikkel Plagborg-Møller (2023). “Full-information estimation of heterogeneous agent models using macro and micro data”. In: *Quantitative Economics* 14.1, pp. 1–35.
- Liu, Laura, Mikkel Plagborg-Møller, and Nelson Matthew P Tan (2026). *Limited-Information Estimation of Heterogeneous Agent Models*. Working paper.
- Melosi, Leonardo, Hanno Kase, and Matthias Rottner (2022). “Estimating Nonlinear Heterogeneous Agents Models with Neural Networks”. In.
- Müller, Ulrich and Andriy Norets (2026). “Locally Robust Semiparametrically Efficient Bayesian Inference”. In: *Econometrica* 94.5, pp. 1761–1777.
- Pesendorfer, Martin and Philipp Schmidt-Dengler (2008). “Asymptotic least squares estimators for dynamic games”. In: *The Review of Economic Studies* 75.3, pp. 901–928.
- Plagborg-Møller, Mikkel (2019). “Bayesian inference on structural impulse response functions”. In: *Quantitative Economics* 10.1, pp. 145–184.
- Rust, John (1987). “Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher”. In: *Econometrica: Journal of the Econometric Society*, pp. 999–1033.
- Smets, Frank and Rafael Wouters (2007). “Shocks and frictions in US business cycles: A Bayesian DSGE approach”. In: *American economic review* 97.3, pp. 586–606.
- van der Laan, Mark J. and Daniel Rubin (2006). “Targeted Maximum Likelihood Learning”. In: *The International Journal of Biostatistics* 2.1, pp. 1–38.
- van der Vaart, A. W. (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.
- Winberry, Thomas (2018). “A method for solving and estimating heterogeneous agent macro models”. In: *Quantitative Economics* 9.3, pp. 1123–1151.