Asset Pricing and Term Structure Models

Seung Hyun Kim

April 30, 2024

Chapter 1

Equilibrium Models of Asset Pricing

The term structure literature is a part of the asset pricing literature, as it aims to study the relationship among the yields of zero-coupon bonds (which, we see below, is equivalent to working with their prices). As such, we first introduce here some fundamental results of asset pricing.

1.1 Fundamentals of Asset Pricing

We primarily work with the following setup. Let there be an asset with a sequence of payoffs $\{X_t\}_{t\in N_+}$. The time $t+1$ **payoff** of the asset X_{t+1} is understood to be the payoff the investor would recieve at time $t+1$ if she were to invest in (purchase) the asset at time t and then sell it at time $t+1$. Consider the following examples:

Stocks If $\{X_t\}_{t\in N_+}$ is a sequence of payoffs for a stock share, then X_{t+1} would be the stock price at time $t+1$ plus the dividends for one period.

Bonds If $\{X_t\}_{t\in N_+}$ is a sequence of payoffs for a bond, then X_{t+1} is the bond price at time $t+1$ plus the coupon paid at time $t+1$. In particular, a zero-coupon bond's time $t+1$ payoff X_{t+1} equals the bond price at time $t+1$.

As the name suggests, asset pricing aims to assign a price p_t to the asset at time t as a function of its payoff X_{t+1} at time $t+1$. Specifically, we want to find the time t price p_t of the asset as a function of X_{t+1}, that is, as

$$p_t = \pi_t(X_{t+1}).$$

The **rate of return** of this asset at time $t+1$ is defined as

$$r_{t+1} = \frac{X_{t+1} - p_t}{p_t} = \frac{X_{t+1}}{p_t} - 1,$$

that is, the rate of return at time $t+1$ from investing in the asset at time t. The asset's **return** at time $t+1$ is defined as $1+r_{t+1}$, which is the payoff the investor stands to receive if she invests exactly one dollar in the asset. Since the price of an asset can be formulated in terms of the rate of return and the payoff as

$$p_t = \frac{X_{t+1}}{1+r_{t+1}},$$

asset pricing is equivalent to the study of the rate of return of the asset.

In general, since the payoff at time $t+1$ is unknown at time t, neither is the rate of return r_{t+1}. This means that the return $1+r_{t+1}$ is a random variable given the information up to time t; it is in this sense that we say assets are **risky**.

An asset whose rate of return at time $t+1$ is known at time t is called a **risk-free asset**, and its rate of return at time $t+1$ is denoted by r^f_{t+1} [1]. The reason this asset is risk-free is because its time $t+1$ return $1+r^f_{t+1}$ is a known at time t. Equivalently, it is a degenerate random variable, that is, a random variable with variance 0, conditional on all the information up to time t.

The amount of risk that an asset possesses at time t is represented by the **variance** $\mathrm{Var}_t(r_{t+1})$ of its one-period ahead rate of return r_{t+1}. In general, the higher this variance, the riskier the asset. On the other hand, the profitability of the asset at time t is represented by its **expected rate of return** $\mathbb{E}_t[r_{t+1}]$. An asset is said to be high-risk and high-return if $\mathbb{E}_t[r_{t+1}]$ and $\mathrm{Var}_t(r_{t+1})$ are both large; likewise, it is low-risk and low-return if these quantities are both small.

The time t **risk premium**, or **expected excess return** of an asset is defined as

$$RP_{t+1} = \mathbb{E}_t[r_{t+1}] - r^f_{t+1};$$

note that this quantity is known at time t, since both the time t expectation of r_{t+1} and the risk-free rate of return r^f_{t+1} are known at time t. Heuristically, the risk premium represents the compensation that an investor receives in exchange for taking on risk. Generally, the riskier the asset, the higher its risk premium.

Broadly speaking, there are two strands of the asset pricing literature. One strand aims to price assets using an equilibrium approach, or in other words, as the result of the utility maximization of rational investors. Notable examples of this approach are the CAPM and C-CAPM. The other strives to price assets under minimal assumptions, usually in the form of the no-arbitrage condition. Examples of this approach are the Black-Scholes

[1] Some authors prefer to use the time subscript t for the time $t+1$ rate of return of a risk-free asset, to indicate that the quantity is known at time t, but for the sake of notational consistency we retain the time subscript $t+1$.

model for option pricing and term structure models. In this chapter we focus on the first type of model, and in the next move onto the second type.

1.2 The Capital Asset Pricing Model (CAPM)

We start with the earliest and simplest asset pricing model, the CAPM. For brevity, we do not discuss the Markowitz portfolio theory[2] that laid the foundations for the CAPM. Our assumptions are as follows:

A1. Many Homogeneous Investors

We assume that there exists a large number of investors who are homogeneous in the sense that they possess the same utility function. This allows for the model to admit a representative investor.

A2. Two-Period Model

The economy lasts for two periods 0 and 1. The investors form their portfolios at time 0 and receive the payoffs at time 1. Thus, we can omit the time subscripts.

A3. Many Risky Assets and One Risk-free Asset

We assume that there exists n risky assets with rates of return $r_1, \cdots, r_n$, and one risk-free asset with rate of return r_f. Investors form their portfolios by choosing the weights $w_1, \cdots, w_n$ on the risky assets, which implies a weight of $1 - \sum_{i=1}^{n} w_i$ on the risk-free asset.

The mean and covariance matrix of the vector of risky asset rates of return $r = (r_1, \cdots, r_n)$ are given by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}$$

As usual, we assume that Σ is a positive definite $n \times n$ matrix.

A portfolio with weights $w = (w_1, \cdots, w_n)$ has expected rate of return

$$\mu_p = w'\mu + (1 - w'\iota) r_f$$

[2] See Markowitz (1952).

and variance

$$\sigma_p^2 = w'\Sigma w,$$

where $\iota \in \mathbb{R}^n$ is a vector of ones.

A4. Mean-Variance Utility

Given a portfolio with expected rate of return r and variance σ^2, the representative investor receives utility equal to

$$u(er, \sigma^2) = er - \frac{A}{2}\sigma^2.$$

In other words, the expected rate of return is a good and the variance is a bad.

The representative investor must choose the vector of weights $w = (w_1, \cdots, w_n)$ as the solution to the following maximization problem:

$$\max_{w \in \mathbb{R}^n} \quad u(\mu_p, \sigma_p^2) = w'\mu + (1 - w'\iota)r_f - \frac{A}{2}w'\Sigma w.$$

Letting w^* be a solution to the above problem, the first order condition for maximization tells us that

$$\mu - \iota \cdot r_f - A \cdot \Sigma w = O_{n\times 1}.$$

Therefore, for any $1 \leq i \leq n$,

$$\mu_i - r_f = A \cdot \sum_{j=1}^{n} \sigma_{ij} w_j^*.$$

The optimal portfolio, also called the market portfolio, has expected rate of return

$$\mu_M = \sum_{i=1}^{n} w_i^*(\mu_i - r_f) + r_f$$

and variance

$$\sigma_M^2 = w^{*\prime}\Sigma w^* = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{ij} w_i^* w_j^*.$$

The first order condition above then implies that

$$\mu_M - r_f = \sum_{i=1}^{n} w_i^*(\mu_i - r_f) = A\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}w_i^*w_j^* = A\cdot\sigma_M^2.$$

Furthermore, for any $1 \leq i \leq n$,

$$\mathrm{Cov}\,(r_i, r_M) = \mathrm{Cov}\left(r_i, \sum_{j=1}^{n} w_j^* \cdot r_j\right) = \sum_{j=1}^{n} w_j^* \cdot \mathrm{Cov}\,(r_i, r_j) = \sum_{j=1}^{n}\sigma_{ij}w_j^*.$$

It follows that

$$\begin{aligned}\mu_i - r_f &= A\cdot\sum_{j=1}^{n}\sigma_{ij}w_j^* \\ &= \frac{\mu_M - r_f}{\sigma_M^2}\cdot\mathrm{Cov}\,(r_i, r_M) \\ &= \frac{\mathrm{Cov}\,(r_i, r_M)}{\sigma_M^2}\left(\mu_M - r_f\right).\end{aligned}$$

In other words, the risk premium of asset i is given as the product of the beta term

$$\beta_i = \frac{\mathrm{Cov}\,(r_i, r_M)}{\sigma_M^2}$$

and the expected excess return of the market portfolio

$$\mu_M - r_f.$$

Note that $\mu_M - r_f$ can be interpreted as a common factor that determines the risk premium of each asset, with the coefficient β_i determining how much the factor loads onto the risk premium of asset i.

1.3 The Consumption-CAPM (C-CAPM)

We now take a more general approach to the equilibrium pricing method studied above. Most notably, we depart from the assumption of mean-variance utility, and consider a multi-period portfolio selection problem. Our assumptions are as follows:

A1. Many Homogeneous Investors

We assume that there exists a large number of investors who are homogeneous in the sense that they possess the same utility function. This allows for the model to admit a representative investor.

A2. Multi-Period Model

The economy starts at time 0, and is populated by infinitely lived investors.

A3. Many Risky Assets and One Risk-free Asset

We assume that there is one consumption good, whose price is normalized to 1. In addition, there exists n assets with time $t+1$ rates of return

$$r_{t+1}^1, \cdots, r_{t+1}^n$$

, of which the first asset is assumed to be the single risk-free asset with time $t+1$ rate of return r_{t+1}^f. At time $t+1$, the representative investor holds a_{t+1}^i units of asset i. At time 0, the representative investor is assumed to be endowed with a_0^i units of asset i.

At time t, the price and dividend of asset i is given as p_t^i and d_t^i, respectively.

A4. General Utility Function

The representative investor has instantaneous utility function $u : \mathbb{R}_+ \to \mathbb{R}$ such that $u'(\cdot) > 0$ and $u''(\cdot) < 0$. This makes it so that the investors are risk-averse and receive positive marginal utilities from consumption. She faces a discount factor of $\beta \in (0,1)$, and receives wage income of y_t each period.

At each time t, the representative investor chooses how much to consume and how much to save by choosing how many units of each asset to hold. Thus, the representative

investor solves the following maximization problem:

$$\begin{aligned} \max_{\{c_t\}_{t\in N_+},\{a_t^i\}_{t\in N_+}} \quad & \mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t\cdot u(c_t)\right] \\ \text{subject to} \quad & c_t+\sum_{i=1}^{n}p_t^i\cdot a_t^i = y_t+\sum_{i=1}^{n}(p_t^i+d_t^i)\cdot a_{t-1}^i. \end{aligned}$$

The Lagrangian for this problem is

$$\mathcal{L}=\mathbb{E}_0\left[\sum_{t=0}^{\infty}\beta^t\left[u(c_t)+\lambda_t\left(y_t+\sum_{i=1}^{n}(p_t^i+d_t^i)\cdot a_{t-1}^i-c_t-\sum_{i=1}^{n}p_t^i\cdot a_t^i\right)\right]\right],$$

and the first order conditions for maximization tell us that

$$\begin{aligned} &\frac{\partial\mathcal{L}}{\partial c_t}=u'(c_t)-\lambda_t=0 \\ &\frac{\partial\mathcal{L}}{\partial a_t^i}=-\lambda_t\cdot p_t^i+\beta\cdot\mathbb{E}_t\left[\lambda_{t+1}\left(p_{t+1}^i+d_{t+1}^i\right)\right]=0 \quad \text{for any } 1\le i\le n. \end{aligned}$$

These conditions then yield the Euler equation

$$u'(c_t)=\beta\mathbb{E}_t\left[u'(c_{t+1})\cdot\frac{p_{t+1}^i+d_{t+1}^i}{p_t^i}\right].$$

Note that $p_{t+1}^i+d_{t+1}^i$ is equal to the payoff X_{t+1}^i of asset i at time $t+1$; therefore,

$$\frac{p_{t+1}^i+d_{t+1}^i}{p_t^i}=\frac{X_{t+1}^i}{p_t^i}=1+r_{t+1}^i,$$

and the Euler equation can be written as

$$u'(c_t)=\beta\mathbb{E}_t\left[u'(c_{t+1})\cdot(1+r_{t+1}^i)\right]$$

for any $1\le i\le n$. This has the usual economic interpretation of equating marginal benefit from an additional unit of consumption at time t with the expected marginal cost of that additional unit of consumption.

In the asset pricing literature, the Euler equation is often interpreted as a pricing equation. Rearranging the Euler equation so that p_t^i is on the left hand side yields

$$p_t^i=\mathbb{E}_t\left[\beta\frac{u'(c_{t+1})}{u'(c_t)}\cdot X_{t+1}^i\right].$$

In other words, the time t price of asset i is given as the expectation of its payoff X_{t+1}^i,

discounted by the **stochsatic discount factor (SDF)**

$$\mathcal{M}_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)}.$$

Heuristically, we can interpret the SDF as follows: if the investor expects to consume more tomorrow than today, then $\mathcal{M}_{t+1}$ become smaller, which means that asset i's payoff is discounted much more and ends up with a lower price today. In other words, if the investor expects to consume more tomorrow, then her assets, which serve as a sort of insurance, become less valuable and thus command a lower price today. This suggest that consumption serves the same role here as the market portfolio does in the CAPM, an idea that we expand on further below.

In any case, the pricing formula is given as

$$p_t^i = \mathbb{E}_t \left[\mathcal{M}_{t+1} \cdot X_{t+1}^i \right],$$

which suggests that, if we define the function π_t as

$$\pi_t(X) = \mathbb{E}_t \left[\mathcal{M}_{t+1} \cdot X \right],$$

then the price of every asset is given as the π_t-value of its one-period ahead payoff:

$$p_t^i = \pi_t(X_{t+1}^i).$$

We show in a later section that this sort of representation also follows from the no-arbitrage assumption alone.

1.3.1 The Beta Representation

Here we derive a convenient and useful representation of the expected excess return of an asset using the pricing formulat derived above. Dividing both sides of the pricing formula by p_t^i yields the equation

$$\begin{aligned} 1 &= \mathbb{E}_t \left[\mathcal{M}_{t+1} \cdot (1 + r_{t+1}^i) \right] \\ &= \mathrm{Cov}_t \left(\mathcal{M}_{t+1}, r_{t+1}^i \right) + \mathbb{E}_t \left[\mathcal{M}_{t+1} \right] \cdot \left(1 + \mathbb{E}_t \left[r_{t+1}^i \right] \right). \end{aligned}$$

Rearranging this equation yields

$$\mathbb{E}_t \left[r_{t+1}^i \right] = \frac{1}{\mathbb{E}_t \left[\mathcal{M}_{t+1} \right]} - 1 - \frac{1}{\mathbb{E}_t \left[\mathcal{M}_{t+1} \right]} \mathrm{Cov}_t \left(\mathcal{M}_{t+1}, r_{t+1}^i \right).$$

Since this holds for any asset, even the risk-free asset, we have

$$r_{t+1}^f = \frac{1}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]} - 1,$$

where the covariance term disappears because r_{t+1}^f is known at time t. Therefore,

$$\begin{aligned}\mathbb{E}_t\left[r_{t+1}^i\right] - r_{t+1}^f &= -\frac{1}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]}\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r_{t+1}^i\right) = -(1+r_{t+1}^f)\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r_{t+1}^i\right) \\ &= \frac{\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r_{t+1}^i\right)}{\mathrm{Var}_t\left(\mathcal{M}_{t+1}\right)} \cdot \left[-(1+r_{t+1}^f)\mathrm{Var}_t\left(\mathcal{M}_{t+1}\right)\right]\end{aligned}$$

for any $1 \leq i \leq n$. Note the similarities with the CAPM: the risk premium of each asset depends on a common factor

$$MPR_t = -(1+r_{t+1}^f)\mathrm{Var}_t\left(\mathcal{M}_{t+1}\right),$$

which loads on the risk premium of asset i with loading

$$\beta_{i,t} = \frac{\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r_{t+1}^i\right)}{\mathrm{Var}_t\left(\mathcal{M}_{t+1}\right)}$$

unique to the asset.

The beta term $\beta_{i,t}$ represents the unique risk associated with the ith asset. The more highly correlated the asset is with the SDF $\mathcal{M}_{t+1}$, the more risky it is (= the less it functions as insurance against macro risks) and therefore asset i commands a higher risk premium.

The term MPR_t is called the **market price of risk**, since it determines how much the risk premium increases in response to a unit increase in the riskiness, or in other words, the beta, of an asset. The higher MPR_t, the more the risk premium of an asset increases in response to a unit increase in its beta; that is, a high MPR_t indicates that an additional unit of risk commands a high price in terms of the risk premium, or compensation for risk. MPR_t is also common to all assets, so it can be interpreted as the common "market price" of risk.

1.3.2 The Sharpe Ratio

The **Sharpe Ratio (SR)** of an asset is an indicator of the profitability of an asset relative to its risk. The SR of asset i from time t to $t+1$ is defined as

$$SR_{t+1}^i = \frac{\mathbb{E}_t\left[r_{t+1}^i\right] - r_{t+1}^f}{\sigma_t(r_{t+1}^i)},$$

where $\sigma_t(r^i_{t+1})$ is the standard deviation of r^i_{t+1}. The higher the Sharpe Ratio, the greater the profitability of the asset compared to other assets with the same amount of risk, quantified in terms of the standard deviation of the rate of return.

We can express the SR of asset i in terms of the correlation between the asset's rate of return and the SDF. To see this, recall that the pricing formula implies

$$\begin{aligned}\mathbb{E}_t\left[r^i_{t+1}\right] - r^f_{t+1} &= -\frac{1}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]}\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r^i_{t+1}\right) \\ &= -\frac{1}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]}\mathrm{Corr}_t\left(\mathcal{M}_{t+1}, r^i_{t+1}\right)\cdot\sigma_t(\mathcal{M}_{t+1})\cdot\sigma_t(r^i_{t+1}),\end{aligned}$$

where the last equality follows from the definition of the correlation coefficient. Therefore,

$$SR^i_{t+1} = \frac{\mathbb{E}_t\left[r^i_{t+1}\right] - r^f_{t+1}}{\sigma_t(r^i_{t+1})} = -\mathrm{Corr}_t\left(\mathcal{M}_{t+1}, r^i_{t+1}\right)\cdot\frac{\sigma_t(\mathcal{M}_{t+1})}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]}.$$

This indicates that the highest and lowest possible SRs are

$$\frac{\sigma_t(\mathcal{M}_{t+1})}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]} \quad \text{and} \quad -\frac{\sigma_t(\mathcal{M}_{t+1})}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]},$$

which are the ratios of assets whose rate of return is perfectly correlated with the SDF.

Note also that the definition of the Sharpe Ratio tells us that the expected rate of return of an asset and its standard deviation satisfies the following trade-off:

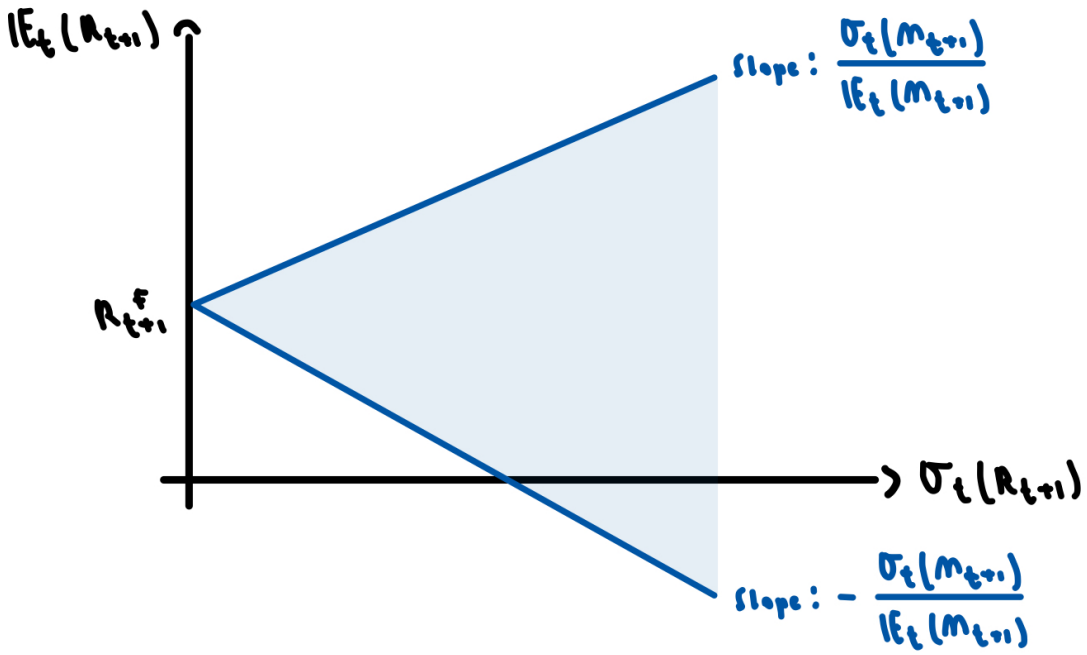
$$\mathbb{E}_t\left[r^i_{t+1}\right] = r^f_{t+1} + SR^i_{t+1}\cdot\sigma_t(r^i_{t+1})$$

The slope of this line, which is called the **capital allocation line (CAL)** in classical portfolio theory, is exactly the Sharpe Ratio. In light of the maximum and minimum ratios derived above, this implies that, for any asset i, the pair $(\mathbb{E}_t\left[r^i_{t+1}\right], \sigma_t(r^i_{t+1}))$ lies in the area surrounded by the mean-variance frontier, pictured below:

1.3.3 The Case of Log-Normal Returns

Suppose now that the log of the SDF $\mathcal{M}_{t+1}$ and return of asset i $1 + r^i_{t+1}$ jointly follow a normal distribution conditional on information up to time t:

$$\begin{pmatrix}\log(\mathcal{M}_{t+1}) \\ \log\left(1 + r^i_{t+1}\right)\end{pmatrix} := \begin{pmatrix} m_{t+1} \\ r^i_{t+1}\end{pmatrix} \sim \mathcal{N},$$



where we used the approximation $\log\left(1+r^i_{t+1}\right) \approx r^i_{t+1}$. Then, the pricing formula can be written as

$$\begin{aligned} 1 &= \mathbb{E}_t\left[\mathcal{M}_{t+1}\cdot(1+r^i_{t+1})\right] \\ &= \mathbb{E}_t\left[\exp\left(m_{t+1}+r^i_{t+1}\right)\right] \\ &= \exp\left(\mathbb{E}_t\left[m_{t+1}+r^i_{t+1}\right]+\frac{1}{2}\mathrm{Var}_t\left(m_{t+1}+r^i_{t+1}\right)\right), \end{aligned}$$

where the last equality used the formula for the MGF of normally distributed variables. Since

$$\mathrm{Var}_t\left(m_{t+1}+r^i_{t+1}\right) = \mathrm{Var}_t\left(m_{t+1}\right)+\mathrm{Var}_t\left(r^i_{t+1}\right)+2\cdot\mathrm{Cov}_t\left(m_{t+1},r^i_{t+1}\right),$$

taking logs on both sides of the equation above yields

$$0 = \mathbb{E}_t\left[m_{t+1}\right]+\mathbb{E}_t\left[r^i_{t+1}\right]+\frac{1}{2}\left(\mathrm{Var}_t\left(m_{t+1}\right)+\mathrm{Var}_t\left(r^i_{t+1}\right)\right)+\mathrm{Cov}_t\left(m_{t+1},r^i_{t+1}\right).$$

Since

$$\begin{aligned} \exp\left(-r^f_{t+1}\right) \approx \frac{1}{1+r^f_{t+1}} &= \mathbb{E}_t\left[\mathcal{M}_{t+1}\right] = \mathbb{E}_t\left[\exp(m_{t+1})\right] \\ &= \exp\left(\mathbb{E}_t\left[m_{t+1}\right]+\frac{1}{2}\mathrm{Var}_t\left(m_{t+1}\right)\right), \end{aligned}$$

taking logs on both sides yields

$$-r_{t+1}^f = \mathbb{E}_t\left[m_{t+1}\right] + \frac{1}{2}\mathrm{Var}_t\left(m_{t+1}\right),$$

which implies

$$\mathbb{E}_t\left[r_{t+1}^i\right] - r_{t+1}^f = -\frac{1}{2}\mathrm{Var}_t\left(r_{t+1}^i\right) - \mathrm{Cov}_t\left(m_{t+1}, r_{t+1}^i\right).$$

This tells us that, if the SDF and asset return are jointly log-normally distributed, then the expected excess return includes an additional variance term alongside the familiar covariance term. This term is called the **Jensen's Inequality term**, and it is often ignored when talking of the expected excess returns of an asset.

In this case, the Sharpe ratio takes into consideration the Jensen's inequality term, and is defined as

$$\begin{aligned} SR_{t+1}^i = \frac{\mathbb{E}_t\left[r_{t+1}^i\right] - r_{t+1}^f + \frac{1}{2}\mathrm{Var}_t\left(r_{t+1}^i\right)}{\sigma_t(r_{t+1}^i)} &= -\mathrm{Cov}_t\left(m_{t+1}, r_{t+1}^i\right)\frac{1}{\sigma_t(r_{t+1}^i)} \\ &= -\mathrm{Corr}_t\left(m_{t+1}, r_{t+1}^i\right)\sigma_t(m_{t+1}). \end{aligned}$$

The maximum Sharpe ratio in this case is equal to the standard deviation $\sigma_t(m_{t+1})$ of the log SDF.

1.3.4 The Case of CRRA Utility

An important special case, which, among other things, will serve as the benchmark for our derivation of the empirical SDF, is the case of CRRA utility. Suppose the utility function is given in the CRRA form

$$u(c) = \frac{c^{1-\theta} - 1}{1-\theta},$$

where $\theta \geq 0$ is the coefficient of relative risk aversion (the case where $\theta = 1$ corresponds to log utility). Then, the time $t+1$ SDF is

$$\mathcal{M}_{t+1} = \beta\frac{u'(c_{t+1})}{u'(c_t)} = \beta \cdot \left(\frac{c_{t+1}}{c_t}\right)^{-\theta}.$$

Defining consumption growth $g_{t+1} = \frac{c_{t+1}-c_t}{c_t}$, we can see that

$$\mathcal{M}_{t+1} = \beta \cdot (1 + g_{t+1})^{-\theta}.$$

An important approximation result in macroeconomics tells us that

$$(1+x)^a \approx 1+ax$$

when x is small, which follows from a first order Taylor expansion applied to the mapping $x \mapsto (1+x)^a$ around 0. Assuming that consumption growth is near 0 (which in most cases it is), we now obtain the approximation

$$\mathcal{M}_{t+1} = \beta \cdot (1+g_{t+1})^{-\theta} \approx \beta(1-\theta g_{t+1}),$$

so that the expected excess return of asset i is

$$\begin{aligned}
\mathbb{E}_t\left[r^i_{t+1}\right] - r^f_{t+1} &= -\frac{1}{\mathbb{E}_t\left[\mathcal{M}_{t+1}\right]}\mathrm{Cov}_t\left(\mathcal{M}_{t+1}, r^i_{t+1}\right) \\
&= -(1+r^f_{t+1}) \cdot \mathrm{Cov}_t\left(\beta(1-\theta g_{t+1}), r^i_{t+1}\right) \\
&= \beta(1+r^f_{t+1}) \cdot \theta \mathrm{Cov}_t\left(g_{t+1}, r^i_{t+1}\right).
\end{aligned}$$

In other words, the risk premium of asset i is proportional to the level of risk aversion of the investors θ and the covariance of consumption growth and the return to asset i, $\mathrm{Cov}_t\left(g_{t+1}, r^i_{t+1}\right)$. As usual, the covariance term (=beta temr) represents the systematic risk present in the asset itself, and differs from asset to asset. Meanwhile, the risk aversion coefficient θ is contained in the market price of risk term, and thus represents how sensitive investors are (=the amount of compensation investors demand) to a unit increase in risk; note how it is not dependent on a specific asset.

We usually call g_{t+1} the risk factor, in other words, the factor that represents the systematic risk present in the economy. The expected excess return of an asset can then be said to be determined by an asset-specific part, namely the correlation of the rate of return with the risk factor, and an non asset-specific part, namely the market price of risk, or investors' attitude to risk.

1.3.5 The Equity Premium Puzzle

Under CRRA utility, the pricing formula suggests that asset risk premia should be determined according to the equation

$$\mathbb{E}_t\left[r^i_{t+1}\right] - r^f_{t+1} = \beta(1+r^f_{t+1}) \cdot \theta \mathrm{Cov}_t\left(g_{t+1}, r^i_{t+1}\right).$$

However, the actual data for the expected excess return on stocks (on average 0.06) indicate that the relative risk aversion coefficient θ should be around 30, a value that is too high to be plausible. This discrepancy between theory and practice is referred to as

the **equity premium puzzle**; here, equity premium refers to the excess return on stocks compared to bonds (=risk-free asset).

Many attempts have been made to resolve the equity premium puzzle; here we present a few popular approaches:

1) **Habit Formation**
This theory, introduced in Campbell and Cochrane (1999), posits that utility is derived not only from consumption but also from adherence to previously formed "habits". This has the effect of raising the degree of risk aversion during recessions and lowering it during booms, so that θ is only required to be high during recessions, a reasonable conclusion.

2) **Distorted Beliefs**
This theory suggests that the conditional expectation $\mathbb{E}_t[\cdot]$ does not accurately reflect investors' attitudes to risk. For instance, it does not reflect how investors' patterns of risk aversion change during recessions and booms.

3) **Survivorship Bias**
Finally, one strand of the literature emphasizes that the left hand side expression $\mathbb{E}_t\left[r^i_{t+1}\right] - r^f_{t+1}$ will be computed using only assets in the United States. However, investors, who consider assets of other less sucessful countries, will actually be facing a risk premium that is much lower than 0.06.

1.4 Models of Stock Prices

Here we briefly study models of stock prices. We focus on two results: the dividend discount model and the dynamic Gordon formula. The first is a direct consequence of the C-CAPM, while the latter is an alternative approach that only makes use of the definition of stock returns.

1.4.1 The Dividend Discount Model

Suppose $\{X_t\}_{t\in N_+}$ is the payoff stream of a share of stock, so that

$$X_{t+1} = p_{t+1} + d_{t+1},$$

where p_{t+1} is the time $t+1$ price of the stock and d_{t+1} its time $t+1$ dividend. In this case, the pricing formula becomes

$$p_t = \mathbb{E}_t\left[\mathcal{M}_{t+1}X_{t+1}\right] = \mathbb{E}_t\left[\mathcal{M}_{t+1}(p_{t+1}+d_{t+1})\right]$$

for any $t \in \mathbb{N}$. The above equation represents a recursion, so that, for any $T > t$, we have

$$\begin{aligned}
p_t &= \mathbb{E}_t\left[\mathcal{M}_{t+1}(p_{t+1}+d_{t+1})\right] \\
&= \mathbb{E}_t\left[\mathcal{M}_{t+1}\mathbb{E}_{t+1}\left[\mathcal{M}_{t+2}(p_{t+2}+d_{t+2})\right]\right] + \mathbb{E}_t\left[\mathcal{M}_{t+1}d_{t+1}\right] \\
&= \mathbb{E}_t\left[\mathcal{M}_{t+1}\mathcal{M}_{t+2}\cdot p_{t+2}\right] + \mathbb{E}_t\left[\mathcal{M}_{t+1}d_{t+1}\right] + \mathbb{E}_t\left[\mathcal{M}_{t+1}\mathcal{M}_{t+2}\cdot d_{t+2}\right] \\
&= \cdots = \mathbb{E}_t\left[\left(\prod_{s=1}^{T}\mathcal{M}_{t+s}\right)p_{t+T}\right] + \sum_{s=1}^{T}\mathbb{E}_t\left[\left(\prod_{r=1}^{s}\mathcal{M}_{t+r}\right)\cdot d_{t+s}\right].
\end{aligned}$$

Assuming that

$$\lim_{T\to\infty}\mathbb{E}_t\left[\left(\prod_{s=1}^{T}\mathcal{M}_{t+s}\right)p_{t+T}\right] = 0,$$

which requires discounted stock prices to converge to 0 in the far future and thus represents a no-bubble condition, we can express p_t as

$$p_t = \sum_{s=1}^{\infty}\mathbb{E}_t\left[\left(\prod_{r=1}^{s}\mathcal{M}_{t+r}\right)\cdot d_{t+s}\right].$$

Here, the s-period ahead dividend is discounted by $\prod_{r=1}^{s}\mathcal{M}_{t+r}$, so that

$$\Lambda_{t,t+s} = \prod_{r=1}^{s}\mathcal{M}_{t+r}$$

is the SDF from time t to $t+s$. Using this more general expression for the discount factor, stock prices are given as

$$p_t = \sum_{s=1}^{\infty} \mathbb{E}_t \left[\Lambda_{t,t+s} \cdot d_{t+s} \right].$$

This tells us that the current stock price is the sum of expected discounted future dividends; this equation forms the centerpiece of the **dividend discount model (DDM)** of stock prices.

1.4.2 The Dynamic Gordon Formula

This is another approach to the modeling of stock prices that relies not on the C-CAPM but only the definition of stock returns. Let $R_{t+1} = 1 + r_{t+1}$ denote the returns at time $t+1$, P_t the time t price, and D_t the time t dividend. We denote by p_t and d_t the logs of P_t and D_t. The ratio $\frac{D_t}{P_t}$ is the **dividend-price ratio (DPR)**, and its log $d_t - p_t$ is referred to as the log DPR.

By definition,

$$R_{t+1} = 1 + r_{t+1} = \frac{P_{t+1} + D_{t+1}}{P_t} = \frac{P_{t+1}}{P_t} \left(1 + \frac{D_{t+1}}{P_{t+1}} \right).$$

Using the fact that

$$\frac{P_{t+1}}{P_t} = \exp(\log(P_{t+1}) - \log(P_t)) = \exp(p_{t+1} - p_t)$$

and

$$1 + \frac{D_{t+1}}{P_{t+1}} = \exp\left(\log\left(1 + \frac{D_{t+1}}{P_{t+1}} \right) \right) = \exp(\log\left(1 + \exp(d_{t+1} - p_{t+1})\right)),$$

the rate of return can be approximated as

$$\begin{aligned} r_{t+1} &\approx \log(1 + r_{t+1}) = \log(R_{t+1}) \\ &= p_{t+1} - p_t + \log\left(1 + \exp(d_{t+1} - p_{t+1})\right). \end{aligned}$$

Letting $\overline{d-p}$ be the mean log DPR, a Taylor expansion of $\log\left(1 + \exp(d_{t+1} - p_{t+1})\right)$ with respect to $d_{t+1} - p_{t+1}$ around $\overline{d-p}$ yields

$$\log\left(1 + \exp(d_{t+1} - p_{t+1})\right) \approx \log\left(1 + \exp\left(\overline{d-p}\right)\right) + \frac{\exp\left(\overline{d-p}\right)}{1 + \exp\left(\overline{d-p}\right)} \left(d_{t+1} - p_{t+1} - \overline{d-p} \right).$$

Defining $\rho = \frac{1}{1+\exp(\overline{d-p})} \in (0,1)$, we have

$$\log\left(1+\exp(d_{t+1}-p_{t+1})\right) \approx -\log(\rho) + (1-\rho)\left(d_{t+1}-p_{t+1}-\log\left(\rho^{-1}-1\right)\right),$$

and the rate of return is approximated by

$$r_{t+1} \approx p_{t+1} - p_t + (1-\rho)\left(d_{t+1}-p_{t+1}\right) - \log(\rho) - (1-\rho)\log\left(\rho^{-1}-1\right).$$

Defining

$$k = -\log(\rho) - (1-\rho)\log\left(\rho^{-1}-1\right),$$

stock prices are now given by

$$\begin{aligned} p_t &= k + p_{t+1} - r_{t+1} + (1-\rho)\left(d_{t+1}-p_{t+1}\right) \\ &= k + (1-\rho)d_{t+1} - r_{t+1} + \rho \cdot p_{t+1}. \end{aligned}$$

Recursively substituting the above formula leads to

$$\begin{aligned} p_t &= k + (1-\rho)d_{t+1} - r_{t+1} + \rho\left(k + (1-\rho)d_{t+2} - r_{t+2} + \rho \cdot p_{t+2}\right) \\ &= \rho^2 \cdot p_{t+2} + k\left(1+\rho\right) + (1-\rho)\left(d_{t+1} + \rho \cdot d_{t+2}\right) - \left(r_{t+1} + \rho \cdot r_{t+2}\right) \\ &= \cdots = \rho^{T+1} \cdot p_{t+T+1} + k\left(\sum_{s=1}^{T} \rho^{s-1}\right) + (1-\rho)\sum_{s=1}^{T} \rho^{s-1} d_{t+s} - \sum_{s=1}^{T} \rho^{s-1} r_{t+s} \end{aligned}$$

for any $T > t$. If we assume that the no-bubble condition

$$\lim_{T\to\infty} \rho^T p_{t+T} = 0$$

holds, then stock prices are given in the limit as

$$p_t = \frac{k}{1-\rho} + (1-\rho) \cdot \sum_{s=1}^{\infty} \rho^{s-1} d_{t+s} - \sum_{s=1}^{\infty} \rho^{s-1} r_{t+s}.$$

This is a similar conclusion to the dividend discount model, where current stock prices are given as the discounted sum of future dividends, but with an additional term invovling future rates of return. It is worth noting that this was derived (albeit approximately) from the definition of the rate of return alone.

Using the above formula for stock prices, we can now obtain an expression for the log

DPR:

$$\begin{aligned} d_t - p_t &= d_t - \frac{k}{1-\rho} - \sum_{s=1}^{\infty} \rho^{s-1} d_{t+s} + \sum_{s=1}^{\infty} \rho^{s} d_{t+s} + \sum_{s=1}^{\infty} \rho^{s-1} r_{t+s} \\ &= \left(d_t - d_{t+1} + \rho(d_{t+1} - d_{t+2}) + \rho^2 (d_{t+2} - d_{t+3}) + \cdots \right) - \frac{k}{1-\rho} + \sum_{s=1}^{\infty} \rho^{s-1} r_{t+s} \\ &= -\frac{k}{1-\rho} + \sum_{s=1}^{\infty} \rho^{s-1} \left(r_{t+s} - \Delta d_{t+s} \right) \end{aligned}$$

where we define $\Delta d_{t+s} = d_{t+s} - d_{t+s-1}$, which is the dividend growth rate from time $t+s-1$ to $t+s$. Therefore, the log DPR is determined as the sum of the discounted differences between future rates of return and dividend growth rates. This formula, derived from the definition of the rate of return and thus possessing great generality, is called the **dynamic Gordon formula**.

Given the close relationship between the log DPR and future rates of return, the DPR has been used, ever since Campbell and Shiller (1988), as a predictor for future stock returns. Another variable that is often used to test the predictability of stock returns is the **consumption-wealth ratio (CAY)**, which is calculated as

$$cay_t = C_t - \beta_1 \cdot a_t - \beta_2 \cdot Y_t,$$

where C_t is consumption, a_t is wealth and Y_t is income.

Chapter 2

Arbitrage-based Models of Asset Pricing

In contrast to the previous chapter, which dealt with equilibrium approaches to asset pricing, the arbitrage-based strand of the asset pricing literature strives to obtain asset pricing results from the weakest possible assumptions. Here, we study this approach to asset pricing, which covers sufficient conditions for the existence of an SDF, the risk-neutral measure, and the form of the empirical SDF.

2.1 Hilbert Spaces and L^p Spaces

Here we briefly introduce and prove results pertaining to Hilbert spaces and L^p spaces. Hilbert spaces will be very useful when studying arbitrage pricing theory, and indeed, has very wide applicability even outside of asset pricing (a notable example is in time series analysis).

2.1.1 Definition of a Hilbert Space

Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space over the complex field. Recall the following definition of an inner product $\langle \cdot, \cdot \rangle : V \to \mathbb{C}$:

a) **Linearity in the First Argument**
For any $u, v, m \in V$ and $z \in \mathbb{C}$,

$$\langle z \cdot u + v, m \rangle = z \cdot \langle u, m \rangle + \langle v, m \rangle.$$

b) **Conjugate Symmetry**

For any $u, v \in V$,

$$\langle u, v \rangle = \overline{\langle v, u \rangle}.$$

c) **Positive Definiteness**
For any $v \in V$,

$$\langle v, v \rangle > 0 \quad \text{if and only if} \quad v \neq 0_V,$$

where 0_V is the zero vector of V.

The following properties follow by definition:

$$\begin{aligned}
\langle \mathbf{0}_V, v \rangle = \langle v, \mathbf{0}_V \rangle = 0 \quad & \text{for any } v \in V \\
\langle v, v \rangle \geq 0 \quad & \text{for any } v \in V \\
\langle v, v \rangle = 0 \quad & \text{if and only if } v = 0_V \\
\langle m, z \cdot u + v \rangle = \overline{z} \cdot \langle m, u \rangle + \langle m, v \rangle \quad & \text{for any } m, u, v \in V \text{ and } z \in \mathbb{C} \\
\langle u + v, u + v \rangle = \langle u, u \rangle + \langle v, v \rangle + 2 \cdot Re\left(\langle u, v \rangle\right) \quad & \text{for any } u, v \in V.
\end{aligned}$$

The Cauchy-Schwarz inequality is trickier to prove (for a proof, consult any linear algebra textbook):

$$|\langle u, v \rangle| \leq \sqrt{\langle u, u \rangle} \cdot \sqrt{\langle v, v \rangle} \quad \text{for any } u, v \in V.$$

Let $\|\cdot\|$ be the norm induced by the inner product, that is, the function $\|\cdot\| : V \to \mathbb{R}_+$ defined as

$$\|v\| = \sqrt{\langle v, v \rangle}$$

for any $v \in V$. We can easily verify that $\|\cdot\|$ satisfies the conditions of a norm using the properties above and the Cauchy-Schwarz inequality:

a) For any $v \in V$, $\|v\| = 0$ if and only if $v = 0_V$.

b) For any $v \in V$ and $z \in \mathbb{C}$, $\|z \cdot v\| = |z| \cdot \|v\|$.

c) For any $u, v \in V$, $\|u + v\| \leq \|u\| + \|v\|$.

Let $d : V \times V \to \mathbb{R}_+$ be the metric induced by the norm $\|\cdot\|$, that is, the function defined as

$$d(u, v) = \|u - v\|$$

for any $u, v \in V$. It is also not difficult to show that d satisfies the conditions of a metric:

a) For any $u, v \in V$, $d(u,v) = 0$ if and only if $u = v$.

b) For any $u, v \in V$, $d(u,v) = d(v,u)$.

c) For any $u, v, m \in V$, $d(u,v) \leq d(u,m) + d(m,v)$.

Therefore, the pair (V,d) is a metric space. We call the inner product space $(V, \langle \cdot, \cdot \rangle)$ a Hilbert space if the metric space (V,d) is a complete metric space: for any sequence $\{x_n\}_{n \in N_+}$ in V that is Cauchy, that is,

$$\lim_{m,n \to \infty} d(x_n, x_m) = 0,$$

$\{x_n\}_{n \in N_+}$ is convergent, that is, there exists some $x \in V$ such that

$$\lim_{n \to \infty} d(x_n, x) = 0.$$

2.1.2 Orthogonal Projections

Given some subset V of an inner product space $(H, \langle \cdot, \cdot \rangle)$, we call $y \in V$ an orthogonal projection of $x \in H$ on V if

$$\|x - y\| = \inf_{z \in V} \|x - z\|.$$

The following are some general results on orthogonal projections:

Theorem (Properties of Orthogonal Projections)
Let $(H, \langle \cdot, \cdot \rangle)$ be an inner product space over the complex field and $\|\cdot\|$ and d the norm and metric induced by $\langle \cdot, \cdot \rangle$. Let V be a subset of H. The following hold true:

i) Let $x \in H$, and suppose that $y \in V$ is an orthogonal projection of x on V, that is,

$$\|x - y\| = \inf_{z \in V} \|x - z\|.$$

Then, y is the unique orthogonal projection of x on y if V is a convex set.

ii) Let $x \in H$, and suppose that V is a subspace of H. Then, $y \in V$ is the unique orthogonal projection of x on V if and only if $\langle x - y, z \rangle = 0$ for any $z \in V$.

iii) Let V be a subspace of H.
Suppose that, for any $x \in H$, there exists a unique orthogonal projection of x on V. Then, $H = V \bigoplus V^{\perp}$.
Moreover, denoting the mapping from x to its unique orthogonal projection on V by P, and the mapping from x to $x - Px$ by Q, P, Q are linear transformations from H into V and $V^{\perp}$, and Qx is the orthogonal projection of x on $V^{\perp}$.
For any $x \in H$, we have

$$\|x\|^2 = \|Px\|^2 + \|Qx\|^2.$$

Proof) i) Let $x \in H$, and suppose that $y \in V$ is an orthogonal projection of x on V, that is,

$$\|x - y\| = \inf_{z \in V} \|x - z\|.$$

Suppose that V is a convex set, and let $y' \in V$ be another orthogonal projection of x on V. Denoting $\delta = \|x - y\| = \|x - y'\|$, by the parallelogram law,

$$\left\| \frac{1}{2}\left(y - y'\right) \right\|^2 + \left\| x - \frac{y + y'}{2} \right\|^2 = 2 \cdot \left\| \frac{1}{2}(x - y) \right\|^2 + 2 \cdot \left\| \frac{1}{2}(x - y') \right\|^2.$$

Multiplying both sides by 4 and noting that $\frac{y+y'}{2} \in V$ because V is convex, we can see that

$$\begin{aligned} \left\| y - y' \right\|^2 &= 2 \cdot \left(\|x - y\|^2 + \left\| x - y' \right\|^2 - 2 \left\| x - \frac{y + y'}{2} \right\|^2 \right) \\ &\leq 2 \cdot \left(2\delta^2 - 2\delta^2 \right) = 0, \end{aligned}$$

since $\left\| x - \frac{y+y'}{2} \right\|^2 \geq \delta^2$. Therefore, $\|y - y'\| = 0$ and $y = y'$, making y the unique orthogonal projection of x on V.

ii) Let V be a subspace of H, and for any $x \in H$, suppose that $y \in V$ is the orthogonal projection of x on V (it is unique because V is convex). Then, by definition,

$$\|x - y\| \leq \|x - z\|$$

for any $z \in V$. Choose any $z \in V$; if $z = 0_H$, then $\langle x - y, z \rangle = 0$ trivially.

Supose that $z \neq 0_H$. For any $a \in \mathbb{C}$, $y + az \in V$ because V is a subspace of H, and thus

$$\|x-y\| \leq \|x-(y+az)\| = \|(x-y)-az\|$$

by the definition of y as the orthogonal projection of x on V. Then, we have

$$\begin{aligned}\|x-y\|^2 \leq \|(x-y)-az\|^2 &= \langle (x-y)-az, (x-y)-az\rangle \\ &= \|x-y\|^2 + |a|^2\|z\|^2 - a\cdot\langle z, x-y\rangle - \bar{a}\cdot\langle x-y, z\rangle,\end{aligned}$$

so that

$$0 \leq |a|^2\|z\|^2 - a\cdot\langle z, x-y\rangle - \bar{a}\cdot\langle x-y, z\rangle.$$

Putting $a = \frac{\langle x-y,z\rangle}{\|z\|^2} \in \mathbb{C}$, the above inequality becomes

$$0 \leq \frac{|\langle x-y,z\rangle|^2}{\|z\|^2} - 2\cdot\frac{|\langle x-y,z\rangle|^2}{\|z\|^2} = -\frac{|\langle x-y,z\rangle|^2}{\|z\|^2},$$

and multiplying both sides by $-\|z\|^2$, we obtain

$$|\langle x-y, z\rangle|^2 \leq 0.$$

This implies that $|\langle x-y,z\rangle|^2 = 0$, or that $\langle x-y, z\rangle = 0$.

Now suppose that $y \in V$ satisfies $\langle x-y, z\rangle$ for any $z \in V$. Then, for any $z \in V$,

$$\begin{aligned}\|x-z\|^2 &= \langle (x-y)+(y-z), (x-y)+(y-z)\rangle \\ &= \|x-y\|^2 + \|y-z\|^2 + 2\cdot Re\left(\langle x-y, y-z\rangle\right).\end{aligned}$$

Since $y-z \in V$ (V is a subspace), by assumption we have $\langle x-y, y-z\rangle = 0$, so that

$$\|x-z\|^2 = \|x-y\|^2 + \|y-z\|^2 \geq \|x-y\|^2.$$

This holds for any $z \in V$, so y is an orthogonal projection of x on V, and by the convexity of V, it is the unique orthogonal projection of x on V.

iii) Let V be a subspace of H, and suppose that, for any $x \in H$, there exists

a unique orthogonal projection of x on V. Define the mapping $P : H \to V$ so that Px is the unique orthogonal projection of x on V for any $x \in H$. For any $x \in H$, by the second result, we can see that $\langle x - Px, z\rangle = 0$ for any $z \in V$. This means that $x - Px \in V^{\perp}$, so defining the mapping $Q : H \to V^{\perp}$ as $Qx = x - Px$ for any $x \in H$,

$$x = Px + Qx,$$

where $Px \in V$ and $Qx \in V^{\perp}$, for any $x \in H$. This shows us that $H = V \bigoplus V^{\perp}$, where the sum becomes a direct sum because V and $V^{\perp}$ are independent.

To see that P and Q are linear, choose any $x, y \in H$, $a \in \mathbb{C}$, and note that

$$\begin{aligned} a \cdot (Px + Qx) + (Py + Qy) &= a \cdot x + y \\ &= P(ax + y) + Q(ax + y) \end{aligned}$$

by the decomposition above. Rearranging terms yields

$$P(ax + y) - a \cdot Px - Py = a \cdot Qx + Qy - Q(ax + y);$$

the left hand side is in V and the right hand side in $V^{\perp}$, and because $V \cap V^{\perp} = 0_H$ (if $z \in V \cap V^{\perp}$, then $\langle z, z\rangle = \|z\|^2 = 0$, or $z = 0_H$), this tells us that

$$P(ax + y) - a \cdot Px - Py = a \cdot Qx + Qy - Q(ax + y) = 0_H.$$

The linearity of P and Q follows immediately.

For any $x \in H$ and $y \in V^{\perp}$,

$$\langle x - Qx, y\rangle = \langle Px, y\rangle = 0$$

because $Px \in V$; by the previous result, this tells us that $Qx \in V^{\perp}$ is the unique orthogonal projection of x on $V^{\perp}$.

Finally, choose any $x \in H$, and note that

$$\|x\|^2 = \|Px + Qx\|^2 = \|Px\|^2 + \|Qx\|^2 + 2 \cdot Re(\langle Px, Qx\rangle) = \|Px\|^2 + \|Qx\|^2$$

because $Px \in V$ and $Qx \in V^{\perp}$.

Q.E.D.

It is well-known that orthogonal projections from a point to a subspace always exists in finite-dimensional inner product spaces. However, this may not be the case for infinite-dimensional inner product spaces. It is one of the most important properties of Hilbert spaces that any closed convex subspace has a unique orthogonal projection.

2.1.3 The Projection Theorem

In general, there does not always exist an orthogonal projection of a vector $x \in H$ onto an arbitrary subset V of H. However, Hilbert spaces are special in that, for any closed convex subset V of H and some $x \in H$, there always exists an orthogonal projection of x onto V.

This property, called the Hilbert projection theorem, allows us to work with orthgonal projections without worrying about their existence in infinite-dimensional spaces (for example, function spaces like L^p spaces), and as such forms the cornerstone of many important mathematical results, including but not limited to the Radon-Nikodym theorem and the characterization of conditional expectations.

The projection theorem is stated below:

Theorem (The Hilbert Projection Theorem)
Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space over the complex field and $\|\cdot\|$ and d the norm and metric induced by $\langle\cdot,\cdot\rangle$. For any nonempty closed convex subset V of H and $x \in H$, there exists a unique $y \in V$ such that

$$\|x - y\| = \inf_{z \in V} \|x - z\|.$$

Furthermore, if V is a closed subspace of H, then the following hold true:

i) $H = V \bigoplus V^{\perp}$.

ii) Defining Px as the unique orthogonal projection of x on V for any $x \in H$, the mapping $x \mapsto Px$ is a linear transformation from H into V.

iii) Defining $Qx = x - Px$ for any $x \in H$, the mapping $x \mapsto Qx$ is a linear transformation from H into $V^{\perp}$, and Qx is an orthogonal projection of x on $V^{\perp}$ for any $x \in H$.

iv) For any $x \in H$, $x = Px + Qx$ and

$$\|x\|^2 = \|Px\|^2 + \|Qx\|^2.$$

Proof) Choose any $x \in H$. Since the set

$$\{z \in V \mid \|x - z\|\}$$

is nonempty due to the nonemptiness of V and bounded below by 0, the infimum

$$\delta = \inf_{z \in V} \|x - z\|$$

exists in $\mathbb{R}_+$. For any $n \in N_+$, by the definition of the infimum there exists a $y_n \in V$ such that

$$\delta \leq \|x - y_n\| < \delta + \frac{1}{n},$$

or equivalently,

$$|\|x - y_n\| - \delta| < \frac{1}{n},$$

so the sequence $\{\|x - y_n\|\}_{n \in N_+}$ converges to δ.
For any $m, n \in N_+$, by the parallelogram law we can see that

$$\left\|\frac{1}{2}(y_n - y_m)\right\|^2 + \left\|x - \frac{y_n + y_m}{2}\right\|^2 = 2 \cdot \left\|\frac{1}{2}(x - y_n)\right\|^2 + 2 \cdot \left\|\frac{1}{2}(x - y_m)\right\|^2,$$

and because $\frac{y_n + y_m}{2} \in V$ by the convexity of V,

$$\delta^2 = \inf_{z \in V} \|x - z\|^2 \leq \left\|x - \frac{y_n + y_m}{2}\right\|^2$$

and we have

$$\|y_n - y_m\|^2 \leq 2\|x - y_n)\|^2 + 2\|x - y_m\|^2 - 4\delta^2.$$

Taking $n, m \to \infty$ on both sides, since

$$\lim_{n \to \infty} \|x - y_n)\|^2 = \lim_{m \to \infty} \|x - y_m\|^2 = \delta^2,$$

the right hand side converges to 0 and thus

$$\lim_{n,m\to\infty} \|y_n - y_m\|^2 = 0.$$

This shows us that $\{y_n\}_{n\in N_+} \subset V$ is Cauchy in the metric d; by the completeness of the metric space (H, d), there exists a $y^* \in H$ such that $y_n \to y^*$ as $n \to \infty$ in the metric d. Finally, because V is a closed subset of H and $\{y_n\}_{n\in N_+}$ is a sequence in V, $y^* \in V$ as well. The continuity of the mapping $y \mapsto \|x - y\|$ on H now tells us that

$$\|x - y^*\| = \lim_{n\to\infty} \|x - y_n\| = \delta = \inf_{z\in V} \|x - z\|.$$

We have shown so far that y^* is an orthogonal projection of x on V. Because V is convex, by the preceding theorem, y^* is the unique orthogonal projection of x on V.

Suppose that V is a closed subspace of H. Then, because V is a closed convex subset of H, by the result above, for any $x \in H$ there exists a unique orthogonal projection of x on V. By the preceding theorem, we can now see that properties i) to iv) above hold true.

Q.E.D.

Corollary to the Hilbert Projection Theorem Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space over the complex field and $\|\cdot\|$ and d the norm and metric induced by $\langle\cdot,\cdot\rangle$. For any nonempty closed convex subset V of H, there exists a unique $y \in V$ of smallest norm, that is, a unique element $y \in V$ such that $\|y\| \leq \|z\|$ for any $z \in V$.

Proof) This follows immediately from the Hilbert Projection Theorem. Specifically, because V is a closed convex subset of the hilbert space H, there exists a unique $y \in V$ such that

$$\|y\| = \|0_H - y\| = \inf_{z\in V} \|0_H - z\| = \inf_{z\in V} \|z\|.$$

Q.E.D.

2.1.4 The Riesz Representation Theorem

A useful application of the Projection Theorem is the Riesz Representation Theorem, which tells us that any linear functional on a Hilbert space can be represented as the inner product with some element of that space.

We first define a property called continuity at 0. Given a real normed vector space $(V, \|\cdot\|)$, we say that a function $f : V \to \mathbb{R}$ is continuous at 0 if, for any $\{x_n\}_{n\in N_+} \subset V$ such that

$$\lim_{n\to\infty} \|x_n\| = 0,$$

we also have

$$\lim_{n\to\infty} f(x_n) = 0.$$

The statement and proof of the main theorem are given below:

Theorem (The Riesz-Fréchet Representation Theorem)
Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space over the complex field and $\|\cdot\|$ and d the norm and metric induced by $\langle\cdot,\cdot\rangle$. For any linear functional $L \in \mathcal{L}(H, \mathbb{C})$ that is continuous at 0, there exists a unique element $\varphi \in H$ (also called the Riesz representation of L) such that

$$L(x) = \langle x, \varphi\rangle$$

for any $x \in H$. $\varphi \neq 0_H$ if there exists at least one $x \in H$ such that $L(x) \neq 0$.

Proof) We first show uniqueness. Suppose that there exist $\varphi_1, \varphi_2 \in H$ such that

$$L(x) = \langle x, \varphi_i\rangle$$

for any $x \in H$ and $i = 1, 2$. Then,

$$L(\varphi_1 - \varphi_2) = \langle \varphi_1 - \varphi_2, \varphi_1\rangle = \langle \varphi_1 - \varphi_2, \varphi_2\rangle,$$

so that

$$\|\varphi_1 - \varphi_2\|^2 = \langle \varphi_1 - \varphi_2, \varphi_1\rangle - \langle \varphi_1 - \varphi_2, \varphi_2\rangle = 0.$$

This implies that $\varphi_1 = \varphi_2$, and that the Riesz representation of L, if it exists, is unique.

To show existence, first define V as the null space of L, that is, as $V = L^{-1}(\{0\})$. We first show that V is a closed subset of H. Let $x \in H$ be a limit point of V; then, there exists a sequence $\{x_n\}_{n \in N_+}$ in V that converges to x, that is,

$$\lim_{n \to \infty} \|x_n - x\| = 0.$$

The continuity of L at 0 now tells us that

$$\lim_{n \to \infty} L(x_n - x) = 0,$$

and by the linearity of L,

$$\lim_{n \to \infty} L(x_n) = L(x).$$

Each x_n is contained in V, the null space of L, so $L(x_n) = 0$; therefore, $L(x) = 0$ as well, which tells us that $x \in V$. By definition, V is a closed set.

Furthermore, V is a linear subspace, so by the Hilbert Projection Theorem, $H = V \bigoplus V^{\perp}$, that is, for any $x \in H$ there exists a unique $P(x) \in V$ such that $x - P(x) \in V^{\perp}$.

If $L(x) = 0$ for any $x \in H$, then we can just put $\varphi = 0_H$. Suppose now that there exists at least one $x \in H$ such that $L(x) \neq 0$, so that V is a proper subset of H. Then, $x - P(x) \in V^{\perp}$ but $x - P(x) \neq 0_H$ because $P(x) \in V$ but $x \notin V$, which tells us that $V^{\perp} \neq \{0_H\}$.

Choose some $z \in V^{\perp}$ such that $|z| = 1$, and for any $x \in H$, define

$$u(x) = L(x) \cdot z - L(z) \cdot x.$$

It follows that

$$L(u(x)) = L(x) \cdot L(z) - L(z) \cdot L(x) = 0$$

by linearity, so $u(x) \in V$ and $\langle u(x), z \rangle = 0$. Therefore,

$$0 = \langle u(x), z \rangle = \langle L(x) \cdot z - L(z) \cdot x, z \rangle = L(x) \cdot \langle z, z \rangle - L(z) \cdot \langle x, z \rangle = L(x) - \langle x, \overline{L(z)} \cdot z \rangle,$$

and rearranging terms, we have

$$L(x) = \langle x, \overline{L(z)} \cdot z \rangle.$$

This holds for any $x \in H$, so it follows that $\varphi = \overline{L(z)} \cdot z$ Finally, $\varphi \neq 0_H$ because $z \neq 0_H$ and $L(z) \neq 0$ by the facts that $\|z\| = 1$ and $z \notin V$.
Q.E.D.

2.1.5 L^p Spaces

For any $p \in [1, +\infty)$, we define the space L^p as the collection of all complex-valued random variables that have finite pth moments; formally,

$$L^p = \{X \mid \mathbb{E}|X|^p < +\infty\}.$$

Since random variables are functions from the outcome space into a metric space, L^p spaces are essentially function spaces. Indeed, we can define L^p spaces for more general kinds of functions, but that is beyond the scope of this text.

The L^p norm $\|\cdot\|_p : L^p \to \mathbb{R}_+$ is defined as

$$\|X\|_p = \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}}$$

for any $X \in L^p$. Note how the finiteness condition is essential for $\|\cdot\|_p$ to be a real-valued function. We can show that the pair $(L^p, \|\cdot\|_p)$ is a normed vector space over the complex field by using the following inequalities:

1) **Hölder's Inequality**
For any random variables X, Y and $p, q \in N_+$ such that $\frac{1}{p} + \frac{1}{q} = 1$,

$$\mathbb{E}|XY| \leq \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} \left(\mathbb{E}|Y|^q\right)^{\frac{1}{q}}.$$

2) **Minkowski's Inequality**
For any random variables X, Y and $p \in [1, +\infty)$,

$$\left(\mathbb{E}|X+Y|^p\right)^{\frac{1}{p}} \leq \left(\mathbb{E}|X|^p\right)^{\frac{1}{p}} + \left(\mathbb{E}|Y|^p\right)^{\frac{1}{p}}.$$

These two inequalities appear often and are very useful, even outside the context of L^p spaces (for instance, baby Rudin asks you to prove them as an exercise in chapter 6).

Let $d : L^p \times L^p \to \mathbb{R}_+$ be the metric induced by the L^p norm $\|\cdot\|_p$. The Riesz-Fischer theorem tells us that the pair (L^p, d) is a complete metric space; in other words, $(L^p, \|\cdot\|_p)$ is a Banach space over the complex field.

Of special interest are L^2 spaces, since they are the only kind of L^p space to admit an inner product. We define the L^2-inner product $\langle \cdot, \cdot \rangle_2 : L^2 \times L^2 \to \mathbb{C}$ as

$$\langle X, Y \rangle_2 = \mathbb{E}\left[X \cdot \overline{Y}\right]$$

for any $X, Y \in L^2$. We can easily verify that $\langle \cdot, \cdot \rangle_2$ satisfies the properties of an inner product on L^2, and note that the norm induced by $\langle \cdot, \cdot \rangle_2$ is precisely the L^2-norm $\|\cdot\|_2$. Since $(L^2, \|\cdot\|_2)$ is a Banach space over the complex field, $(L^2, \langle \cdot, \cdot \rangle_2)$ becomes a Hilbert space over the complex field. This property, in addition to the fact that almost every random variable of interest has a finite variance, is the reason for the popularity of L^2 spaces in economic analysis.

2.2 The No-Arbitrage Condition

So far, we have derived asset prices as equilibrium prices in an economy populated by utility maximizing rational investors. However, we also saw that the assumptions imposed in such general equilibrium models are often restrictive. **Aribtrage Pricing Theory (APT)**, pioneered by Ross (1976), offers an elegant solution. Instead of starting at fundamental assumptions such as utility-maximizing investors, APT starts at the pricing formula [1]

$$p_t = \mathbb{E}_t \left[\mathcal{M}_{t+1} X_{t+1} \right]$$

It then investigates sufficient conditions for a strictly positive SDF $\mathcal{M}_{t+1}$ to exist. These conditions, which are much weaker than the assumptions of a general equilibrium pricing model, become the starting point of APT, and results concerning asset pricing are derived on the basis of the pricing formula alone. Because the no-arbitrage condition is the key condition for the existence of a positive SDF, the pricing formula itself is often called the no-arbitrage condition [2].

Here we analyze the two-period case [3]. As in the exposition on the CAPM, we assume that investors make investment decisions at time 0 and receive their payoffs at time 1. Therefore, assets may be identified with their payoffs at time 1.

The set of all assets, or payoffs, is denoted P, and is taken to be a subset of L^2. Suppose there exists a pricing function $\pi : P \to \mathbb{R}$; that is, a function that assigns an asset with payoff X the price $p = \pi(X)$. We want to find sufficient conditions for there to exist a strictly positive SDF $\mathcal{M}$ such that

$$\pi(X) = \mathbb{E} \left[\mathcal{M} \cdot X \right],$$

so that the price of any asset is given as its expected discounted payoff.

[1] In Ross' original paper, he used the CAPM representation

$$\mathbb{E}_t \left[r_{t+1} \right] = r^f_{t+1} + \beta_i \cdot \lambda_m$$

as the starting point and studies sufficient conditions for such a factor representation to exist. Since the beta representation and the pricing formula are equivalent, as we saw above, we instead start from the pricing formula. Indeed, this is the exposition chosen in Cochrane (2011).

[2] For an example, consult my paper.

[3] The multi-period case with conditioning information is studied in Hansen and Richard (1987).

2.2.1 The Law of One Price

As a first step, we make the following assumptions:

A1. Complete and Linear Payoff Space

Let $\langle \cdot, \cdot \rangle_2$, $\|\cdot\|_2$ be the L^2-inner product and norm, and d the metric induced by $\|\cdot\|_2$. We assume that P is a linear subspace of L^2 such that (P, d) is a complete metric space. This ensures that $(P, \langle \cdot, \cdot \rangle_2)$ is a Hilbert space over the real field.

A2. Law of One Price (LOP)

We assume that the pricing function $\pi : P \to \mathbb{R}$ is linear. Specifically, for two payoffs $X, Y \in P$ and weights $w_1, w_2 \in \mathbb{R}$,

$$\pi(w_1 \cdot X + w_2 \cdot Y) = w_1 \cdot \pi(X) + w_2 \cdot \pi(Y).$$

A3. Continuity at 0

We assume that π is continuous at 0, that is, for any sequence $\{X_n\}_{n \in N_+}$ of payoffs such that $\|X_n\|_2 \to 0$ as $n \to +\infty$, we also have $\pi(X_n) \to 0$ as $n \to \infty$.

A4. Risk-Free Asset

We assume that P contains the risk-free payoff $X_f = 1$, which yields a payoff of 1 with certainty. In addition, $\pi(X_f) > 0$.

The first condition is a technical assumptions. The third assumption has the interpretation that it requires asset prices to shrink to 0 if the payoff shrink to 0, which seems reasonable. Finally, the fourth condition simply assumes that there exists a risk-free asset, a standard assumption we have made in the previous sections.

The second condition is referred to as the law of one price because it requires the repackaged asset with payoff $w_1 \cdot X + w_2 \cdot Y$ to have "one price". Suppose the LOP does not hold, so that, for instance,

$$\pi(w_1 \cdot X + w_2 \cdot Y) > w_1 \cdot \pi(X) + w_2 \cdot \pi(Y).$$

In this case, an investor would be able to earn a riskless positive profit by simply purchasing w_1 and w_2 units of the assets X and Y, packaging them as $w_1 \cdot X + w_2 \cdot Y$, and then selling them at the price $\pi(w_1 \cdot X + w_2 \cdot Y)$. This would cause the demand of the assets X and Y to increase, which has the effect of raising their prices $\pi(X)$ and $\pi(Y)$, so that ultimately equality holds.

An implication of the LOP is that the price of the risk-free asset should equal its payoff, since

$$\pi(X_f) = \pi(1) = w \cdot \pi(1) + (1-w) \cdot \pi(1)$$

We first show that, under the two assumptions above, there exists a non-zero SDF $\mathcal{M}$:

Theorem (Existence of Non-zero SDF)
Under assumptions A1 to A4, there exists $\mathcal{M} \in P$ such that $\mathcal{M} \neq 0$ and

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X]$$

for any $X \in P$.

Proof) Note that, under the two assumptions, $(P, \langle \cdot, \cdot \rangle_2)$ is a Hilbert space and $\pi : P \to \mathbb{R}$ a linear functional on P that is continuous at 0 and not equal to 0 everywhere on P. By the Riesz representation theorem, there exists a unique non-zero $\mathcal{M} \in P$ such that

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X]$$

for any $X \in P$.
Q.E.D.

Mathematically, we can say that the SDF $\mathcal{M}$ is simply the Riesz representation of the pricing function π when the LOP holds. Note also that

$$\pi(X_f) = \mathbb{E}[\mathcal{M}] > 0,$$

and since $\pi(X_f) = \frac{1}{1+r_f}$, where r_f is the risk-free rate of return, it follows that

$$1 + r_f = \frac{1}{\mathbb{E}[\mathcal{M}]},$$

as was derived from the C-CAPM.

2.2.2 The No-Arbitrage Assumption

We now want to find sufficient conditions for there to exist a strictly positive SDF. To this end, we make the following additional assumptions:

A5. The Payoff Space Includes all Derivatives

We assume that the payoff space P includes all derivatives with fundamentals whose payoffs are in P. In other words, if $X \in P$, then $f \circ X \in P$ for any measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $f \circ X$ is square integrable.

A6. No-Arbitrage Opportunities

We say there exist arbitrage opportunities if a non-negative payoff X that is positive with non-zero probability has a non-negative price, that is, if for any non-negative $X \in P$ such that $\mathbb{P}(X > 0) > 0$, we have $\pi(X) \leq 0$.

We assume that there are no arbitrage opportunities, that is,

$$\mathbb{P}(\{X > 0\} \cap \{\pi(X) \leq 0\}) = 0$$

for any $X \in P$ such that $X \geq 0$.

The no-arbitrage assumption tells us that, if an investor incurs no losses from an asset and there is a non-negligible chance for that asset to deliver a positive payoff, then the price of that payoff should be positive. This is clearly much stronger than the usual no arbitrage requirement that an asset with an assured positive payoff should have positive price.

Under these additional assumptions, we can show that there exists a positive SDF:

Theorem (Existence of Positive SDF)
Under assumptions A1 to A6, there exists a $\mathcal{M} \in P$ such that $\mathcal{M} > 0$ and

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X]$$

for any $X \in P$.

Proof) We already showed above that, if assumptions A1 to A4 hold, then there exists an $\mathcal{M} \in P$ such that $\mathcal{M} \neq 0$ and

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X].$$

It remains to show that $\mathcal{M} > 0$ when we also assume A5 and A6.

Suppose that $\mathcal{M} \leq 0$ with positive probability. Then, defining

$$X = I_{\{\mathcal{M} \leq 0\}},$$

$X \in P$ because P contains all derivatives with fundamentals in P, $X \geq 0$, and $X > 0$ with the same positive probability that $\mathcal{M} \leq 0$. It follows from the no-arbitrage assumption that

$$\mathbb{P}(\{X > 0\} \cap \{\pi(X) \leq 0\}) = 0.$$

Since $\mathcal{M}$ is the Riesz representation of π, we can see that

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X] = \mathbb{E}\left[\mathcal{M} \cdot I_{\{\mathcal{M} \leq 0\}}\right] \leq 0.$$

In other words,

$$\mathbb{P}(X > 0) = \mathbb{P}(\{X > 0\} \cap \{\pi(X) \leq 0\}) = 0,$$

which allows us to conclude that

$$\mathbb{P}(\mathcal{M} \leq 0) = \mathbb{P}(X = 1) = \mathbb{P}(X > 0) = 0.$$

This contradicts our initial assumption, so it must be the case that $\mathcal{M} > 0$ with probability 1.
Q.E.D.

Going forward, we collectively refer to assumptions A1 to A6 as the no-arbitrage condition, so that the no-arbtirage equation

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X]$$

holds for any $X \in P$.

2.3 The Risk-Neutral Measure

Under the no-aribtrage condition above, we saw that the no-arbitrage equation

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X]$$

holds for any $X \in P$. Here we develop an alternative representation of the no-arbitrage equation.

2.3.1 Some (Really Rudimentary) Measure Theory

First, we introduce some very rudimentary measure theoretic concepts. Denote the sample space by Ω, and the set of all events by $\mathcal{H}$; obviously, $\mathcal{H}$ is a collection of subset of Ω, a set of sets. An example of an event set is the power set 2^{Ω}, which collects all the subsets of Ω. We require the set of all events to be a σ-algebra on Ω:

a) **Inclusion of Empty Set and Entire Set**
$\mathcal{H}$ includes the empty and entire sets $\emptyset$ and Ω.

b) **Closed under Complements**
$\mathcal{H}$ is closed under complementation, that is, for any $H \in \mathcal{H}$,

$$H^c = \Omega \setminus H \in \mathcal{H}.$$

c) **Closed under Countable Unions**
$\mathcal{H}$ is closed under countable unions, that is, for any countable collection $\{H_n\}_{n \in N_+}$ of sets in $\mathcal{H}$,

$$H = \bigcup_n H_n \in \mathcal{H}.$$

A probability measure $\mu : \mathcal{H} \to [0,1]$ is a function defined on the set of all events $\mathcal{H}$ that satisfies the following properties:

a) **Empty Sets have measure 0**
μ assigns measure 0 to the empty set; $\mu(\emptyset) = 0$.

b) **Countable Additivity**
For any disjoint countable collection $\{H_n\}_{n \in N_+}$ of sets in $\mathcal{H}$,

$$\mu\left(\bigcup_n H_n\right) = \sum_{n=1}^{\infty} \mu(H_n).$$

c) **Total Mass is 1**
μ assigns a total mass of 1 to the entire set; $\mu(\Omega) = 1$.

We can easily show that any probability measure is also finitely additive, countably subadditive and monotonic. The triple $(\Omega, \mathcal{H}, \mu)$ is called a probability space.

Measure theory was originally developed as the foundation for a system of integration more general and abstract than Riemann integration. Here we briefly introduce some concepts related to abstract integration. Let $(\Omega, \mathcal{H}, \mu)$ be a probability space. Consider the simple function $f : \Omega \to \mathbb{R}_+$ defined as

$$f = \sum_{i=1}^{n} a_i \cdot I_{H_i},$$

where $H_1, \cdots, H_n \in \mathcal{H}$ and $a_1, \cdots, a_n$ are non-negative real numbers. In other words, f is a function that takes on finitely many values. The integral of f with respect to μ is defined as

$$\int_\Omega f d\mu = \sum_{i=1}^{n} a_i \cdot \mu(H_i).$$

Note that this is just the expected value of the random variable f.

Now consider an aribtrary non-negative function $f : \Omega \to [0, +\infty]$. We say that f is measurable if there exists an increasing sequence $\{f_n\}_{n \in N_+}$ of simple functions that converges pointwise to f. It turns out that this is equivalent to requiring that

$$\{\omega \in \Omega \mid f(\omega) \leq x\} \in \mathcal{H}$$

for any $x \in \mathbb{R}$, and that most functions are measurable (including continuous functions that those with countably many discontinuities). The integral of f with respect to μ is defined as

$$\int_\Omega f d\mu = \sup_{n \in N_+} \int_\Omega f_n d\mu.$$

In other words, since $\{f_n\}_{n \in N_+}$ approximates f from below, the integral of f is approximated by the integrals of f_n. The integral, defined in this way, satisfies many of the useful properties of integration, such as monotonicity and linearity.

Finally, let $f : \Omega \to \mathbb{R}$ be an aribtrary real-valued function. The positive and negative parts of f are defined as

$$f^+ = \max(f, 0) \quad \text{and} \quad f^- = -\min(f, 0).$$

We say that f is integrable with respect to μ if

$$\int_\Omega |f| d\mu < +\infty,$$

and the integral of f in this case is defined as

$$\int_\Omega f d\mu = \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu.$$

Again, the integral defined in this way satisfies properties of integration such as montonicity and linearity.

We can now cast probability theory and expectations in a measure-theoretic light. Let $(\Omega, \mathcal{H}, \mu)$ be our probability space. A random variable X is a real-valued measurable function defined on the sample space Ω. X is said to be μ-integrable if the non-negative variable $|X|$ has finite expectation, or integral. In this case, its expected value is defined as the integral of X with respect to μ, that is, as

$$\mathbb{E}[X] := \int_\Omega X d\mu.$$

2.3.2 Mathematical Definition of the Risk-Neutral Measure

Usually, we work with the **physical measure**, or the P-measure, $\mathbb{P}$. This is the probability measure that, for a given event, yields the actual probability of the event occuring. The expectation in the no-arbitrage equation

$$\pi(X) = \mathbb{E}[\mathcal{M} \cdot X] := \int_\Omega (\mathcal{M} \cdot X) d\mathbb{P}$$

is taken with respect to the P-measure.

Now consider an alternative probability measure $\mathbb{Q}$, or the Q-measure, defined as

$$\mathbb{Q}(H) := \mathbb{E}\left[(1+r_f)\mathcal{M} \cdot I_H\right] = \int_H (1+r_f)\mathcal{M} d\mathbb{P}$$

for any $H \in \mathcal{H}$, where I_H is the indicator function that equals 1 if an outcome is included in H and 0 otherwise. Note that $\mathbb{Q}$ assigns 0 to the empty set and, since

$$\mathbb{Q}(\Omega) = (1+r_f)\mathbb{E}[\mathcal{M}] = 1,$$

1 to the entire sample space. It can also be shown that $\mathbb{Q}$ satisfies the countable additivity condition, so that $\mathbb{Q}$ is a proper probability measure on $\mathcal{H}$ with all the properties expected of one.

An important result in measure theory tells us that, for any random variable X such that $\mathcal{M}X$ is $\mathbb{P}$-integrable,

$$\int_\Omega X d\mathbb{Q} = \mathbb{E}\left[(1+r_f)\mathcal{M}\cdot X\right] = \int_\Omega \left((1+r_f)\mathcal{M}X\right) d\mathbb{P}.$$

Denote expectations with respect to $\mathbb{Q}$ by $\mathbb{E}^{\mathbb{Q}}[\cdot]$. Then, this result tells us that, for any payoff X that belongs to the payoff space P,

$$\mathbb{E}^{\mathbb{Q}}[X] = (1+r_f)\cdot\mathbb{E}[\mathcal{M}X] = (1+r_f)\pi(X),$$

or equivalently,

$$\pi(X) = \frac{1}{1+r_f}\mathbb{E}^{\mathbb{Q}}[X].$$

This result is referred to as the **first fundamental theorem of asset pricing**: under our assumptions, there exists a measure $\mathbb{Q}$ under which the price of any asset is equal to the present value of its expected payoff.

The equation also reveals why $\mathbb{Q}$ is called the **risk-neutral measure**; if investors are risk-neutral, then under the no arbitrage the expected return from selling an asset ($\pi(X)$) equals the expected discounted return from holding the asset and selling it next period $\left(\mathbb{E}^{\mathbb{Q}}\left[\frac{1}{1+r_f}X\right]\right)$.

It is also sometimes called the **equivalent martingale measure**, since it is the measure "equivalent" to the physical measure $\mathbb{P}$ that turns asset price processes into martingales.

2.3.3 Intuitive Meaning of the Risk-Neutral Measure

While the risk-neutral measure is a mathematical construct, it also has an appealing intuitive meaning. For simplicity, assume that the asset market is complete, so that any L^2 random variable is contained in the payoff space P. By design, the probability of some event $H \in \mathcal{H}$ under the risk-neutral measure is

$$\mathbb{Q}(H) = \mathbb{E}\left[(1+r_f)\mathcal{M}\cdot I_H\right] = (1+r_f)\cdot\pi(I_H),$$

where we used the fact that $\pi(X) = \mathbb{E}[\mathcal{M}\cdot X]$ forany $X \in P$. In other words, the probability of H under the risk-neutral measure is proportional to the price of an asset that yields a payoff of 1 if and only if the event H occurs.

Suppose there are two events, H_1 and H_2, of equal (phyiscal) probability. Let the event H_2 be riskier than H_1; for instance, H_1 can be the event that war breaks out far from

home and H_2 the event that war breaks out near home. If investors are risk-neutral, then they choose assets based only on their expected payoff; since the two assets have the same expected payoffs

$$\mathbb{E}[I_{H_1}] = \mathbb{P}(H_1) = \mathbb{P}(H_2) = \mathbb{E}[I_{H_2}],$$

a risk-neutral investor would be indifferent to either asset. However, if investors are risk averse, then they would prefer the asset with payoff I_{H_1} over the asset with payoff I_{H_2}, since they have the same expected payoff but I_{H_2} is riskier than I_{H_1}.

In reality, investors are risk-neutral and thus asset 1 will command a higher demand and higher price than asset 2, or equivalently, the expected rate of return for asset 2 will be higher than asset 1 to compensate for the additional risk. In terms of the risk-neutral measure, this means that

$$\mathbb{Q}(H_1) = (1+r_f)\cdot\pi(I_{H_1}) > (1+r_f)\cdot\pi(I_{H_2}) = \mathbb{Q}(H_2).$$

Thus, if risk averse investors lived in a world with $\mathbb{Q}$ as the probability measure, they could make the same choice as in a world with $\mathbb{P}$ based only on the expected payoff of the assets I_{H_1} and I_{H_2}. Another way to put this is that the risk-neutral measure implements information about the risk of an asset into its expected payoff, making it possible for risk averse investors to rely only on an asset's expected payoff when making investment decisions.

Since the expected payoff corresponds to the first moments of an asset's payoff, and the risk, being the variance, corresponds to the second moments, the risk-neutral measure can also be said to collapses a two-dimensional problem into a one-dimensional problem.

2.4 The Empirical SDF

We now have on hand two probability measures, the physical measure and the risk-neutral measure. The two measures are related through the SDF $\mathcal{M}$ as follows:

$$\mathbb{Q}(H) = \exp\Big(-r_f\Big) \cdot \mathbb{E}\left[\mathcal{M} \cdot I_H\right],$$

for any event H, where we have approximated $1 + r_f$ with $\exp\big(r_f\big)$[4]. In this section, we choose a general form for the SDF that has interesting implications for normally distributed variables under the risk-neutral and physical measures.

2.4.1 The Empirical SDF and Girsanov's Theorem

Let Z be an n-dimensional standard normally distributed random vector under the physical measure. We define the **empirical SDF** as

$$\mathcal{M} = \exp\Big(-r_f - \frac{1}{2}\lambda'\lambda - \lambda' Z\Big),$$

where λ is some n-dimensional vector. This empirical SDF satisfies, in the first place, the property that $\mathbb{E}[\mathcal{M}] = \exp\big(-r_f\big)$; this can be seen by noting that

$$\begin{aligned}\mathbb{E}[\mathcal{M}] &= \exp\Big(-r_f - \frac{1}{2}\lambda'\lambda\Big) \cdot \mathbb{E}\left[\exp\big(-\lambda' Z\big)\right] \\ &= \exp\Big(-r_f - \frac{1}{2}\lambda'\lambda\Big) \cdot \exp\Big(\frac{1}{2}\lambda'\lambda\Big) = \exp\big(-r_f\big)\end{aligned}$$

using the formula for the MGF of normally distributed variables.

In addition, define $Z^* = \lambda + Z$, so that Z^* is an n-dimensional normally distributed random vector with mean λ and variance I_n under the physical measure. We can show that Z^* is a standard normal random vector under the risk-neutral measure: for any $t \in \mathbb{R}^n$,

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}\left[\exp\big(t' Z^*\big)\right] &= \exp\big(r_f\big) \cdot \mathbb{E}\left[\mathcal{M} \cdot \exp\big(t' Z^*\big)\right] \\ &= \mathbb{E}\left[\exp\Big(-\frac{1}{2}\lambda'\lambda - \lambda' Z\Big) \cdot \exp\big(t' Z^*\big)\right] \\ &= \exp\Big(-\frac{1}{2}\lambda'\lambda + t'\lambda\Big) \cdot \mathbb{E}\left[\exp\big(-(\lambda - t)' Z\big)\right] \\ &= \exp\Big(-\frac{1}{2}\lambda'\lambda + t'\lambda\Big) \cdot \exp\Big(\frac{1}{2}(\lambda - t)'(\lambda - t)\Big) \\ &= \exp\Big(\frac{1}{2}t't\Big),\end{aligned}$$

[4]This follows from a first order Taylor expansion, and if the approximation error bothers you, then we can always define r_f as $r_f = \log\big(\pi(X_f)\big)$.

where we again used the formula for the MGF of normally distributed variables. Therefore, the MGF of Z^* under the risk-neutral measure is exactly the MGF of the the n-dimensional standard normal distribution. Since two random vectors with the same MGF are identically distributed, this implies that $Z^* \sim \mathcal{N}[O_{n\times 1}, I_n]$.

This result, known in continuous time as **Girsanov's theorem**, shows us that, if the SDF assumes the (stochastic exponential) form above, then simply changing the location of a random vector that is standard normally distributed under the physical measure can produce a random vector that has the same distribution under the risk-neutral measure. This monumental result allows us to shift between the physical and risk-neutral measures via a simple change in mean, and is one of the reasons for the widespread use of Gaussian innovations in financial models.

2.4.2 Intuitive Meaning of the Empirical SDF

At first glance, the empirical SDF might seem like a purely mathematical construct, designed to ease the transition between the risk-neutral and physical measures. However, the form of the empirical SDF can also be motivated by general equilibrium pricing models such as the C-CAPM, and in fact, making this connection elucidates the economic meaning of λ.

Recall that, in the C-CAPM, the SDF is given as

$$\mathcal{M} = \beta \cdot \frac{u'(C_1)}{u'(C_0)},$$

where the time subscripts have been modified to accomodate our two-period environment. Under a CRRA utility function given by

$$u(C) = \frac{C^{1-\theta}}{1-\theta},$$

the SDF becomes

$$\mathcal{M} = \beta \cdot \left(\frac{C_1}{C_0}\right)^{-\theta},$$

and if the subjective discount rate is given as ρ, we can express

$$\beta = \frac{1}{1+\rho} \approx \exp(-\rho).$$

Taking logs on both sides thus yields

$$\log(\mathcal{M}) = -\rho - \theta\left(\log(C_1) - \log(C_0)\right) \approx -r_f - \theta \cdot \Delta c$$

$$= -\rho - \theta \cdot \mathbb{E}[\Delta c] - \theta\sigma(\Delta c) \cdot z,$$

where Δc is consumption growth, $\sigma(\Delta c)$ is its standard deviation and $z = \frac{\Delta c - \mathbb{E}[\Delta c]}{\sigma(\Delta c)}$.

Suppose that $z \sim \mathcal{N}(0,1)$ under the physical measure. The SDF satisfies the condition $\mathbb{E}[\mathcal{M}] = \exp\Big(-r_f\Big)$, so

$$\begin{aligned}\exp\Big(-r_f\Big) = \mathbb{E}[\mathcal{M}] &= \mathbb{E}[\exp(-\rho - \theta \cdot \mathbb{E}[\Delta c] - \theta\sigma(\Delta c) \cdot z)] \\ &= \exp(-\rho - \theta \cdot \mathbb{E}[\Delta c]) \cdot \mathbb{E}[\exp(-\theta\sigma(\Delta c) \cdot z)] \\ &= \exp\Big(-\rho - \theta \cdot \mathbb{E}[\Delta c] + \frac{1}{2}\theta^2\sigma(\Delta c)^2\Big).\end{aligned}$$

Therefore,

$$\mathcal{M} = \exp\Big(-r_f - \frac{1}{2}\theta^2\sigma(\Delta c)^2 - \theta\sigma(\Delta c) \cdot z\Big),$$

which is exactly the form of the empirical SDF. Recall that, in our analysis of the C-CAPM, the relative risk aversion coefficient θ represented the market price of risk, that is, it determined how much compensation investors demand for an additional unit of risk. Meanwhile, (normalized) consumption growth z represented the systematic risk factor, with the correlation of an asset's rate of return with z determining the amount of risk present in the factor.

This suggests that Z and λ in the general empirical SDF

$$\mathcal{M} = \exp\Big(-r_f - \frac{1}{2}\lambda'\lambda - \lambda' Z\Big)$$

can be interpreted as the vector of risk factors and the market prices of risk, respectively.

2.5 Extension to a Multi-Period Setting

So far, we have studied the no-arbitrage condition, the risk-neutral measure and the empirical SDF in a two-period setting. The extension of this model to a multi-period setting is straightforward.

As in the beginning, we assume that to every asset is an associated sequence of payoffs $\{X_t\}_{t\in N_+}$, where X_{t+1} is the payoff the investor receives at time $t+1$ if she invests in the asset at time t and sells it at time $t+1$. Each time $t+1$ payoff X_{t+1} is contained in the time $t+1$ payoff space P_{t+1}, a subset of L^2.

We assume that asset prices at time t are a function of its payoff at time $t+1$, that is, we assume the existence of a function $\pi_t : P_{t+1} \to \mathbb{R}$ such that the time t price p_t of an asset with time $t+1$ payoff X_{t+1} is given as

$$p_t = \pi_t(X_{t+1}).$$

Suppose assumptions A1 to A6 hold for each period, appropriately tailored to accomodate the fact that information is accumulated starting from the first period onward[5]. Then, for any time t there exists an SDF $\mathcal{M}_{t+1} \in P_{t+1}$ such that

$$p_t = \mathbb{E}_t\left[\mathcal{M}_{t+1} \cdot X_{t+1}\right],$$

where the subscript t denotes that the expectation is conditional on information up to time t. Putting $\mathcal{M}_0 = 1$, the sequence $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ is referred to as our **SDF process**.

As in the previous sections, under the additional assumption that $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ is uniformly integrable[6], there exists a risk-neutral measure $\mathbb{Q}$ such that

$$p_t = \mathbb{E}_t\left[\mathcal{M}_{t+1} \cdot X_{t+1}\right] = \mathbb{E}_t^{\mathbb{Q}}\left[\exp\left(-r_{t+1}^f\right) \cdot X_{t+1}\right]$$

for any $X_{t+1} \in P_{t+1}$. The intuitive meaning and implication of the risk-neutral measure remain unchanged.

Finally, $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ is said to be an **empirical SDF process** if

$$\mathcal{M}_{t+1} = \exp\left(-r_{t+1}^f - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t \cdot v_{t+1}^{\mathbb{P}}\right)$$

for any $t \in \mathbb{N}$, where $v_{t+1}^{\mathbb{P}}$ is an n-dimensional standard normally distributed random vector under the physical measure and λ_t is some n-dimensional random vector known at time

[5] We omit explicitly stating the extension of the assumptions. If curious, consult me directly.

[6] This is a technical assumption that ensures the SDF process is well-behaved in the far future. If the time index is bounded, then we can omit the uniform integrability assumption.

t. Defining

$$v_{t+1}^{\mathbb{Q}} = \lambda_t + v_{t+1}^{\mathbb{P}},$$

under the above empirical SDF process $v_{t+1}^{\mathbb{Q}}$ follows an n-dimensional standard normal distribution thanks to Girsanov's theorem. Here, too, λ_t represents the market price of risk, or the sensitivity of investors to an additional unit of risk, and $v_{t+1}^{\mathbb{P}}$ the normalized systematic risk factors.

Chapter 3

Empirical Models of the Yield Curve

In this chapter, we first introduce some fundamental concepts concerning bonds and yields, and then investigate the components of affine term structure models. Afterward, we study some empirical models used to model the yield curve.

3.1 Bonds and Yields

A bond with face/par value A and maturity T is an asset that pays the fixed amount A at time T. For this reason, it is sometimes called a fixed-income security. A **coupon bond** is a bond that pays a dividend c in fixed intervals until the time of maturity, while a **zero-coupon bond** is a bond that pays no coupons. Since a coupon bond that pays a coupon c each period until maturity can be seen as a composite of various zero-coupon bonds with face value c (this is called the STRIPS principle), we focus only on zero-coupon bonds. In addition, we focus only on government bonds such as treasury bills, as the analysis of corporate bonds requires the consideration of default risk.

Going forward, we will denote by $P_t(\tau)$ the price of a zero-coupon bond at time t with face value 1 and τ periods left to maturity. A bond with τ periods left to maturity at time t will also be called a τ-period bond at time t. The **yield** of this bond is defined as

$$Y_t(\tau) = -\frac{1}{\tau}\log(P_t(\tau)).$$

Note that

$$Y_t(\tau) = \frac{1}{\tau}\left(\log(1) - \log(P_t(\tau))\right) \approx \frac{1}{\tau}\cdot\frac{1-P_t(\tau)}{P_t(\tau)}.$$

Since 1 is the amount that the investor stands to receive at maturity, $\frac{1-P_t(\tau)}{P_t(\tau)}$ is the rate of return from investing in the bond at time t and holding it to maturity. In other words, $Y_t(\tau)$ is the rate of return of the bond upon maturity, calculated at time t and divded by

the time left to maturity. This is why the time t yield $Y_t(\tau)$ is said to be the average rate of return of the bond until maturity.

Note that $Y_t(1)$ is the yield of the zero-coupon bond with one period left to maturity. Since a zero-coupon bond with one period left to maturity provides a risk-free one-period ahead payoff equal to 1, we can see that $Y_t(1)$ is essentially the risk-free rate of return. In the term structure literature, the risk-free (one period ahead) rate of return is called **the short rate** and is denoted

$$r_t = Y_t(1).$$

Clearly, $Y_t(0) = 0$, since the bond price at the time of maturity is $P_t(0) = 1$.

While we defined bond yields using bond prices, we can conversely recover bond prices from yields:

$$P_t(\tau) = \exp(-\tau \cdot Y_t(\tau)) \approx \frac{1}{1 + \tau \cdot Y_t(\tau)}.$$

This tells us that the time t yield is also the average discount rate under which the time t bond price is its discounted face value.

The h-period ahead holding period return for a τ-period bond at time t is given as

$$r_{t,t+h}^{(\tau)} = \log(P_{t+h}(\tau - h)) - \log(P_t(\tau)) \approx \frac{P_{t+h}(\tau - h) - P_t(\tau)}{P_t(\tau)}.$$

This is the rate of return from holding a bond with maturity in τ periods from time t to time $t + h$, and then selling it. The h-period ahead excess return for a τ-period bond is now given as

$$exr_{t,t+h}^{(\tau)} = r_{t,t+h}^{(\tau)} - h \cdot Y_t(h),$$

that is, as the difference in the rates of return from investing in a τ-period bond at time t and selling it at time $t + h$, and that from an h-period bond with a guaranteed payoff of 1 at time $t + h$. Its expected value is the h-period ahead expected excess return, or h-period ahead risk premium, of a τ-period bond.

The one-period ahead holding period return and excess return from time t to $t + 1$ are denoted by

$$r_{t+1}^{(\tau)} \quad \text{and} \quad exr_{t+1}^{(\tau)},$$

so that the usual (one period ahead) risk premium is equal to

$$\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] = \mathbb{E}_t\left[r_{t+1}^{(\tau)}\right] - r_t.$$

3.1.1 The Forward Rate

This is a related, but somewhat more complicated concept. At time t, suppose we are interested in the future risk-free (one period ahead) rate. Specifically, we are interested in the value of r_{t+h}. One way to investigate the current level of the future risk-free rate is as follows:

Step 1: Issue a zero-coupon bond with maturity at time $t+h$. This will give us $P_t(h)$ units of the numeraire.

Step 2: Buy $P_t(h)$ worth of zero-coupon bonds with maturity at time $t+h+1$. This will leave us with $\frac{P_t(h)}{P_t(h+1)}$ units of zero-coupon bonds maturing at time $t+h+1$.

The asset constructed as such will require:

Cost of 1 at time $t+h$, since that is when the bond we issued will mature.

Payoff of $\frac{P_t(h)}{P_t(h+1)}$ at time $t+h+1$, since that is when the bonds we purchased will mature.

This asset yields a risk-free payoff of $\frac{P_t(h)}{P_t(h+1)}$ at time $t+h+1$ in exchange for a cost of 1 at time $t+h$, and requires no other costs, nor does it yield any other benefits. As such, the rate of return of this asset can be viewed as the current level of the risk-free rate of return h periods from now. It is given as

$$f_t^{(h)} = \frac{P_t(h)}{P_t(h+1)} - 1 = \frac{P_t(h) - P_t(h+1)}{P_t(h+1)},$$

and is called **the forward rate**. A first order Taylor approximation yields

$$f_t^{(h)} \approx \log(P_t(h)) - \log(P_t(h+1)),$$

so that

$$-\log(P_t(\tau)) = \sum_{h=0}^{\tau-1} (\log(P_t(h)) - \log(P_t(h+1))) = \sum_{h=0}^{\tau-1} f_t^{(h)}.$$

Therefore,

$$Y_t(\tau) = -\frac{1}{\tau}\log(P_t(\tau)) = \frac{1}{\tau}\sum_{h=0}^{\tau-1} f_t^{(h)}.$$

In other words, the time t yield of a bond with maturity in τ periods is the average of the forward rates across the remaining life of the bond.

Another useful expression for the forward rate can be obtained in terms of $h+1$-period ahead holding period returns:

$$\begin{aligned} f_t^{(h)} &= \log(P_t(h)) - \log(P_t(h+1)) \\ &= -\log(P_{t+h}(1)) + [\log(P_{t+h}(1)) - \log(P_t(h+1))] - [\log(P_{t+h}(0)) - \log(P_t(h))] \\ &= r_{t+h} + r_{t,t+h}^{(h+1)} - r_{t,t+h}^{(h)}, \end{aligned}$$

where the we used the fact that

$$r_{t+h} = Y_{t+h}(1) = -\log(P_{t+h}(1)).$$

In other words, the difference between the forward rate $f_t^{(h)}$ from time $t+h$ to time $t+h+1$ and the risk-free rate across the same time interval equals the difference in holding period returns from time t to time $t+h$ between an $h+1$-period bond and an h-period bond.

3.1.2 The Expectations Hypothesis

The **expectations hypothesis (EH)** refers to a series of equalities that must hold under the law of one price when investors are risk-neutral. In this case, the law of one price dictates that any two assets with the same price (expected payoff) must have the same expected payoff (price). Following the lead of Cochrane and Piazzesi (2008), we formulate the expectations hypothesis in three equivalent forms:

1) **For Long Term Yields**
   The expected payoffs from investing a unit of the numeraire in a τ-period bond until maturity and that from successively investing it in a one-period bond for τ periods must be equal:

$$1 + \tau \cdot Y_t(\tau) = \prod_{h=0}^{\tau-1} (1 + Y_{t+h}(1)).$$

   Under a first order Taylor expansion, the above equation can be formulated as

$$Y_t(\tau) = \frac{1}{\tau} \sum_{h=0}^{\tau-1} r_{t+h}.$$

   Since the right hand side is unknown at time t, under risk neutrality we equate expected returns:

$$Y_t(\tau) = \frac{1}{\tau} \sum_{h=0}^{\tau-1} \mathbb{E}_t \left[r_{t+h} \right].$$

2) **For Short Rates**
The risk premium of a τ-period bond is 0:

$$\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] = 0,$$

or equivalently,

$$\mathbb{E}_t\left[r_{t+1}^{(\tau)}\right] = r_t.$$

3) **For Forward Rates**
The h-period ahead forward rate is equal to the expected h-period ahead risk-free rate:

$$f_t^{(h)} = \mathbb{E}_t\left[r_{t+h}\right].$$

Empirically, the expectations hypotheses have found little support, most likely due to the risk aversion of investors. When taking into consideration that investors are risk averse, the hypotheses above can be reformulated with the appropriate risk premia appended to either side:

$$Y_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[r_{t+h}\right] + TP_t(\tau)$$

$$\mathbb{E}_t\left[r_{t+1}^{(\tau)}\right] = r_t + RP_t(\tau)$$

$$f_t^{(h)} = \mathbb{E}_t\left[r_{t+h}\right] + FRP_t(h),$$

where $TP_t(\tau)$, $RP_t(\tau)$ and $FRP_t(h)$ are referred to as the **Term Premium**, (one-period ahead) **Risk Preimum** and **Forward Premium**. In particular, the term

$$EH_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[r_{t+h}\right]$$

is referred to as the EH component of a τ-period yield, so that a long term yield can be expressed as the sum of its EH component and the term premium.

3.2 Principal Components of the Yield Curve

We will now investigate some empirical models of the yield curve, before imposing no-arbitrage restrictions in the next chapter. The simplest and most ubiquitous means of modeling the yield curve is to summarize the variation present in the sample yields using three factors: the level, slope and curvature. These factors are taken as the first three principal components of a given panel of yields, and they are usually understood as representing the short, long and middle ends of the yield curve. Here we briefly introduce principal component analysis, before applying it to the yield curve.

3.2.1 Principal Component Analysis

Principal component analysis (PCA) is a means of summarizing the variation present in multiple random variables using linear combinations of said variables. Suppose there are k square integrable random variables $X_1, \cdots, X_k$ with mean zero that are collected into a k-dimensional real random vector $X = (X_1, \cdots, X_k)$. Denote the covariance matrix of X by

$$\Sigma = \mathbb{E}\left[XX'\right],$$

and assume that it is positive definite.

If k is large, then instead of using all k variables for our analysis, it may be more advantageous (and of course, parsimonious) to use a select few linear combinations of the k variables that represent their co-movement. Intuitively, we can think of the information contained in the k variables as divided into **signal** and **noise**. Signal refers to the unique information stored in each variable, represented by the variance of a certain variable. On the other hand, noise is redundant information contained in a certain variable, represented by the covariance of this variable with others (the covariance represents, heuristically, the overlap between information in different variables). Therefore, we want to find linear combinations of the k variables that best capture the signal, that is, maximize the variance, while minimizing the noise, or covariance. Mathematically, this problem can be formulated as follows:

First, we want to find coefficients $v \in \mathbb{R}^k$ such that the variance of the linear combination

$$v'X = \sum_{i=1}^{k} v_i \cdot X_i$$

is maximized. The variance in question is given by

$$\text{Var}\left(v'X\right) = v'\mathbb{E}\left[XX'\right]v = v'\Sigma v.$$

Second, if there are two such coefficients $v_1, v_2 \in \mathbb{R}^k$, then we want to ensure that the linear combination $v_1'X$ and $v_2'X$ are orthogonal, or uncorrelated. The covariance in question is given by

$$\text{Cov}\left(v_1'X, v_2'X\right) = v_1'\mathbb{E}\left[XX'\right]v_2 = v_1'\Sigma v_2.$$

Third, we wnat to normalize the coefficients v so that $|v| = 1$, in order to preclude trivial results (such as sending the coefficients to ∞ to maximize the variance).

We now search for these coefficients v one by one. We first search for the first coefficient $v_1 \in \mathbb{R}^k$, which solves the constrained maximization problem

$$\begin{aligned} \max_{v \in \mathbb{R}^k} \quad & v'\Sigma v \\ \text{subject to} \quad & |v| = 1. \end{aligned}$$

For the time being, we need not worry about the 0 covariance condition because v_1 is the first set of conditions. Since the objective function is strictly quasiconcave and the mapping $v \mapsto v'v$ is quasiconcave, there exists a unique solution $v_1 \in \mathbb{R}^k$ to the maximization problem. This solution must satisfy the first order condition

$$\Sigma v_1 = \lambda_1 v_1$$

for some $\lambda_1 \in \mathbb{R}$. We can immediately see that v_1 must be an unit eigenvector of the positive definite matrix Σ, and that the value of the objective function at v_1 is $v_1'\Sigma v_1 = \lambda_1$, the eigenvalue of Σ corresponding to v_1. It follows that v_1 must be a unit eigenvector of Σ associated with its largest eigenvalue λ_1 [1]. The linear combination $v_1'X$ is called the first **Principal Component (PC)** of X.

Suppose now that we have found the first $1 \leq m < k$ coefficients $v_1, \cdots, v_m \in \mathbb{R}^k$ with unit 1 such that the PCs $v_1'X, \cdots, v_m'X$ have ordered variances

$$\lambda_1 \geq \cdots \geq \lambda_m > 0,$$

[1] Recall that the eigenvalues of symmetric matrices can be ordered because they are all real, and that in particular the eigenvalues of positive semidefinite matrices are all non-negative

and the covariances of any two PCs $v_i'X$ and $v_j'X$ is 0:

$$v_i'\Sigma v_j = 0 \quad \text{for any } 1 \leq i \neq j \leq m.$$

The $m+1$th coefficient $v_{m+1} \in \mathbb{R}^k$ solves the following maximization problem:

$$\begin{aligned} \max_{v \in \mathbb{R}^k} \quad & v'\Sigma v \\ \text{subject to} \quad & |v| = 1, \\ & v'\Sigma v_i = 0 \quad \text{for any } 1 \leq i \leq m \end{aligned}$$

The Lagrangian for this problem is

$$\mathcal{L} = v'\Sigma v + \lambda\left(1 - v'v\right) - \sum_{i=1}^{m} \gamma_i \cdot v'\Sigma v_i$$

and by the first order conditions of maximization, there exist $\lambda_{m+1}, \gamma_1, \cdots, \gamma_m \in \mathbb{R}$ such that

$$\Sigma v_{m+1} = \lambda_{m+1} \cdot v_{m+1} + \sum_{i=1}^{m} \gamma_i \cdot \Sigma v_i.$$

Since $v_1, \cdots, v_m$ are orthonormal eigenvectors of Σ with non-zero eigenvalues, we can see that

$$0 = v_{m+1}'\Sigma v_i = \lambda_i \cdot v_{m+1}' v_i,$$

and as such that $v_{m+1}'v_i = 0$. Of course, the zero covariance restriction tells us that $v_i'\Sigma v_{m+1} = 0$ for any $1 \leq i \leq n$. By implication, for any $1 \leq i \leq m$, premultipyling the first order condition on both sides by v_i yields

$$0 = v_i'\Sigma v_{m+1} = \sum_{j=1}^{m} \gamma_j \cdot v_i'\Sigma v_j = \gamma_i \cdot v_i'\Sigma v_i = \gamma_i \cdot \lambda_i,$$

where the third inequality follows from the fact that $v_i'\Sigma v_j = 0$ for any $j \neq i$. Therefore, each γ_i is equal to 0, using which the first order conditions can be rewritten as

$$\Sigma v_{m+1} = \lambda_{m+1} \cdot v_{m+1}.$$

Therefore, v_{m+1} is a unit eigenvector of Σ with eigenvalue λ_{m+1} that is orthogonal to the eigenvectors $v_1, \cdots, v_m$. Note that, if λ_{m+1} is greater than any one eigenvalue in $\lambda_1, \cdots, \lambda_m$, then v_{m+1} would have been chosen as one of the earlier coefficients. Therefore, λ_{m+1} must be smaller than or equal to λ_m. $v_{m+1}'X$ is our $m+1$th PC, and it has variance λ_{m+1}.

By induction, letting $\{v_1, \cdots, v_k\}$ be a set of orthonormal eigenvectors of Σ that have been ordered so that their eigenvalues $\lambda_1, \cdots, \lambda_k$ satisfy

$$\lambda_1 \geq \cdots \geq \lambda_k > 0,$$

the k PCs of X are given as the linear combinations

$$v_1'X, \cdots, v_k'X.$$

Note that we can find exactly k PCs in this case because Σ is a symmetric and positive definite matrix; by the principal axis theorem, it has an orthonormal eigenbasis. Collecting the coefficients $v_1, \cdots, v_k$ into the $k \times k$ matrix

$$P = \begin{pmatrix} v_1 & \cdots & v_k, \end{pmatrix}$$

we can easily see that

$$P'\Sigma P = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k. \end{pmatrix}$$

Thus, P can be seen as a basis of $\mathbb{R}^k$ that rotates the variables in X so that they become uncorrelated and are arranged in the order of highest variance. In other words, P is a rotation of X that eliminates all the noise among the variables in X and retains the signal contained in them, ordered from the strongest to the weakest signal. Since $\lambda_1, \cdots, \lambda_k$ represent the strength of the signals contained in the first to the last PC, we can say that

$$p_i = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \in (0, 1)$$

represents the proportion of the co-movement among $X_1, \cdots, X_k$ explained by the ith PC. Usually, we reduce the dimension of $X_1, \cdots, X_k$ via PCA by choosing the first m PCs that explain more than 80% of the co-movement among $X_1, \cdots, X_k$.

So far, we have used the population covariance matrix Σ of $X = (X_1, \cdots, X_k)$ to extract the PCs. Since Σ is not known, in practice we use a simple sample analogue of Σ. Suppose we have n observations $x_{1i}, \cdots, x_{ni}$ of variable $1 \leq i \leq k$. Then, collecting these into the

data matrix

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk}, \end{pmatrix}$$

and denoting $x_i = (x_{i1}, \cdots, x_{ik})$, or the value of the variables for the ith observation, we can express

$$\frac{1}{n}X'X = \frac{1}{n}\begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}\begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'.$$

If the sample $\{x_i\}_{1 \le i \le n}$ satisfies certain independence or limited dependence assumptions, as well as distributional assumptions, then the law of large numbers tells us that $\frac{1}{n}X'X$ converges in probability to Σ. As such, we can extract the sample PCs by choosing an orthonormal eigenbasis $\{v_1^{(n)}, \cdots, v_k^{(n)}\}$ of $\frac{1}{n}X'X$ with ordered eigenvalues $\lambda_1^{(n)} \ge \cdots \ge \lambda_k^{(n)}$; we can succinctly express this relationship as

$$\begin{pmatrix} v_1^{(n)\prime} \\ \vdots \\ v_k^{(n)\prime} \end{pmatrix}\left(\frac{1}{n}X'X\right)\begin{pmatrix} v_1^{(n)} & \cdots & v_k^{(n)} \end{pmatrix} = \begin{pmatrix} \lambda_1^{(n)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k^{(n)} \end{pmatrix}.$$

Given these eigenvectors, the rows of the matrix

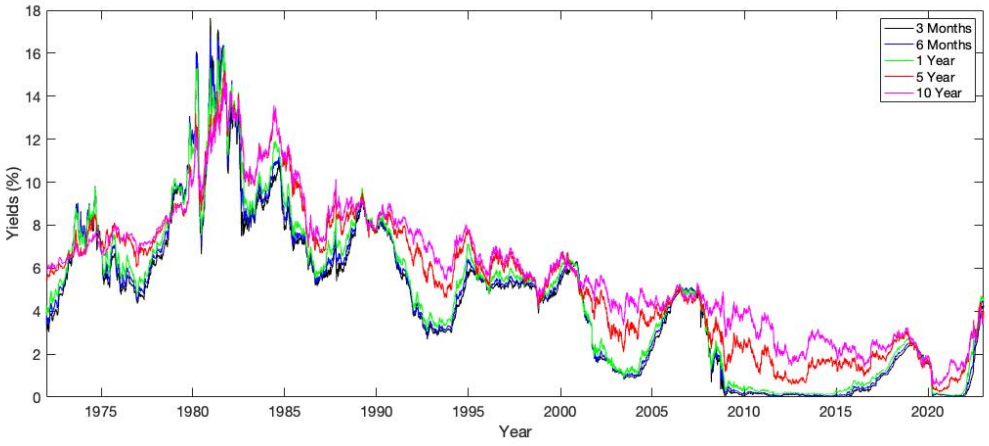
$$X\begin{pmatrix} v_1^{(n)} & \cdots & v_k^{(n)} \end{pmatrix}$$

represents the ith observation of the k sample PCs. Since $\lambda_1^{(n)}, \cdots, \lambda_k^{(n)}$ are the sample variances of each sample PC, we usually normalize the PCs by dividing the ith PC by the square root of $\lambda_i^{(n)}$:

$$\frac{1}{\sqrt{\lambda_1^{(n)}}}Xv_1^{(n)}, \cdots, \frac{1}{\sqrt{\lambda_k^{(n)}}}Xv_k^{(n)}.$$

A final point we need to touch upon is that of identifying the PCs. In general, the orthonormal eigenbasis $\{v_1, \cdots, v_k\}$ of some $k \times k$ positive definite matrix Σ is not unique, since an eigenspace of an eigenvalue with geometric multiplicity greater than 1 may have infinitely many orthonormal bases. Therefore, identification of the PCs in this case requires additional restrictions.

Figure 3.1: **Yields of Various Maturities** The figure below maps daily yields of maturity 3, 6, 12, 60 and 120 months.



A special case is when Σ has k distinct eigenvalues

$$\lambda_1 > \cdots > \lambda_k.$$

In this case, the geometric multiplicity of each eigenvalue is equal to 1. This means that the orthonormal basis of each eigenspace is unique up to sign changes, so that the PCs are now unique up to sign changes. The signs of the PCs in this case are left up to the researcher's discretion.

3.2.2 The Level, Slope and Curvature Factors

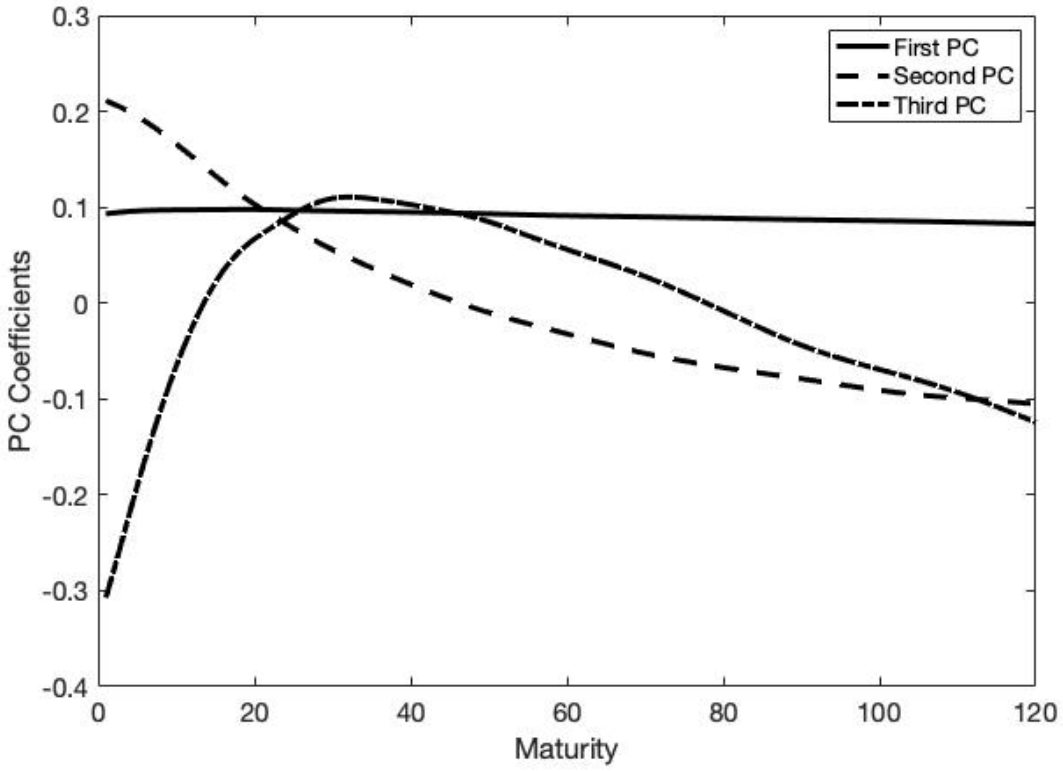
Data on the yield curve is often viewed as a panel of data. Specifically, suppose that we have data on m yields with maturities $1 \leq \tau_1 < \cdots < \tau_m$, from time 1 to T. Then, the data on the yields are collected in the $T \times m$ matrix

$$\mathcal{Y} = \begin{pmatrix} Y_1(\tau_1) & \cdots & Y_1(\tau_m) \\ \vdots & \ddots & \vdots \\ Y_T(\tau_1) & \cdots & Y_T(\tau_m), \end{pmatrix} = \begin{pmatrix} \mathcal{Y}_1' \\ \vdots \\ \mathcal{Y}_T' \end{pmatrix},$$

where

$$\mathcal{Y}_t = \begin{pmatrix} Y_t(\tau_1) \\ \vdots \\ Y_t(\tau_m) \end{pmatrix}$$

Figure 3.2: **Loadings of the First Three PCs** The figure below maps the orthornomal eigenvectors corresponding to the first three PCs.



for each $1 \leq t \leq T$. In most cases, the number of maturities collected in this panel, m, is quite high; therefore, we want to extract the time series of a few factors that explain most of the variation in yields in order to facilitate analysis.

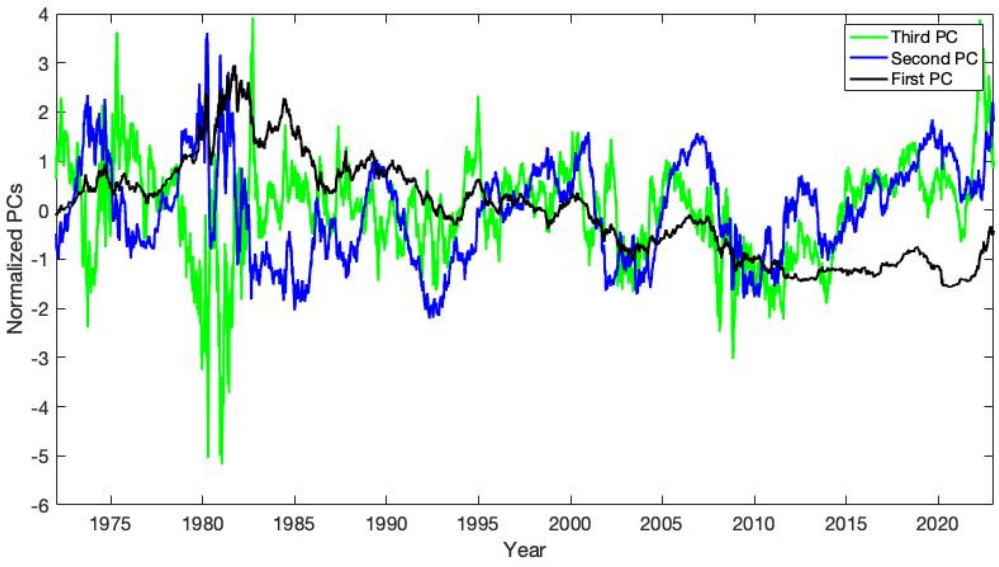
One way to do this is through the use of principal components, as exemplified by Litterman and Scheinkman (1991). In this paper, it is found that the yield curve is special in that most of the co-movement of the yields are captured in the first three PCs of the yield curve. In fact, the first factor predominantly explains the variation in the yield curve. The tables and figures below provide an illustration. Monthly yields of maturities 1 to 120 months obtained from the Liu and Wu database[2] are plotted in Figure 3.1. We can immediately see that yields of longer maturities are often higher than those of shorter

[2]Usually, yield data are constructed by taking yields of maturities 1 year and up from Gurkaynak, Sack, and Wright (2007), and yields of lower maturities from the H-15 release of the Federal Reserve. The reason we take yields of lower maturities from the Fed is because it has been found empirically that yields of short maturities found in GSW are generally unreliable.

Here, we use an alternative data source, namely the LW database of Liu and Wu (2021). This database has an advantage over the traditional GSW database because it offers yields of maturities of every month from 1 to 360, and because the way it constructed the yield data is more stable than that of GSW. However, having been only recently introduced, the LW database is still not as widely used as the GSW database.

In both cases, we take the data in daily frequencies first, and then construct either end-of-month or monthly averages.

Figure 3.3: **First Three PCs** The figure below maps the time series of the first three PCs.



maturities, reflecting the risk associated with holding longer term bonds. Furthermore, following the Great Inflation of the 1970s and early 1980s, we can see that yields exhibit a marked downward trend; the problem of non-stationarity in yields will be dealt with in more depth in later sections. Finally, it can be seen how, during the financial crisis of 2008 and the COVID pandemic, short term yields (3 month and 6 month maturities) were bound by the zero lower bound. Modeling yields at the lower bound is also something that will be addressed in a later section.

Now we extract the principal components of the (demeaned) yield curve. It turns out that the first PC explains 98.47%, the second PC 1.38%, the third PC 0.12%, and the fourth PC 0.02% of the variation in yields, with the remaining $m-4$ PCs not explaining any proportion of this variation. This indicates that three, or maybe four PCs (if we follow Cochrane and Piazzesi (2008)) summarize all the variation in the cross section of the yields.

We study the shape of the coefficients v_1, v_2, v_3 associated with the first three PCs to identify them. Figure 3.2 shows that the coefficients for the first factor are almost uniform across maturities. This means that the first PC acts as sort of an average of yields of various maturities; this is why we call the first PC the **level factor**.

On the other hand, the coefficients for the second factor start high and then fall below 0. This means that yields of lower maturities load more heavily on the second PC than yields of higher maturities; therefore, the second PC represents the short end of the yield curve. It can actually be seen from Figure 3.3 that the shape of the second PC is similar to that of the yield spread, that is, the difference between long and short term yields. The

second PC is thus called the **slope factor**, for its role in capturing how steep the yield curve is.

Finally, the coefficients for the third factor peak at a maturity of around 36 months, or 3 years. This suggests that the third factor represents the middle end of the yield curve, and is called the **curvature factor** because it determines how curved the yield curve is. Since the slope and curvature factors explain the short and middle end of the yield curve, it naturally follows that the level factor represents the long end of the yield curve.

While principal components are a simple and intuitive way of summarizing the information contained in the yield curve, they are still inadequate for our purposes in a number of ways. First, the coefficients of the PCs are dependent on the number of yields m and the sample size T, which means that they contain a degree of real-time instability. This instability proves fatal when estimating term structure models, with models using the PCs as proxies for latent factors proving to be very difficult to estimate. In addition, it is unclear specifically how the three PCs determine the shape of the yield curve, and if they really can be interpreted as representing the long, short and medium end of the yield curve.

An alternative to the principal components approach is detailed starting from the next section. In this alternative approach, we assume from the outset that the yields are determined by three latent factors, and the loadings of these factors are determined so that the three factors always represent the long, short and medium end of the yield curve. The shape of the yield curve is, in addition, determined by a single decay parameter, which helps make the model parsimonious. This approach is that of the famous Nelson-Siegel model.

3.3 The Nelson-Siegel Model

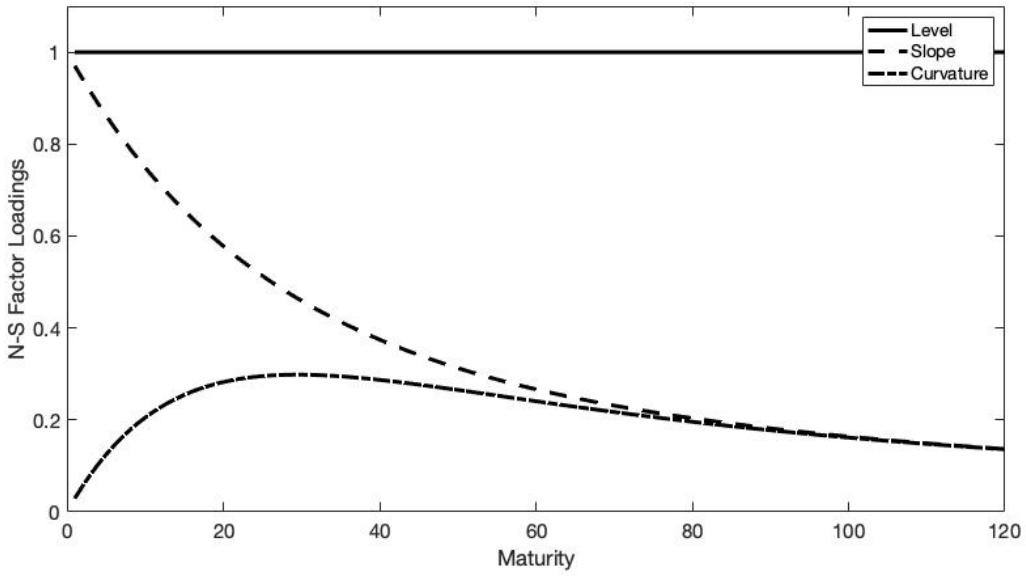
In the **Nelson-Siegel model**, a yield that has τ periods left to maturity at time t is determined as the linear combination of the level factor L_t, the slope factor S_t, and the curvature factor C_t in the following manner:

$$Y_t(\tau) = L_t + \frac{1-\exp(-\tau\kappa)}{\tau\kappa} S_t + \left(\frac{1-\exp(-\tau\kappa)}{\tau\kappa} - \exp(-\tau\kappa) \right) C_t.$$

Here, $\kappa \in (0,1)$ is a decay parameter whose role will be made clear shortly. The factor loadings

$$\beta(\tau;\kappa)' = \begin{pmatrix} 1 & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} - \exp(-\tau\kappa) \end{pmatrix}$$

Figure 3.4: **Nelson-Siegel Factor Loadings** The figure below maps the N-S factor loadings by maturity. The decay parameter is given as $\kappa = 0.0609$.



for the τ-maturity bond are derived by solving a second-order differential equation involving the forward rate; for details, consult Nelson and Siegel (1987).

The Nelson-Siegel factor loadings are displated in Figure 3.4. As with the coefficients associated with the first three yield curve PCs, it is immediately clear why the three factors are referred to as level, slope and curvature; they represent the long, short and middle end of the yield curve, respectively. Unlike the first three PCs, the Nelson-Siegel factors represent the long, short and middle ends of the yield curve by design. To see this, note that the loading of the level factor is equal to 1 for all maturities, and that the mapping

$$\beta_2(\tau;\kappa) = \frac{1-\exp(-\tau\kappa)}{\tau\kappa}$$

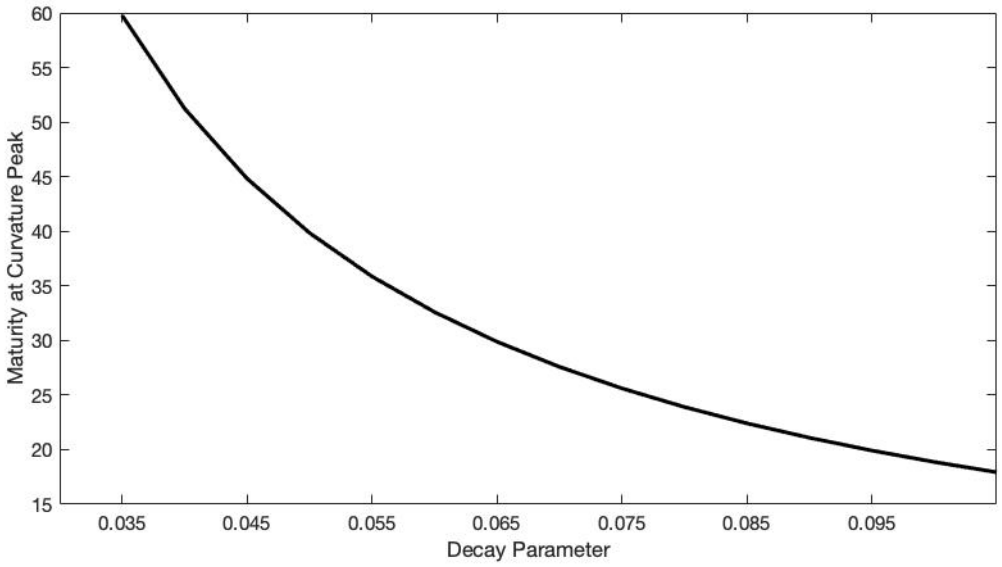
has derivative

$$\begin{aligned}\frac{\partial\beta_2(\tau;\kappa)}{\partial\tau} &= \frac{\exp(-\tau\kappa)}{\tau} - \frac{1-\exp(-\tau\kappa)}{\tau^2\kappa} \\ &= \frac{1}{\tau^2\kappa}\left(\tau\kappa\exp(-\tau\kappa) - 1 + \exp(-\tau\kappa)\right) \\ &= \frac{1}{\tau^2\kappa}\left((\tau\kappa+1)\exp(-\tau\kappa) - 1\right) < 0,\end{aligned}$$

with respect to τ, where the last inequality follows because

$$x + 1 < \exp(x)$$

Figure 3.5: **Peaks of the Curvature Factor Loading** The figure below maps the maturities at which the curvature factor loading peaks for various values of κ.



for any $x > 0$. This revelas that $\beta_2(\tau;\kappa)$ is decreasing in τ for any fixed κ.

As for the third facctor loading, note that

$$\beta_3(\tau;\kappa) = \frac{1-\exp(-\tau\kappa)}{\tau\kappa} - \exp(-\tau\kappa)$$

has derivative

$$\begin{aligned}\frac{\partial\beta_3(\tau;\kappa)}{\partial\tau} &= \frac{1}{\tau^2\kappa}\left((\tau\kappa+1)\exp(-\tau\kappa)-1\right)+\kappa\exp(-\tau\kappa)\\ &= \frac{\exp(-\tau\kappa)}{\tau^2\kappa}\left(\tau^2\kappa^2+\tau\kappa+1\right)-\frac{1}{\tau^2\kappa}\end{aligned}$$

with respect to τ. Let $x^* > 0$ be the non-zero solution to the equation

$$x^2 + x + 1 - \exp(x) = 0.$$

Then, since $x^2+x+1 > \exp(x)$ for any $0 < x < x^*$ and $x^2+x+1 < \exp(x)$ for any $x > x^*$, we can see that $\beta_3(\tau;\kappa)$ is maximized when

$$\tau^*(\kappa) = \frac{x^*}{\kappa}.$$

The larger κ, the shorter the maturity $\tau^*(\kappa)$ at which the curvature factor loading peaks. This tells us that the parameter κ determines where the curvature factor loading peaks and thus which maturity yield best represents the middle end of the yield curve. Figure

3.5 plots $\tau^*(\kappa)$ for values of κ between 0.03 and 0.1. Since yields of maturities bewteen 24 months and 36 months are usually taken to be representative of the middle end of the yield curve, a standard value for the decay parameter presented in Diebold and Li (2006) is 0.0609, which corresponds to a peak of around 30 months.

In short, under the NS model, each factor has a clearly defined role that identifies it against the other factors, and this role is invariant to the sample size given the decay parameter κ. It also turns out that the Nelson-Siegel model fits the yield curve extremely well, with minimal measurement errors. This property will be demonstrated in the upcoming subsections.

3.3.1 Estimating Static Approximate Factor Models

In this section we briefly introduce methods of estimating **static factor models**, or factor models in which the dynamics of the factors are left unspecified. Consider a general macro panel model with T time series observations and N macroeconomic variables. The observation of the ith variable at time t is denoted by x_{it}. In the factor model framework, we assume that there exist $r < \min(T, N)$ common factors whose value at time t is denoted f_t such that the ith variable at time t is determined as a fixed linear combination of the factors at time t plus an error term: formally, we assume that there exists an r-dimensional vector of factor loadings λ_i such that

$$x_{it} = \lambda_i' f_t + e_{it}$$

for any $1 \leq i \leq N$ and $1 \leq t \leq T$. An example of this type of model is the APT introduced in Ross (1976), in which x_{it} represents the risk premium of asset i at time t, λ_i the factor loadings associated with asset i and f_t the common factors at time t.

Factor models such as the APT are examples of **exact factor models**, in which the factors f_t completely explain the cross-sectional and temporal co-movement in the macroeconomic variables, rendering the error term e_{it} as purely idiosyncratic terms. In other words, in exact factor models, the errors e_{it} are uncorrelated across i and t, in addition to being independent of the factors f_t and loadings λ_i. This proves to be very restrictive in practice, so recent papers in the factor model literature have focused on **approximate factor models**, in which limited cross-sectional and temporal correlation among the errors e_{it} is allowed, and they are allowed to be correlated with the factors f_t to an extent.

Below we study the estimation of **large static approximate factor models (LSAFM)**, including the estimation of the factors, factor loadings, and the number of factors. These models are large in that both the cross-sectional dimension N and the time dimension T

are allowed to go to infinity when studying the asymptotic properties of the estimators. The exposition mainly follows Bai and Ng (2002) and Bai (2003).

To study static approximate factor models, we first arrange the data into tractable matrices. One way of organizing the data is to organize it by variable; for any $1 \leq i \leq N$, define

$$\tilde{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix} = \underbrace{\begin{pmatrix} f_1' \\ \vdots \\ f_T' \end{pmatrix}}_{F} \lambda_i + \underbrace{\begin{pmatrix} e_{i1} \\ \vdots \\ e_{iT} \end{pmatrix}}_{\tilde{e}_i}.$$

The data on each variable can then be collected as

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1' \\ \vdots \\ \tilde{x}_N' \end{pmatrix} = \underbrace{\begin{pmatrix} \lambda_1' \\ \vdots \\ \lambda_N' \end{pmatrix}}_{\Lambda} F' + \underbrace{\begin{pmatrix} \tilde{e}_1' \\ \vdots \\ \tilde{e}_N' \end{pmatrix}}_{\tilde{e}}.$$

In summation, we can organize the data in terms of the variables first to obtain

$$\underbrace{\tilde{X}}_{N \times T} = \underbrace{\Lambda}_{N \times r} \cdot \underbrace{F'}_{r \times T} + \underbrace{\tilde{e}}_{N \times T}.$$

An alternative means of organizing the data is to organize it by time. For any $1 \leq t \leq t$, define

$$x_t = \begin{pmatrix} x_{1t} \\ \vdots \\ x_{Nt} \end{pmatrix} = \underbrace{\begin{pmatrix} \lambda_1' \\ \vdots \\ \lambda_N' \end{pmatrix}}_{\Lambda} f_t + \underbrace{\begin{pmatrix} e_{1t} \\ \vdots \\ e_{Nt} \end{pmatrix}}_{e_t},$$

which leads to

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_T' \end{pmatrix} = \underbrace{\begin{pmatrix} f_1' \\ \vdots \\ f_T' \end{pmatrix}}_{F} \Lambda' + \underbrace{\begin{pmatrix} e_1' \\ \vdots \\ e_T' \end{pmatrix}}_{e},$$

so that we are left with

$$\underbrace{X}_{T \times N} = \underbrace{F}_{T \times r} \cdot \underbrace{\Lambda'}_{r \times N} + \underbrace{e}_{T \times N}.$$

Note that $\tilde{X} = X'$.

The **Principal Components (PC) Estimator**, or **Least Squares Estimator** of Λ and F is a non-parametric estimator that is found as the minimizer of the objective function

$$\begin{aligned} V(\Lambda, F) &= \frac{1}{NT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left|x_{it} - \lambda_i' f_t\right|^2 \\ &= \frac{1}{NT}\sum_{t=1}^{T}(x_t - \Lambda f_t)'(x_t - \Lambda f_t) \\ &= \frac{1}{NT}\sum_{i=1}^{N}(\tilde{x}_i - F\lambda_i)'(\tilde{x}_i - F\lambda_i) \\ &= \frac{1}{NT}\operatorname{tr}\left(\left(X - F\Lambda'\right)'\left(X - F\Lambda'\right)\right), \end{aligned}$$

or the sample mean squared error. Note that the organization of the data suggests two methods of minimizing the above objective function; either by concentrating out λ_i or f_t first. We detail each approach below.

1) **Concentrating Out the Factor Loadings**
We proceed in multiple steps.

Step 1: Obtaining the Concentrated Objective Function

The objective function can be expressed as

$$V(\Lambda, F) = \frac{1}{NT}\sum_{i=1}^{N}(\tilde{x}_i - F\lambda_i)'(\tilde{x}_i - F\lambda_i),$$

For any possible value of the factors F, the minimizer of $V(\Lambda, F)$ with respect to λ_i, denoted $\tilde{\lambda}_i(F)$, satisfies the first order conditions

$$F'(\tilde{x}_i - F \cdot \tilde{\lambda}_i(F)) = O_{r\times 1},$$

so that

$$\tilde{\lambda}_i(F) = (F'F)^{-1}F'\tilde{x}_i.$$

Collecting the estimators of $\lambda_1, \cdots, \lambda_N$ into the $N \times r$ matrix

$$\tilde{\Lambda}(F) = \begin{pmatrix} \tilde{\lambda}_1(F)' \\ \vdots \\ \tilde{\lambda}_N(F)' \end{pmatrix} = \begin{pmatrix} \tilde{x}_1' \\ \vdots \\ \tilde{x}_N' \end{pmatrix} F(F'F)^{-1} = X'F(F'F)^{-1},$$

we can see that Λ is estimated as if it were the coefficient in a linear regression in the hypothetical case where F is known.

Substituting this expression back into the objective function, we obtain the concentrated objective function

$$\begin{aligned} \tilde{V}(F) = V(\hat{\Lambda}(F), F) &= \frac{1}{NT} \operatorname{tr}\left(\left(X - F\tilde{\Lambda}(F)'\right)'\left(X - F\tilde{\Lambda}(F)'\right)\right) \\ &= \frac{1}{NT} \operatorname{tr}\left(X' M_F X\right), \end{aligned}$$

where $M_F = I_T - F(F'F)^{-1}F'$ is the residual maker associated with the columns of the $T \times r$ matrix F. Further simplifying the concentrated objective function yields

$$\begin{aligned} \tilde{V}(F) &= \frac{1}{NT} \operatorname{tr}\left(X'X\right) - \frac{1}{NT} \operatorname{tr}\left(X'F(F'F)^{-1}F'X\right) \\ &= \frac{1}{NT} \operatorname{tr}\left(X'X\right) - \frac{1}{NT} \operatorname{tr}\left(F'XX'F(F'F)^{-1}\right). \end{aligned}$$

Step 2: Obtaining Estimates of Factors

Now we impose the identification restriction that $\frac{F'F}{T} = I_r$. This means that the columns of F form a set of r orthonormal vectors, and is imposed to normalize the magnitude and co-dependence among the factors. Our minimization problem can now be written as

$$\begin{aligned} \min_{F \in \mathbb{R}^{T \times r}} \quad & \bar{V}(F) = \frac{1}{NT} \operatorname{tr}\left(X'X\right) - \frac{1}{NT} \operatorname{tr}\left(F'XX'F(F'F)^{-1}\right) \\ \text{subject to} \quad & \frac{F'F}{T} = I_r. \end{aligned}$$

Substituting the constraint into the objective function and noting that $\frac{1}{NT} \operatorname{tr}(X'X)$ does not involve F in any way, we can see that the problem reduces to

$$\begin{aligned} \max_{F \in \mathbb{R}^{T \times r}} \quad & \operatorname{tr}\left(\left(\frac{1}{\sqrt{T}}F\right)' \frac{1}{NT} XX' \left(\frac{1}{\sqrt{T}}F\right)\right) \\ \text{subject to} \quad & \left(\frac{1}{\sqrt{T}}F\right)' \left(\frac{1}{\sqrt{T}}F\right) = I_r. \end{aligned}$$

To solve this minimization problem, we take a brief detour. A useful result in linear algebra reveals that, for any positive semidefinite matrix $M \in \mathbb{R}^{T\times T}$ with ordered eigenvalues $\mu_1 \geq \cdots \geq \mu_T \geq 0$ and a matrix $A \in \mathbb{R}^{T\times k}$ such that $A'A = I_k$, the trace $\operatorname{tr}(A'MA)$ is bounded above as follows[3]:

$$\operatorname{tr}\left(A'MA\right) \leq \sum_{i=1}^{k} \mu_i.$$

This maximum can be attained by letting the columns of A equal orthonormal eigenvectors of M corresponding to the eigenvalues $\mu_1 \geq \cdots \geq \mu_k$.

Letting $\tilde{F}$ be the solution to the above minimization problem, since $\left(\frac{1}{\sqrt{T}}\tilde{F}\right)'\left(\frac{1}{\sqrt{T}}\tilde{F}\right) = I_r$ and $\frac{1}{NT}XX'$ is positive semidefinite, we can see that the columns of $\frac{1}{\sqrt{T}}\tilde{F}$ must equal orthonormal eigenvectors corresponding to the r largest eigenvalues $\mu_1 \geq \cdots \geq \mu_r$ of $\frac{1}{NT}XX'$, and that the maximized value of the objective function is

$$MSE(r) := \tilde{V}(\tilde{F}) = \frac{1}{NT}\operatorname{tr}\left(XX'\right) - \sum_{i=1}^{r}\mu_i.$$

To summarize, the estimators of the factors and factor loadings, as well as the minimized mean squared errors are given as

$$\tilde{F} = \sqrt{T} \times \text{any } r \text{ orthonormal eigenvectors corresponding to the } r \text{ largest eigenvalues } \mu_1 \geq \cdots \geq \mu_r \text{ of } \frac{1}{NT}XX'$$

$$\tilde{\Lambda} = \tilde{\Lambda}(\tilde{F}) = \frac{1}{T}X'\tilde{F}$$

$$MSE(r) = \frac{1}{NT}\operatorname{tr}\left(XX'\right) - \sum_{i=1}^{r}\mu_i.$$

2) **Concentrating Out the Factors**

This alternative approach proceeds similarly to the previous one.

Step 1: Obtaining the Concentrated Objective Function

[3]For a proof, consult my factor model text.

The objective function can be expressed alternatively as

$$V(\Lambda, F) = \frac{1}{NT}\sum_{t=1}^{T}(x_t - \Lambda f_t)'(x_t - \Lambda f_t).$$

For any possible values of Λ, the minimizer of $V(\Lambda, F)$ with respect to f_t, denoted $\overline{f}_t(\Lambda)$, satisfies the first order conditions

$$\Lambda'(x_t - \Lambda \cdot \overline{f}_t(\Lambda)) = O_{r\times 1},$$

so that

$$\overline{f}_t(\Lambda) = (\Lambda'\Lambda)^{-1}\Lambda' x_t.$$

Collecting the factor estimators into the $T \times r$ matrix

$$\overline{F}(\Lambda) = \begin{pmatrix} \overline{f}_1(\Lambda)' \\ \vdots \\ \overline{f}_T(\Lambda)' \end{pmatrix} = X\Lambda\left(\Lambda'\Lambda\right)^{-1},$$

we can see that, this time, the factors are actually estimated as we would the coefficients from a linear regression.

The concentrated objective function is now given as

$$\begin{aligned}\overline{V}(\Lambda) = V(\Lambda, \overline{F}(\Lambda)) &= \frac{1}{NT}\operatorname{tr}\left(\left(X' - \Lambda\overline{F}(\Lambda)'\right)'\left(X' - \Lambda\overline{F}(\Lambda)'\right)\right) \\ &= \frac{1}{NT}\operatorname{tr}\left(XM_\Lambda X'\right),\end{aligned}$$

where $M_\Lambda = I_N - \Lambda(\Lambda'\Lambda)^{-1}\Lambda'$ is the residual maker associated with the columns of Λ. Expanding this expression further yields

$$\overline{V}(\Lambda) = \frac{1}{NT}\operatorname{tr}\left(X'X\right) - \frac{1}{NT}\operatorname{tr}\left(\Lambda'X'X\Lambda(\Lambda'\Lambda)^{-1}\right),$$

which is an expression analogous to the previous case.

Step 2: Obtaining Estimates of Factor Loadings

This time, we impose the identification restriction that $\frac{\Lambda'\Lambda}{N} = I_r$. Our minimization problem can now be written as

$$\max_{\Lambda\in\mathbb{R}^{N\times r}} \quad \frac{1}{NT}\operatorname{tr}\left(\Lambda'X'X\Lambda(\Lambda'\Lambda)^{-1}\right)$$

$$\text{subject to} \quad \frac{\Lambda'\Lambda}{N} = I_r.$$

Substituting the constraint into the objective function changes the problem into

$$\max_{\Lambda \in \mathbb{R}^{N \times r}} \quad \operatorname{tr}\left(\left(\frac{1}{\sqrt{N}}\Lambda\right)' \frac{1}{NT} X'X \left(\frac{1}{\sqrt{N}}\Lambda\right)\right)$$
$$\text{subject to} \quad \left(\frac{1}{\sqrt{N}}\Lambda\right)' \left(\frac{1}{\sqrt{N}}\Lambda\right) = I_r.$$

The matrix inequality shown in the previous section reveals now that the solution $\overline{\Lambda}$ to the above problem, as well as the factor estimates $\overline{F} = \overline{F}(\tilde{\Lambda})$, and the minimized mean squared errors, are given as

$$\overline{\Lambda} = \sqrt{N} \times \text{any } r \text{ orthonormal eigenvectors corresponding to the } r \text{ largest eigenvalues } \mu_1 \geq \cdots \geq \mu_r \text{ of } \frac{1}{NT} X'X$$

$$\overline{F} = \frac{1}{N} X \hat{\Lambda}$$

$$MSE(r) = \frac{1}{NT} \operatorname{tr}\left(X'X\right) - \sum_{i=1}^{r} \mu_i.$$

Note that the eigenvalues $\mu_1 \geq \cdots \geq \mu_r$ here are exactly those of the preceding approach, since the positive semidefinite matrices $\frac{1}{NT} XX'$ and $\frac{1}{NT} X'X$ share the same set of eigenvalues.

Perusing the solution for the factor estimates $\overline{F}$, we can see why this estimation method is called PC estimation. Specifically, $\overline{F}$ are here given as $\frac{1}{\sqrt{N}}$ times the first r PCs of the N variables collected in the columns of X.

Since the first approach requires the computation of the eigenvectors of a $T \times T$ matrix, while the second requires the eigenvectors of an $N \times N$, the first approach is less computationally burdensome if $T < N$, while the second is preferred if $T > N$. We often deal with the case $T > N$, so the factors are often estimated as PCs rather than eigenvectors themselves.

It turns out that the factor estimators $\tilde{F}$ and $\overline{F}$ obtained from either approach are closely related, as shown in Bai and Ng (2002). Suppose that the first r eigenvalues of $\frac{1}{NT} XX'$ (and equivalently, of $\frac{1}{NT} X'X$) are distinct and non-zero, so that the columns of $\frac{1}{\sqrt{T}} \tilde{F}$ and $\frac{1}{\sqrt{N}} \overline{\Lambda}$ are unique up to sign changes. Fixing the signs of each column, we can see that, in this case, the columns of $\frac{1}{\sqrt{T}} \tilde{F}$ and $\frac{1}{\sqrt{N}} \overline{\Lambda}$ are the unique orthonormal eigenvectors of $\frac{1}{NT} XX'$ and $\frac{1}{NT} X'X$ corresponding to $\mu_1 > \cdots > \mu_r > 0$.

Letting V_{NT} be the diagonal matrix with diagonal entries equal to $\mu_1, \cdots, \mu_r$,

$$\left(\frac{1}{NT}X'X\right)\frac{1}{\sqrt{N}}\overline{\Lambda} = \frac{1}{\sqrt{N}}\overline{\Lambda}V_{NT}$$

and

$$\left(\frac{1}{NT}XX'\right)\frac{1}{\sqrt{T}}\tilde{F} = \frac{1}{\sqrt{T}}\tilde{F}V_{NT}$$

by the definition of eigenvectors. Premultiplying both sides of the first equation by $\frac{1}{\sqrt{NT}}X$ shows us that

$$\left(\frac{1}{NT}XX'\right)\frac{1}{\sqrt{T}}\overline{F} = \frac{1}{\sqrt{T}}\overline{F}V_{NT},$$

where

$$\frac{\overline{F}'\overline{F}}{T} = \left(\frac{1}{\sqrt{N}}\overline{\Lambda}\right)'\left(\frac{1}{NT}X'X\right)\left(\frac{1}{\sqrt{N}}\overline{\Lambda}\right) = V_{NT},$$

using $\frac{\overline{\Lambda}'\overline{\Lambda}}{N} = I_r$. By implication,

$$\left(\frac{1}{\sqrt{T}}\overline{F}V_{NT}^{-\frac{1}{2}}\right)'\left(\frac{1}{\sqrt{T}}\overline{F}V_{NT}^{-\frac{1}{2}}\right) = I_r.$$

The uniqueness of $\frac{1}{\sqrt{T}}\tilde{F}$ now tells us that

$$\frac{1}{\sqrt{T}}\tilde{F} = \frac{1}{\sqrt{T}}\overline{F}V_{NT}^{-\frac{1}{2}},$$

or equivalently,

$$\overline{F}V_{NT}^{\frac{1}{2}} = \tilde{F}V_{NT} = \left(\frac{1}{NT}XX'\right)\tilde{F} = \frac{1}{N}X\tilde{\Lambda}.$$

Therefore, we have the relationships

$$\frac{1}{N}X\overline{\Lambda} = \overline{F}$$
$$\frac{1}{N}X\tilde{\Lambda} = \overline{F}\left(\frac{\overline{F}'\overline{F}}{T}\right)^{\frac{1}{2}}.$$

Note that, because the post-multiplication of $\overline{F}$ by $\left(\frac{\overline{F}'\overline{F}}{T}\right)^{\frac{1}{2}}$ is effectively a normalization of the scale of the columns of $\overline{F}$ by the square root of the corresponding eigenvalues, $\frac{1}{N}X\tilde{\Lambda}$ are the normalized PCs studied earlier.

Bai and Ng (2002) shows that the normalized PCs

$$\hat{F} = \frac{1}{N} X\tilde{\Lambda} = \overline{F}\left(\frac{\overline{F}'\overline{F}}{T}\right)^{\frac{1}{2}}$$

consistently estimate a rotation of the factors as both the cross-sectional and time dimensions go to infinity:

$$\frac{1}{T}\sum_{t=1}^{T}\left|\hat{F}_t - H'F_t^0\right|^2 = O_p\left(\frac{1}{\min(N,T)}\right)$$
$$\left|\hat{F}_t - H'F_t^0\right|^2 = O_p\left(\frac{1}{\min(N,T)}\right) \quad \text{for any } t \in N_+$$

as $N,T \to \infty$, where $\|\cdot\|$ is the trace norm, F^0 is the true value of the factors F, and

$$H = \left(\frac{\Lambda^{0\prime}\Lambda^0}{N}\right)\left(\frac{F^{0\prime}\tilde{F}}{T}\right)$$

is an $O_p(1)$ rotation. This shows us that even though $\hat{F}_t$ does not converge to the true factors themselves, it does converge to a rotation of the true factors, and that the speed of convergence is $\min(\sqrt{N},\sqrt{T})$. This is a result that can provide a theoretical basis for the use of principal components to summarize the co-movement among different variables. The asymptotic normality of the PC estimators under additional assumptions is shown in Bai (2003).

Another area of interest is the estimation of the number of factors r. Let $MSE(k)$ be the minimum mean squared error under the assumption of k factors. We saw above that

$$MSE(k) = \frac{1}{NT}\operatorname{tr}\left(XX'\right) - \sum_{i=1}^{k}\mu_i,$$

where $\mu_1 \geq \cdots \geq \mu_k$ are the k largest eigenvalues of the positive semidefinite matrix $\frac{1}{NT}XX'$. By positive semidefiniteness, the eigenvalues of XX' are all non-negative, and as such the greater the number of factors k, the smaller the minimum mean squared error $MSE(k)$. This indicates that $MSE(k)$ can serve a role similar to the maximized log-likelihood in information criteria such as the AIC and BIC.

Various criteria for determining the number of factors based on $MSE(k)$ have been introduced. One such criterion is the **eigenvalue ratio (ER) test** introduced in S. C.

Ahn and Horenstein (2013), which proposes to choose the number of factors as

$$r_{max} = \underset{1 \leq k \leq k_{max}-1}{\operatorname{argmax}} \quad \frac{\mu_k}{\mu_{k+1}}.$$

This is inspired by the intuition that the eigenvalue μ_{k+1} represents the amount of information added (the amount by which the MSE falls) by the inclusion of the $k+1$th factor. Intuitively, the true number of factors must be the number k such that μ_k is large, so that the kth factor contributes meaningfully to explaining the co-movement among the macro variables, but μ_{k+1} is low, so that additional factors do not contribute much to explaining this co-movement.

Another popular criterion for the determination of the number of factors is the **panel information criterion (PIC)** introduced in Bai and Ng (2002). The number of factors under the PIC is chosen as the minimizer of

$$PIC(k) = \log(MSE(k)) + k \cdot g(N,T)$$

across $\{1, \cdots, k_{max}\}$, where $g(N,T)$ is a penalty function such that

$$g(N,T) \to 0 \quad \text{and} \quad \min(N,T) \cdot g(N,T) \to +\infty$$

as $N,T \to \infty$. It is shown in Bai and Ng (2002) that using the PIC allows for the consistent estimation of the true number of factors r. A popular penalty function is

$$g(N,T) = \frac{N+T}{NT} \log\left(\frac{NT}{N+T}\right).$$

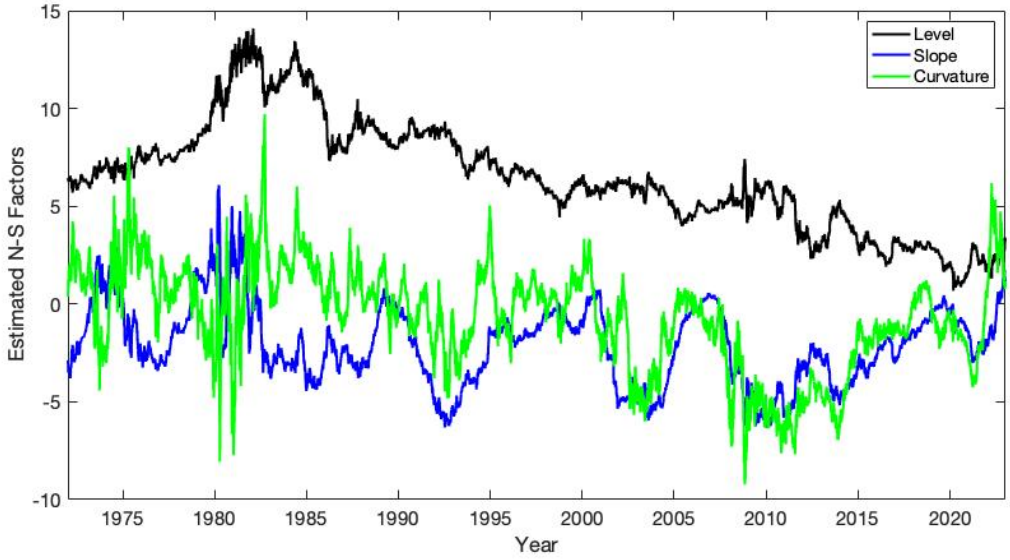
3.3.2 Estimating the Nelson-Siegel Model

The Nelson-Siegel model can be estimated similarly to LSAFM. Given a large enough selection of maturities m and periods T, the panel of yields $\mathcal{Y}$ certainly quantifies as large. Furthermore, the yields of time t can be written as

$$\mathcal{Y}_t = \Lambda(\kappa) f_t + e_t,$$

where e_t contains the measurement errors at time t, and f_t is a 3×1 vector comprising the level L_t, slope S_t, and curvature C_t. The factor loadings are determined by the decay

Figure 3.6: **Estimated Nelson-Siegel Factors** The figure below maps the estimated Nelson-Siegel factors, given daily yield curve data from 1972 to 2023 and maturities 1 to 120 months.



parameter κ as

$$\Lambda(\kappa) = \begin{pmatrix} 1 & \frac{1-\exp(-\tau_1\kappa)}{\tau_1\kappa} & \frac{1-\exp(-\tau_1\kappa)}{\tau_1\kappa} - \exp(-\tau_1\kappa) \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-\exp(-\tau_m\kappa)}{\tau_m\kappa} & \frac{1-\exp(-\tau_m\kappa)}{\tau_m\kappa} - \exp(-\tau_m\kappa). \end{pmatrix}$$

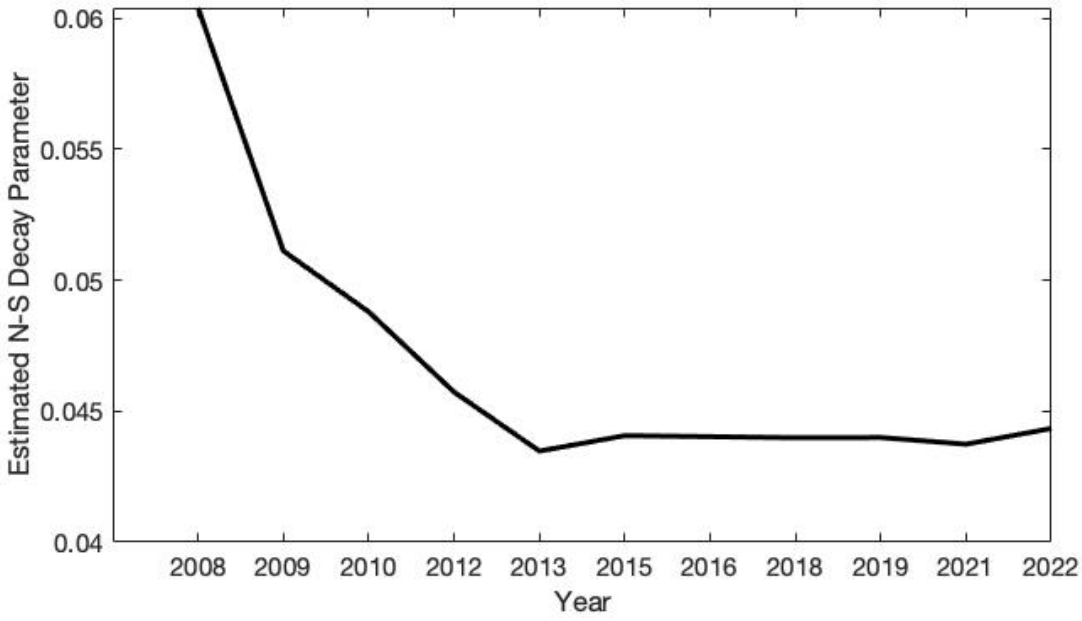
Estimation of the model can be done non-parametrically, as in the previous section. The only difference is that, instead of the factor loadings being unrestricted $m \times 3$ matrices, they depend only on a single parameter κ. This means that it is convenient to first concentrate out the factors, and then obtain an estimator of κ. The mean squared error is given as

$$V(\kappa, F) = \frac{1}{mT} \sum_{t=1}^{T} (\mathcal{Y}_t - \Lambda(\kappa) f_t)'(\mathcal{Y}_t - \Lambda(\kappa) f_t).$$

For any value of the decay parameter κ, the minimizer of $V(\kappa, F)$ with respect to f_t, denoted $\overline{f}_t(\kappa)$, is given as

$$\overline{f}_t(\kappa) = \left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1} \Lambda(\kappa)'\mathcal{Y}_t,$$

Figure 3.7: **Estimated Nelson-Siegel Decay Parameter** The figure below maps the N-S decay parameter estimated using samples with endpoints from 2008 to 2022.



and these are collected as

$$\overline{F}(\kappa) = \begin{pmatrix} \overline{f}_1(\kappa)' \\ \vdots \\ \overline{f}_T(\kappa)' \end{pmatrix} = \mathcal{Y}\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}.$$

The concentrated objective function is

$$\begin{aligned} \overline{V}(\kappa) = V(\kappa, \overline{F}(\kappa)) &= \frac{1}{mT}\operatorname{tr}\left(\left(\mathcal{Y}' - \Lambda(\kappa)\overline{F}(\kappa)\right)'\left(\mathcal{Y}' - \Lambda(\kappa)\overline{F}(\kappa)\right)\right) \\ &= \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}'\mathcal{Y}\right) - \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\mathcal{Y}'\right). \end{aligned}$$

We find our estimator of κ as the solution to the following problem:

$$\max_{\kappa\in[\epsilon,1-\epsilon]} \quad \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\mathcal{Y}'\right).$$

Here, $\epsilon > 0$ is a small positive value that ensures the above problem has a solution $\overline{\kappa}$. Since a value of the decay parameter below 0.02 indicates that yields of maturity 10 years or longer represent the middle end of the yield curve, which is not really plausible, in practice we let $\epsilon = 0.02$. The factor estimators are then given as

$$\overline{F} = \mathcal{Y}\Lambda(\overline{\kappa})\left[\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right]^{-1}.$$

We can show that, under certain assumptions, the estimators above are consistent for the true factors as both the cross-sectional and time dimensions go to infinity. For details, consult the appendix.

The estimated value of the decay parameter is 0.0447, which corresponds to a peak of roughly 40 months, which is slightly on the higher side. The estimated factors are displayed in Figure 3.6. We can see that the shape of the three factors are similar to those of the first three PCs, which is unsurprising given that they play the same role in the N-S model as in the naive PC model. One difference is that the level factor is now very high; this reflects the fact that information on the mean of the yields is contained in the level factor.

The real-time estimates of the decay parameter every year from 2008 to 2022 are given in Figure 3.7. There are significant fluctuations in the estimate, with the maturity that represents the middle end of the yield curve increasing over time.

So far, the N-S model that we have studied does not specify the dynamics of the factors, instead relying only on the relationship between the factors and the yields. In the next section, we study how specifying the factor dynamics enrichens the model, and how to incorporate the additional information on the factor dynamics in the estimation process.

3.4 The Dynamic Nelson-Siegel Model

The **Dynamic Nelson-Siegel (DN-S) model** was developed by Diebold and Li (2006) as a special version of the N-S model where the dynamics of the factors is specified. As before, let the sample comprise yields of m maturities $1 \leq \tau_1 < \cdots < \tau_m$, and suppose that a yield of τ maturities depends on the level, slope and curvature factors L_t, S_t and C_t in the following manner:

$$Y_t(\tau) = \underbrace{\begin{pmatrix} 1 & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} - \exp(-\tau\kappa) \end{pmatrix}}_{\beta(\tau;\kappa)'} \begin{pmatrix} L_t \\ S_t \\ C_t \end{pmatrix}.$$

This is simply the specification of the static N-S model. In contrast, the standard DN-S model is formulated in the following state-space form, with an explicit measurement error e_t and the assumption that the factors follow a VAR(1) specification[4].

$$\mathcal{Y}_t = \Lambda(\kappa) f_t + \Sigma e_t \qquad \text{(Measurement Equation)}$$

$$f_t = c + G f_{t-1} + H u_t, \qquad \text{(Transition Equation)}$$

[4]As usual, we can assume that the VAR lag order is greater than 1, but this does not alter the model in any major way because we need only choose the companion form of the model in this case

where $\mathcal{Y}_t$ collects the sample yields at time t and $\Lambda(\kappa)$ the factor loadings. Here, Σe_t is a vector of mean 0 time t measurement error terms with variance $\Sigma\Sigma'$, and Hu_t a vector of mean 0 factor innovations with variance HH'. We allow the dimension of u_t is allowed to be smaller than the number of factors f_t, to accomodate more general cases, including those where the lag order is greater than 1.

Below we study generic methods of estimating dynamic factor models, or state-space models. We focus on two methods: a two-step method devised by Doz, Giannone, and Reichlin (2011) based on the PC estimators of the static factor model, and a QMLE-EM algorithm method, proposed in Doz, Giannone, and Reichlin (2012) and refined in Barigozzi and Luciani (2019).

3.4.1 Estimating Small Dynamic Factor Models

As in the static factor model case, consider panel data of N macroeconomic variables and T time series observations. The time t observation of the ith variable is, as before, denoted by x_{it}, and the time t variables are collected into the vector

$$x_t = \begin{pmatrix} x_{1t} \\ \vdots \\ x_{Nt} \end{pmatrix}.$$

Suppose there are r factors that explain the co-movement of the N macroeconomic variables, whose time t values are collected in the random vector f_t.

A generic **dynamic factor model (DFM)** is given in the following state space model form:

$$x_t = \underbrace{\Lambda}_{N\times r} f_t + \underbrace{\Sigma}_{N\times N} e_t \qquad \text{(Measurement Equation)}$$

$$f_t = c + Gf_{t-1} + \underbrace{H}_{r\times q} u_t, \qquad \text{(Transition Equation)}$$

where e_t is an N-dimensional random vector with mean 0 and variance I_N representing the idiosyncratic errors, and u_t is a $q \leq r$ dimensional random vector with mean 0 and variance I_q representing the factor innovations. We allow the dimension of u_t to be smaller than r to accomodate more general cases that arise in the literature, such as the case where the lag order of the transition equation is greater than 1.

In **small dynamic factor models (SDFM)**, that is, models in which N is small and fixed, while T goes to infinity, it is customary to estimate the model via Gaussian Quasi-Maximum Likelihood Estimation (QMLE). In other words, we find the values of the parameters that maximize the likelihood derived under the assumption that the idiosyn-

cratic errors and factor innovations $\{e_t\}_{t\in\mathbb{Z}}$ and $\{u_t\}_{t\in\mathbb{Z}}$ are i.i.d. Gaussian. If the true idiosyncratic errors and factor innovations are not i.i.d. Gaussian, then the log-likelihood derived under the i.i.d. Gaussian assumption is an approximation to the true likelihood, which is why this method is, strictly speaking, only "Quasi"-MLE.

The Kalman Filter

To derive the Gaussian log-likelihood, we make use of the **Kalman filter**. The filter allows us, under the assumption of Gaussian errors, to derive the best real-time estimates of the factors f_t. That is, it yields closed-form recursions for the conditional expectations

$$\mathbb{E}\left[f_t \mid x_t, \cdots, x_1\right],$$

which is the point estimate of f_t that minimizes the mean squared error, provided that we have information on the macro variables up to time t.

To compute the Kalman filter, let

$$\theta = \{\Lambda, \Sigma, c, G, H\}$$

be the vector collecting the model parameters. First, we define the following quantities:

$$\begin{aligned}
\mathcal{F}_t &= \sigma\{x_t, \cdots, x_1\}, \quad \text{or the information contained up to the } t\text{th sample period} \\
f_{t|t-1}(\theta) &= \mathbb{E}\left[f_t \mid \mathcal{F}_{t-1}, \theta\right] \\
f_{t|t}(\theta) &= \mathbb{E}\left[f_t \mid \mathcal{F}_t, \theta\right] \\
x_{t|t-1}(\theta) &= \mathbb{E}\left[x_t \mid \mathcal{F}_{t-1}, \theta\right] \\
P_{t|t-1}(\theta) &= \mathrm{Var}\left(f_t \mid \mathcal{F}_{t-1}, \theta\right) = \mathbb{E}\left[(f_t - f_{t|t-1}(\theta))(f_t - f_{t|t-1}(\theta))' \mid \mathcal{F}_{t-1}, \theta\right] \\
P_{t|t}(\theta) &= \mathrm{Var}\left(f_t \mid \mathcal{F}_t, \theta\right) = \mathbb{E}\left[(f_t - f_{t|t}(\theta))(f_t - f_{t|t}(\theta))' \mid \mathcal{F}_t, \theta\right] \\
V_{t|t-1}(\theta) &= \mathrm{Var}\left(x_t \mid \mathcal{F}_{t-1}, \theta\right) = \mathbb{E}\left[(x_t - x_{t|t-1}(\theta))(x_t - x_{t|t-1}(\theta))' \mid \mathcal{F}_{t-1}, \theta\right].
\end{aligned}$$

The filter is initialized as follows:

$$f_0 \mid \theta \sim \mathcal{N}\left[f_{0|0}(\theta), P_{0|0}(\theta)\right],$$

where $f_{0|0}(\theta)$ and $P_{0|0}(\theta)$ are the unconditional mean and variance of the initial factor f_0[5]. Due to the assumption that $\{e_t\}_{t\in\mathbb{Z}}$ and $\{u_t\}_{t\in\mathbb{Z}}$ are i.i.d. Gaussian, we can also claim

[5]If the factors are non-stationary, then neither the unconditional mean nor the uncondition variance of f_0 exists. In this case, $f_{0|0}$ and $P_{0|0}$ are interpreted as initial values for the non-stationary process $\{f_t\}_{t\in\mathbb{N}}$, and they are chosen accordingly during estimation. The choice of initial values will be studied

that

$$\begin{pmatrix} f_0 \\ e_1 \\ u_1 \end{pmatrix} \mid \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{0|0}(\theta) \\ O_{(N+q)\times 1} \end{pmatrix}, \quad \operatorname{diag}\left(P_{0|0}(\theta), I_{N+q}\right) \right].$$

Now suppose, for some $1 \leq t \leq T$, that we have obtained $f_{t-1|t-1}(\theta)$ and $P_{t-1|t-1}(\theta)$, and that

$$\begin{pmatrix} f_{t-1} \\ e_t \\ u_t \end{pmatrix} \mid \mathcal{F}_{t-1}, \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{t-1|t-1}(\theta) \\ O_{(N+q)\times 1} \end{pmatrix}, \quad \operatorname{diag}\left(P_{t-1|t-1}(\theta), I_N, I_q\right) \right].$$

We can see that

$$\begin{aligned} f_{t|t-1}(\theta) &= \mathbb{E}\left[f_t \mid \mathcal{F}_{t-1}, \theta\right] = \mathbb{E}\left[c + G f_{t-1} + H u_t \mid \mathcal{F}_{t-1}, \theta\right] \\ &= c + G f_{t-1|t-1}(\theta) \end{aligned}$$

$$\begin{aligned} P_{t|t-1}(\theta) &= \mathbb{E}\left[(f_t - f_{t|t-1}(\theta))(f_t - f_{t|t-1}(\theta))' \mid \mathcal{F}_{t-1}, \theta\right] \\ &= \mathbb{E}\left[\left(G(f_{t-1} - f_{t-1|t-1}(\theta)) + H u_t\right)\left(G(f_{t-1} - f_{t-1|t-1}(\theta)) + H u_t\right)' \mid \mathcal{F}_{t-1}, \theta\right] \\ &= G P_{t-1|t-1}(\theta) G' + H H'. \end{aligned}$$

Since

$$\begin{aligned} \begin{pmatrix} f_t \\ e_t \end{pmatrix} &= \begin{pmatrix} f_{t|t-1}(\theta) \\ O_{N\times 1} \end{pmatrix} + \begin{pmatrix} G(f_{t-1} - f_{t-1|t-1}(\theta)) + H u_t \\ e_t \end{pmatrix} \\ &= \begin{pmatrix} f_{t|t-1}(\theta) \\ O_{N\times 1} \end{pmatrix} + \begin{pmatrix} G & O_{r\times N} & H \\ O_{N\times r} & I_N & O_{N\times q} \end{pmatrix} \begin{pmatrix} f_{t-1} - f_{t-1|t-1}(\theta) \\ e_t \\ u_t \end{pmatrix}, \end{aligned}$$

we have

$$\begin{pmatrix} f_t \\ e_t \end{pmatrix} \mid \mathcal{F}_{t-1}, \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{t|t-1}(\theta) \\ O_{N\times 1} \end{pmatrix}, \quad \operatorname{diag}\left(P_{t|t-1}(\theta), I_N\right) \right].$$

In addition, the measurement equation

$$x_t = \Lambda f_t + \Sigma e_t$$

in more detail during Professor Kang's lecture.

implies that

$$
\begin{aligned}
x_{t|t-1}(\theta) &= \mathbb{E}\left[x_t \mid \mathcal{F}_{t-1}, \theta\right] = \Lambda f_{t|t-1}(\theta) \\
V_{t|t-1}(\theta) &= \mathbb{E}\left[(x_t - x_{t|t-1}(\theta))(x_t - x_{t|t-1}(\theta))' \mid \mathcal{F}_{t-1}, \theta\right] \\
&= \mathbb{E}\left[\left(\Lambda(f_t - f_{t|t-1}(\theta)) + \Sigma e_t\right)\left(\Lambda(f_t - f_{t|t-1}(\theta)) + \Sigma e_t\right)' \mid \mathcal{F}_{t-1}, \theta\right] \\
&= \Lambda P_{t|t-1}(\theta)\Lambda' + \Sigma\Sigma'.
\end{aligned}
$$

Finally, since

$$
\begin{aligned}
\begin{pmatrix} f_t \\ x_t \end{pmatrix} &= \begin{pmatrix} f_{t|t-1}(\theta) \\ x_{t|t-1}(\theta) \end{pmatrix} + \begin{pmatrix} f_t - f_{t|t-1}(\theta) \\ \Lambda\left(f_t - f_{t|t-1}(\theta)\right) + \Sigma e_t \end{pmatrix} \\
&= \begin{pmatrix} f_{t|t-1}(\theta) \\ x_{t|t-1}(\theta) \end{pmatrix} + \begin{pmatrix} I_r & O_{r\times N} \\ \Lambda & \Sigma \end{pmatrix} \begin{pmatrix} f_t - f_{t|t-1}(\theta) \\ e_t \end{pmatrix},
\end{aligned}
$$

we have

$$
\begin{pmatrix} f_t \\ x_t \end{pmatrix} \mid \mathcal{F}_{t-1}, \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{t|t-1}(\theta) \\ x_{t|t-1}(\theta) \end{pmatrix}, \quad \begin{pmatrix} P_{t|t-1}(\theta) & P_{t|t-1}(\theta)\Lambda' \\ \Lambda P_{t|t-1}(\theta) & V_{t|t-1}(\theta) \end{pmatrix}\right].
$$

By the updating formula for jointly normally distributed random variables[6],

$$
f_t \mid \mathcal{F}_t, \theta \sim f_t \mid x_t, \mathcal{F}_{t-1}, \theta \sim \mathcal{N}\left[f_{t|t}(\theta), P_{t|t}(\theta)\right],
$$

where

$$
\begin{aligned}
f_{t|t}(\theta) &= f_{t|t-1}(\theta) + P_{t|t-1}(\theta)\Lambda' V_{t|t-1}(\theta)^{-1}\left(x_t - x_{t|t-1}(\theta)\right) \\
P_{t|t}(\theta) &= P_{t|t-1}(\theta) - P_{t|t-1}(\theta)\Lambda' V_{t|t-1}(\theta)^{-1}\Lambda P_{t|t-1}(\theta) \\
&= \left[I_N - P_{t|t-1}(\theta)\Lambda' V_{t|t-1}(\theta)^{-1}\Lambda\right] P_{t|t-1}(\theta).
\end{aligned}
$$

The **Kalman gain** is defined as

$$
K_{t|t-1}(\theta) = P_{t|t-1}(\theta)\Lambda' V_{t|t-1}(\theta)^{-1},
$$

using which we can write the above quantities as

$$
f_{t|t}(\theta) = f_{t|t-1}(\theta) + K_{t|t-1}(\theta)\left(x_t - x_{t|t-1}(\theta)\right)
$$

[6]For reference, see the section "Conditional Distributions" in `https://en.wikipedia.org/wiki/Multivariate_normal_distribution`.

$$P_{t|t}(\theta) = \left[I_N - K_{t|t-1}(\theta)\Lambda\right] P_{t|t-1}(\theta).$$

Heuristically, the Kalman gain represents how much of the information on the time t data x_t to use when updating our estimate of the factors f_t from $f_{t|t-1}$ to $f_{t|t}$.

To complete our derivation, note that, because $\mathcal{F}_t$ and $f_t - f_{t|t}(\theta)$ are independent of e_{t+1} and u_{t+1}, $f_t - f_{t|t}(\theta)$ is conditionally independent of e_{t+1} and u_{t+1} given $\mathcal{F}_t$[7]. Furthermore, they are all Gaussian given $\mathcal{F}_t$, so

$$\begin{pmatrix} f_t \\ e_{t+1} \\ u_{t+1} \end{pmatrix} \mid \mathcal{F}_t, \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{t|t}(\theta) \\ O_{(N+q)\times 1} \end{pmatrix}, \quad \operatorname{diag}\left(P_{t|t}(\theta), I_{N+q}\right)\right].$$

By induction, the Kalman filtered values are given as follows:

$$f_{t|t-1}(\theta) = c + G f_{t-1|t-1}(\theta) \tag{3.1}$$

$$P_{t|t-1}(\theta) = G P_{t-1|t-1}(\theta) G' + HH' \tag{3.2}$$

$$x_{t|t-1}(\theta) = \Lambda f_{t|t-1}(\theta) \tag{3.3}$$

$$V_{t|t-1}(\theta) = \Lambda P_{t|t-1}(\theta)\Lambda' + \Sigma\Sigma' \tag{3.4}$$

$$K_{t|t-1}(\theta) = P_{t|t-1}(\theta)\Lambda' V_{t|t-1}(\theta)^{-1} \tag{3.5}$$

$$f_{t|t}(\theta) = f_{t|t-1}(\theta) + K_{t|t-1}(\theta)\left(x_t - x_{t|t-1}(\theta)\right) \tag{3.6}$$

$$P_{t|t}(\theta) = \left[I_r - K_{t|t-1}(\theta)\Lambda\right] P_{t|t-1}(\theta). \tag{3.7}$$

Note that, while $P_{t|t-1}(\theta)$ is likely to be singular when $q < r$ due to the singularity of HH', $V_{t|t-1}$ remains non-singular because it is the sum of a positive semidefinite matrix $\Lambda P_{t|t-1}(\theta)\Lambda'$ and a positive definite matrix $\Sigma\Sigma'$. Thus, the inverse matrix that appears in the Kalman gain exists regardless of whether $q < r$ or not.

Using the values above, we can compute the Gaussian Quasi-log likelihood. The log-likelihood can be decomposed as

$$l(x_T, \cdots, x_1 \mid \theta) = \sum_{t=1}^{T} \log\left(f(x_t \mid \mathcal{F}_{t-1}, \theta)\right)$$

where $f(x_t \mid \mathcal{F}_{t-1}, \theta)$ is the density of x_t given $\mathcal{F}_{t-1}$ and θ. For any $1 \le t \le T$, since

$$x_t = \Lambda f_t + \Sigma e_t,$$

[7] If you want to know the reason why, consult me.

where (f_t, e_t) are independent and jointly Gaussian given $\mathcal{F}_{t-1}$,

$$x_t \mid \mathcal{F}_{t-1}, \theta \sim \mathcal{N}\left[x_{t|t-1}(\theta), V_{t|t-1}(\theta)\right],$$

and

$$\log\left(f(x_t \mid \mathcal{F}_{t-1}, \theta)\right) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\left|V_{t|t-1}(\theta)\right| - \frac{1}{2}(x_t - x_{t|t-1}(\theta))' V_{t|t-1}(\theta)^{-1}(x_t - x_{t|t-1}(\theta)).$$

Therefore,

$$\begin{aligned} l(x_T, \cdots, x_1 \mid \theta) = &-\frac{NT}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log\left|V_{t|t-1}(\theta)\right| \\ &- \frac{1}{2}\sum_{t=1}^{T}(x_t - x_{t|t-1}(\theta))' V_{t|t-1}(\theta)^{-1}(x_t - x_{t|t-1}(\theta)). \end{aligned}$$

The Gaussian QMLE of θ is now found by maximizing this Gaussian Quasi log likelihood with respect to θ. However, in the current form $l(x_T, \cdots, x_1 \mid \theta)$ is a very complicated function of θ, which makes numerical maximization intractable. In order to find the maximizer in a more stable manner, we must turn to the EM algorithm, for which we require the Kalman smoother. This will be the next topic we focus on.

The Kalman Smoother

If the Kalman filter allowed us to obtain expressions for the best (mean squared error minimizing) estimate of the time t factor f_t given the information up to time t, the **Kalman smoother** utilizes the information available in the entire sample to estimate f_t. Consequently, we are able to obtain more precise estimates of f_t.

As above, we start by defining the following terms:

$$\begin{aligned} f_{t|T}(\theta) &= \mathbb{E}\left[f_t \mid \mathcal{F}_T, \theta\right] \\ P_{t|T}(\theta) &= \mathrm{Var}\left(f_t \mid \mathcal{F}_T, \theta\right) = \mathbb{E}\left[(f_t - f_{t|T}(\theta))(f_t - f_{t|T}(\theta))' \mid \mathcal{F}_T, \theta\right], \end{aligned}$$

where $\mathcal{F}_T$ represents the information present in the entire sample. As with the Kalman filter, the values of $f_{t|T}(\theta)$ and $P_{t|T}(\theta)$ are also obtained recursively, this time starting from $f_{T|T}(\theta)$, $P_{T|T}(\theta)$ and moving backward in time. For this reason, the Kalman filtering process is often called the **forward pass** and the smoothing process the **backward pass**.

Note that the values of $f_{T|T}(\theta)$ and $P_{T|T}(\theta)$ have already been computed as the last step of the forward pass. Suppose that, for some $0 \leq t < T$, we have $f_{t+1|T}(\theta)$ and $P_{t+1|T}(\theta)$.

We want to compute the quantities

$$f_{t|T}(\theta) \quad \text{and} \quad P_{t|T}(\theta).$$

We address two cases: the case when $q = r$, so that $P_{t+1|t}(\theta)$ is invertible, and $q < r$, where $P_{t+1|t}(\theta)$ is most likely singular.

i) **The Non-singular Case**

In this case, the derivation is relatively straightforward. Note that

$$\begin{aligned}\begin{pmatrix} f_t \\ f_{t+1} \end{pmatrix} &= \begin{pmatrix} f_{t|t}(\theta) \\ c + G f_{t|t}(\theta) \end{pmatrix} + \begin{pmatrix} f_t - f_{t|t}(\theta) \\ G\left(f_t - f_{t|t}(\theta)\right) + H u_{t+1} \end{pmatrix} \\ &= \begin{pmatrix} f_{t|t}(\theta) \\ f_{t+1|t}(\theta) \end{pmatrix} + \begin{pmatrix} I_r & O_{r\times q} \\ G & H \end{pmatrix} \begin{pmatrix} f_t - f_{t|t}(\theta) \\ u_{t+1} \end{pmatrix},\end{aligned}$$

so that

$$\begin{pmatrix} f_t \\ f_{t+1} \end{pmatrix} \mid \mathcal{F}_t, \theta \sim \mathcal{N}\left[\begin{pmatrix} f_{t|t}(\theta) \\ f_{t+1|t}(\theta) \end{pmatrix}, \begin{pmatrix} P_{t|t}(\theta) & P_{t|t}(\theta)G' \\ G P_{t|t}(\theta) & P_{t+1|t}(\theta) \end{pmatrix} \right].$$

In addition, we can see that

$$x_{t+1} = \Lambda f_{t+1} + \Sigma e_{t+1}$$

is determined by f_{t+1} and e_{t+1}. Likewise,

$$\begin{aligned} x_{t+2} &= \Lambda f_{t+2} + \Sigma e_{t+2} \\ &= \Lambda c + \Lambda G f_{t+1} + \Lambda H u_{t+2} + \Sigma e_{t+2} \end{aligned}$$

is determined by f_{t+1} and the error terms u_{t+2} and e_{t+2}. In other words, the information set $\mathcal{F}_T$ is contained in the information set

$$\sigma\{\mathcal{F}_t, f_{t+1}, e_{t+1}, \cdots, e_T, u_{t+1}, \cdots, u_T\}.$$

The updating formula for jointly normal variables tells us that

$$f_t \mid f_{t+1}, \mathcal{F}_t, \theta \sim \mathcal{N}\left[f_{t|t+1}(\theta), P_{t|t+1}(\theta)\right],$$

where

$$f_{t|t+1}(\theta) = f_{t|t}(\theta) + P_{t|t}(\theta) G' P_{t+1|t}(\theta)^{-1} \left(f_{t+1} - f_{t+1|t}(\theta)\right)$$

$$P_{t|t+1}(\theta) = P_{t|t}(\theta) - P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta).$$

Letting E_{t+1} be defined as

$$E_{t+1} = \sigma\{e_{t+1}, \cdots, e_T, u_{t+1}, \cdots, u_T\},$$

the observation about the information sets above tells us that

$$\begin{aligned}
f_{t|T}(\theta) &= \mathbb{E}\left[f_t \mid \mathcal{F}_T, \theta\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[f_t \mid E_{t+1}, f_{t+1}, \mathcal{F}_t, \theta\right] \mid \mathcal{F}_T, \theta\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[f_t \mid f_{t+1}, \mathcal{F}_t, \theta\right] \mid \mathcal{F}_T, \theta\right] \\
&= \mathbb{E}\left[f_{t|t+1}(\theta) \mid \mathcal{F}_T, \theta\right] \\
&= f_{t|t}(\theta) + P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1}\left(f_{t+1|T}(\theta) - f_{t+1|t}(\theta)\right),
\end{aligned}$$

where the second equality follows from the law of iterated expectations, and the third from the independence of E_{t+1} and f_t given f_{t+1} and $\mathcal{F}_t$. Similarly,

$$\begin{aligned}
P_{t|T}(\theta) &= \mathbb{E}\left[(f_t - f_{t|T}(\theta))(f_t - f_{t|T}(\theta))' \mid \mathcal{F}_T, \theta\right] \\
&= \mathbb{E}\left[f_t f_t' \mid \mathcal{F}_T, \theta\right] - f_{t|T}(\theta) f_{t|T}(\theta)' \\
&= \mathbb{E}\left[\mathbb{E}\left[f_t f_t' \mid f_{t+1}, \mathcal{F}_t, \theta\right] \mid \mathcal{F}_T, \theta\right] - f_{t|T}(\theta) f_{t|T}(\theta)' \\
&= \mathbb{E}\left[P_{t|t+1}(\theta) + f_{t|t+1}(\theta) f_{t|t+1}(\theta) \mid \mathcal{F}_T, \theta\right] - f_{t|T}(\theta) f_{t|T}(\theta)' \\
&= P_{t|t}(\theta) - P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta) \\
&\quad + \mathrm{Var}\left(f_{t|t+1}(\theta) \mid \mathcal{F}_T, \theta\right) \\
&= P_{t|t}(\theta) - P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta) \\
&\quad + P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1} P_{t+1|T}(\theta) P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta) \\
&= P_{t|t}(\theta) - P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1}\left[P_{t+1|t}(\theta) - P_{t+1|T}(\theta)\right] P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta).
\end{aligned}$$

In summary, the smoothed factors and the smoothed factor variance under the parameter values θ are given as

$$f_{t|T}(\theta) = f_{t|t}(\theta) + P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1}\left(f_{t+1|T}(\theta) - f_{t+1|t}(\theta)\right) \tag{3.8}$$

$$P_{t|T}(\theta) = P_{t|t}(\theta) - P_{t|t}(\theta)G' P_{t+1|t}(\theta)^{-1}\left[P_{t+1|t}(\theta) - P_{t+1|T}(\theta)\right] P_{t+1|t}(\theta)^{-1} G P_{t|t}(\theta). \tag{3.9}$$

ii) **The Singular Case**

This case is trickier to deal with. If $q < r$, then $P_{t+1|t}(\theta)^{-1}$ is likely to not exist, so that the formulas above become inadmissible. As such, we pursue a different smoothing method. The Kalman smoother in this case is given as

$$f_{t|T}(\theta) = f_{t|t-1}(\theta) + P_{t|t-1}(\theta) r_{t-1} \tag{3.10}$$

$$P_{t|T}(\theta) = \Big(I_r - P_{t|t-1}(\theta) N_{t-1}(\theta)\Big) P_{t|t-1}(\theta) \tag{3.11}$$

$$r_{t-1}(\theta) = \Lambda' V_{t|t-1}(\theta)^{-1} \Big(x_t - x_{t|t-1}(\theta)\Big) + L_t(\theta)' r_t(\theta) \tag{3.12}$$

$$N_{t-1}(\theta) = \Lambda' V_{t|t-1}(\theta)^{-1} \Lambda + L_t(\theta)' N_t(\theta) L_t(\theta) \tag{3.13}$$

$$r_T(\theta) = O_{r \times 1} \tag{3.14}$$

$$N_T(\theta) = O_{r \times r} \tag{3.15}$$

$$L_t(\theta) = G\Big(I_r - K_{t|t-1}(\theta)\Lambda\Big) \tag{3.16}$$

for $1 \leq t \leq T$. For details, consult the appendix.

Assuming that the parameter values are known, the Kalman smoothed factors $f_{t|T}(\theta)$ represent the most accurate estimators of the factors f_t given the data in the sample. Since the cross-sectional dimension N does not go to infinity in small DFMs, the asymptotic convergence of $f_{t|T}(\theta)$ to the true factor values f_t is not guaranteed; for this reason, the smoothed values are often used to estimate the factors, since they are at least the most accurate.

The filtered values and smoothed values derived above are mean-square optimal when the error terms are Gaussian. However, even when the error terms are non-Gaussian, they retain some value as the best linear projections, that is, $f_{t|t}(\theta)$ represents the best linear projection of f_t on the variables collected in $\mathcal{F}_t$, $P_{t|t}(\theta)$ is the mean squared projection error associated with $f_{t|t}(\theta)$, and so on. This is why their use is so widespread even beyond Gaussian systems. For more details on the linear projection approach to the Kalman filter and smoother, consult Hamilton (2020).

The EM Algorithm

The **expecation-maximization (EM) algorithm** is an iterative method that is designed to find the local maxima of the log likelihood function $l(x_T, \cdots, x_1 \mid \theta)$. Denote $X = (x_1, \cdots, x_T)$ and $F = (f_0, f_1, \cdots, f_T)$. Given initial parameter values $\theta^{(0)}$, the algorithm consists of two steps:

1) **The E-Step**
In the E-step, which is short for the "expectation" step, we calculate the expectation

$$\begin{aligned} Q(\theta \mid \theta^{(i)}) &= \mathbb{E}\left[\log f(X, F \mid \theta) \mid X, \theta^{(i)}\right] \\ &= \int \log f(X, F \mid \theta) \cdot f(F \mid X, \theta^{(i)}) dF. \end{aligned}$$

2) **The M-Step**
In the M-step, which is short for the "maximization" step, we update the parameter estimates by maximizing $Q(\theta \mid \theta^{(i)})$ with respect to θ:

$$\theta^{(i+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}} \quad Q(\theta \mid \theta^{(i)}).$$

The algorithm stops once it has converged according to the convergence criterion

$$c_k = \frac{\left| l(X \mid \theta^{(k)}) - l(X \mid \theta^{(k-1)}) \right|}{\frac{1}{2}\left| l(X \mid \theta^{(k)}) + l(X \mid \theta^{(k-1)}) \right|}.$$

This is the convergence criterion used in Doz, Giannone, and Reichlin (2012), where the algorithm is said to have converged if $c_k < 10^{-4}$. The formal reasoning as to why the EM algorithm helps find local maxima is contained in the appendix.

For linear state space models, there exist closed form solutions for both the E-and M-steps invovling the Kalman smoothed quanities. In what follows, we assume that $c = O_{r \times 1}$ for simplicity.

In the E-step, we must calculate the expectation

$$Q(\theta \mid \theta^{(i)}) = \int \log f(X, F \mid \theta) \cdot f(F \mid X, \theta^{(i)}) dF.$$

To do so, we first decompose the integrand as follows:

$$\begin{aligned}\log f(X,F \mid \theta) &= \sum_{t=1}^{T} \log f(x_t, f_t \mid \mathcal{F}_{t-1}, f_{t-1}, \cdots, f_0, \theta) \\ &= \sum_{t=1}^{T} \log f(x_t \mid f_t, \theta) + \sum_{t=1}^{T} \log f(f_t \mid f_{t-1}, \theta).\end{aligned}$$

Since

$$\begin{aligned}x_t \mid f_t, \theta &\sim \mathcal{N}\left[\Lambda f_t, \Sigma\Sigma'\right] \\ f_t \mid f_{t-1}, \theta &\sim \mathcal{N}\left[G f_{t-1}, HH'\right],\end{aligned}$$

we can write the above expression as

$$\begin{aligned}\log f(X,F \mid \theta) \simeq &-\frac{T}{2}\log\left|\Sigma\Sigma'\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t - \Lambda f_t)'(\Sigma\Sigma')^{-1}(x_t - \Lambda f_t) \\ &-\frac{T}{2}\log\left|HH'\right| - \frac{1}{2}\sum_{t=1}^{T}(f_t - G f_{t-1})'(HH')^{-1}(f_t - G f_{t-1}),\end{aligned}$$

where the equality is up to constants and initial values. If HH' is singular, as it is in the case $q < r$, then we can replace $|HH'|$ by $|H'H|$ and $(HH')^{-1}$ by $H^{\dagger\prime}H^{\dagger}$, where

$$H^{\dagger} = (H'H)^{-1}H'$$

is the left multiplication pseudo-inverse of H. In this case, $H^{\dagger\prime}H^{\dagger}$ has rank equal to q, since H is a full rank matrix.

Now the expectation can be evaluated as

$$\begin{aligned}Q(\theta \mid \theta^{(i)}) = &-\frac{T}{2}\log\left|\Sigma\Sigma'\right| - \frac{T}{2}\log\left|HH'\right| \\ &-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left((\Sigma\Sigma')^{-1}\cdot\mathbb{E}\left[(x_t - \Lambda f_t)(x_t - \Lambda f_t)' \mid \mathcal{F}_T, \theta^{(i)}\right]\right) \\ &-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left((HH')^{-1}\cdot\mathbb{E}\left[(f_t - c - G f_{t-1})(f_t - c - G f_{t-1})' \mid \mathcal{F}_T, \theta^{(i)}\right]\right).\end{aligned}$$

We start with the term invovling x_t. For any $1 \le t \le T$,

$$\begin{aligned}\mathbb{E}\left[(x_t - \Lambda f_t)(x_t - \Lambda f_t)' \mid \mathcal{F}_T, \theta^{(i)}\right] &= x_t x_t' - \Lambda\mathbb{E}\left[f_t \mid \mathcal{F}_T, \theta^{(i)}\right]x_t' \\ &\quad - x_t\mathbb{E}\left[f_t' \mid \mathcal{F}_T, \theta^{(i)}\right]\Lambda' + \Lambda\mathbb{E}\left[f_t f_t' \mid \mathcal{F}_T, \theta^{(i)}\right]\Lambda' \\ &= x_t x_t' - \Lambda f_{t|T}^{(i)} x_t'\end{aligned}$$

$$
\begin{aligned}
&- x_t f_{t|T}^{(i)\prime}\Lambda' + \Lambda\left[P_{t|T}^{(i)} + f_{t|T}^{(i)} f_{t|T}^{(i)\prime}\right]\Lambda' \\
&= \left(x_t - \Lambda f_{t|T}^{(i)}\right)\left(x_t - \Lambda f_{t|T}^{(i)}\right)' + \Lambda P_{t|T}^{(i)}\Lambda'.
\end{aligned}
$$

Meanwhile, the term invovling f_t can be evaluated as

$$
\begin{aligned}
&\mathbb{E}\left[(f_t - Gf_{t-1})(f_t - Gf_{t-1})' \mid \mathcal{F}_T, \theta^{(i)}\right] \\
&\qquad = \mathbb{E}\left[f_t f_t' \mid \mathcal{F}_T, \theta^{(i)}\right] - \mathbb{E}\left[f_t f_{t-1}' \mid \mathcal{F}_T, \theta^{(i)}\right] G' \\
&\qquad - G\mathbb{E}\left[f_{t-1} f_t' \mid \mathcal{F}_T, \theta^{(i)}\right] + G\mathbb{E}\left[f_{t-1} f_{t-1}' \mid \mathcal{F}_T, \theta^{(i)}\right] G \\
&\qquad = P_{t|T}^{(i)} + f_{t|T}^{(i)} f_{t|T}^{(i)\prime} - \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)}\right) G' \\
&\qquad\quad - G\cdot\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)}\right)' + G P_{t-1|T}^{(i)} G' + G f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} G',
\end{aligned}
$$

where we define[8]

$$
C_{t,t-1|T}^{(i)} = \mathbb{E}\left[\left(f_t - f_{t|T}\right)\left(f_{t-1} - f_{t-1|T}\right)' \mid \mathcal{F}_T, \theta^{(i)}\right].
$$

The expectation is thus computed as

$$
\begin{aligned}
Q(\theta \mid \theta^{(i)}) = &\frac{T}{2}\left(\log\left|(\Sigma\Sigma')^{-1}\right| + \log\left|(HH')^{-1}\right|\right) \\
&- \frac{1}{2}\sum_{t=1}^{T} \operatorname{tr}\left[(\Sigma\Sigma')^{-1}\left(x_t - \Lambda f_{t|T}^{(i)}\right)\left(x_t - \Lambda f_{t|T}^{(i)}\right)'\right] \\
&- \frac{1}{2}\sum_{t=1}^{T} \operatorname{tr}\left[(\Sigma\Sigma')^{-1}\Lambda P_{t|T}^{(i)}\Lambda'\right] \\
&- \frac{1}{2}\sum_{t=1}^{T} \operatorname{tr}\left[(HH')^{-1}\left(P_{t|T}^{(i)} + f_{t|T}^{(i)} f_{t|T}^{(i)\prime}\right)\right] \\
&+ \sum_{t=1}^{T} \operatorname{tr}\left[(HH')^{-1}\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)}\right) G'\right]
\end{aligned}
$$

[8]For the computation of $C_{t,t-1|T}$, we first obtain the smoothed factor variance $\mathcal{P}_{t|T}$ from the augmented state space model

$$
x_t = \begin{pmatrix}\Lambda & O_{N\times r}\end{pmatrix}\begin{pmatrix}F_t \\ F_{t-1}\end{pmatrix} + \Sigma e_t
$$

$$
\begin{pmatrix}F_t \\ F_{t-1}\end{pmatrix} = \begin{pmatrix}G & O_{r\times r} \\ I_r & O_{r\times r}\end{pmatrix}\begin{pmatrix}F_{t-1} \\ F_{t-2}\end{pmatrix} + \begin{pmatrix}H \\ O_{r\times q}\end{pmatrix} u_t.
$$

and take the block matrix in its (1,2) position. Since the factor innovation covariance matrix of this augmented model is necessarily singular, we implement the recursions for the Kalman smoother in the case $q < r$ to compute $\mathcal{P}_{t|T}$.

$$-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left[(HH')^{-1}G\left(P_{t-1|T}^{(i)}+f_{t-1|T}^{(i)}f_{t-1|T}^{(i)\prime}\right)G'\right].$$

For the sake of completeness, we state the form of the expectation when the variance HH' is singular:

$$\begin{aligned}
Q(\theta\mid\theta^{(i)}) = \frac{T}{2}&\left(\log\left|(\Sigma\Sigma')^{-1}\right|+\log\left|H^{\dagger}H^{\dagger\prime}\right|\right)\\
&-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left[(\Sigma\Sigma')^{-1}\left(x_t-\Lambda f_{t|T}^{(i)}\right)\left(x_t-\Lambda f_{t|T}^{(i)}\right)'\right]\\
&-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left[(\Sigma\Sigma')^{-1}\Lambda P_{t|T}^{(i)}\Lambda'\right]\\
&-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left[\left(H^{\dagger\prime}H^{\dagger}\right)\left(P_{t|T}^{(i)}+f_{t|T}^{(i)}f_{t|T}^{(i)\prime}\right)\right]\\
&+\sum_{t=1}^{T}\operatorname{tr}\left[\left(H^{\dagger\prime}H^{\dagger}\right)\left(C_{t,t-1|T}^{(i)}+f_{t|T}^{(i)}f_{t-1|T}^{(i)\prime}\right)G'\right]\\
&-\frac{1}{2}\sum_{t=1}^{T}\operatorname{tr}\left[\left(H^{\dagger\prime}H^{\dagger}\right)G\left(P_{t-1|T}^{(i)}+f_{t-1|T}^{(i)}f_{t-1|T}^{(i)\prime}\right)G'\right].
\end{aligned}$$

Moving onto the M-step, we must maximize $Q(\theta\mid\theta^{(i)})$ with respect to θ if we are to find the next iterates $\theta^{(i+1)}$. To find θ, we must maximize $Q(\theta\mid\theta^{(i)})$ with respect to:

$$\Lambda,\underbrace{\Sigma\Sigma'}_{\Gamma^x},c,G,\underbrace{HH'}_{\Gamma^f}.$$

First, note that

$$\begin{aligned}
&-\frac{1}{2}\sum_{t=1}^{T}\left(\operatorname{tr}\left[(\Sigma\Sigma')^{-1}\left(x_t-\Lambda f_{t|T}^{(i)}\right)\left(x_t-\Lambda f_{t|T}^{(i)}\right)'\right]-\operatorname{tr}\left[(\Sigma\Sigma')^{-1}\Lambda P_{t|T}^{(i)}\Lambda'\right]\right)\\
&=-\frac{1}{2}\sum_{t=1}^{T}\left(x_t-\Lambda f_{t|T}^{(i)}\right)'(\Gamma^x)^{-1}\left(x_t-\Lambda f_{t|T}^{(i)}\right)-\frac{1}{2}\sum_{t=1}^{T}\operatorname{vec}(\Lambda)'\left(P_{t|T}^{(i)}\bigotimes(\Gamma^x)^{-1}\right)\operatorname{vec}(\Lambda)\\
&=-\frac{1}{2}\sum_{t=1}^{T}\left[x_t-\left(f_{t|T}^{(i)\prime}\bigotimes I_N\right)\operatorname{vec}(\Lambda)\right]'(\Gamma^x)^{-1}\left[x_t-\left(f_{t|T}^{(i)\prime}\bigotimes I_N\right)\operatorname{vec}(\Lambda)\right]\\
&\qquad-\frac{1}{2}\sum_{t=1}^{T}\operatorname{vec}(\Lambda)'\left(P_{t|T}^{(i)}\bigotimes(\Gamma^x)^{-1}\right)\operatorname{vec}(\Lambda),
\end{aligned}$$

so we have

$$\frac{\partial Q(\theta \mid \theta^{(i)})}{\partial \mathrm{vec}\,(\Lambda)} = \sum_{t=1}^{T} \left(f_{t|T}^{(i)} \bigotimes I_N \right) (\Gamma^x)^{-1} \left[x_t - \left(f_{t|T}^{(i)\prime} \bigotimes I_N \right) \mathrm{vec}\,(\Lambda) \right] - \sum_{t=1}^{T} \left(P_{t|T}^{(i)} \bigotimes (\Gamma^x)^{-1} \right) \mathrm{vec}\,(\Lambda).$$

$\Lambda^{(i+1)}$ equates the above first order condition with 0, so that

$$\begin{aligned} \mathrm{vec}\left(\Lambda^{(i+1)}\right) &= \left[\left(\sum_{t=1}^{T} \left(f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)} \right) \right)^{-1} \bigotimes I_N \right] \mathrm{vec} \left(\sum_{t=1}^{T} x_t f_{t|T}^{(i)\prime} \right) \\ &= \mathrm{vec} \left(\left(\sum_{t=1}^{T} x_t f_{t|T}^{(i)\prime} \right) \left(\sum_{t=1}^{T} \left(f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)} \right) \right)^{-1} \right), \end{aligned}$$

and by implication,

$$\Lambda^{(i+1)} = \left(\sum_{t=1}^{T} x_t f_{t|T}^{(i)\prime} \right) \left(\sum_{t=1}^{T} \left(f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)} \right) \right)^{-1}.$$

Likewise, since

$$\begin{aligned} &\sum_{t=1}^{T} \mathrm{tr} \left[(HH')^{-1} \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) G' \right] - \frac{1}{2} \sum_{t=1}^{T} \mathrm{tr} \left[(HH')^{-1} G \left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) G' \right] \\ &\quad = -\frac{1}{2} \sum_{t=1}^{T} \mathrm{vec}\,(G)' \left[\left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) \bigotimes \left(\Gamma^f \right)^{-1} \right] \mathrm{vec}\,(G) \\ &\qquad + \sum_{t=1}^{T} \mathrm{vec}\,(G)' \left[I_r \bigotimes \left(\Gamma^f \right)^{-1} \right] \mathrm{vec} \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right), \end{aligned}$$

we have

$$\begin{aligned} \frac{\partial Q(\theta \mid \theta^{(i)})}{\partial \mathrm{vec}\,(G)} &= \sum_{t=1}^{T} \left[\left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) \bigotimes \left(\Gamma^f \right)^{-1} \right] \mathrm{vec}\,(G) \\ &\quad + \sum_{t=1}^{T} \left[I_r \bigotimes \left(\Gamma^f \right)^{-1} \right] \mathrm{vec} \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right). \end{aligned}$$

$G^{(i+1)}$ must then satisfy

$$\mathrm{vec}\left(G^{(i+1)}\right) = \left[\left(\sum_{t=1}^{T} \left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) \right)^{-1} \bigotimes I_r \right]^{-1} \mathrm{vec} \left(\sum_{t=1}^{T} \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) \right)$$

$$= \text{vec}\left(\left[\sum_{t=1}^{T}\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime}\right)\right]\left[\sum_{t=1}^{T}\left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime}\right)\right]^{-1}\right),$$

so that

$$G^{(i+1)} = \left[\sum_{t=1}^{T}\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime}\right)\right]\left[\sum_{t=1}^{T}\left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime}\right)\right]^{-1}.$$

Moving onto the covariance terms, we first obtain an expression for Γ^x. Since

$$\frac{\partial Q(\theta \mid \theta^{(i)})}{\partial (\Gamma^x)^{-1}} = \frac{T}{2}(\Gamma^x)^{-1} - \frac{1}{2}\sum_{t=1}^{T}\left[\left(x_t - \Lambda f_{t|T}^{(i)}\right)\left(x_t - \Lambda f_{t|T}^{(i)}\right)' + \Lambda P_{t|T}^{(i)} \Lambda'\right],$$

$\Gamma^{x(i+1)}$ should be given as

$$\begin{aligned}
\Gamma^{x(i+1)} &= \frac{1}{T}\sum_{t=1}^{T}\left[\left(x_t - \Lambda^{(i+1)} f_{t|T}^{(i)}\right)\left(x_t - \Lambda^{(i+1)} f_{t|T}^{(i)}\right)' + \Lambda^{(i+1)} P_{t|T}^{(i)} \Lambda^{(i+1)\prime}\right] \\
&= \frac{1}{T}\sum_{t=1}^{T} x_t x_t' - \frac{1}{T}\sum_{t=1}^{T}\Lambda^{(i+1)} f_{t|T}^{(i)} x_t' - \frac{1}{T}\sum_{t=1}^{T} x_t f_{t|T}^{(i)} \Lambda^{(i+1)\prime} \\
&\quad + \frac{1}{T}\sum_{t=1}^{T}\Lambda^{(i+1)}\left(f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)}\right)\Lambda^{(i+1)\prime}.
\end{aligned}$$

$\Sigma^{(i+1)}$ can then be defined as the Cholesky factor of $\Gamma^{x(i+1)}$.

The updated value of Γ^f is trickier to obtain. If $q = r$, then analogously to the preceding result,

$$\begin{aligned}
\Gamma^{f(i+1)} &= \frac{1}{T}\sum_{t=1}^{T}\left[f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)}\right] \\
&\quad - \frac{1}{T}\sum_{t=1}^{T}\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime}\right) G^{(i+1)\prime} - \frac{1}{T}\sum_{t=1}^{T} G^{(i+1)}\left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime}\right)' \\
&\quad + \frac{1}{T}\sum_{t=1}^{T} G^{(i+1)}\left(P_{t-1|T}^{(i)} + f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime}\right) G^{(i+1)\prime}.
\end{aligned}$$

$H^{(i+1)}$ can then be taken as the square root[9] of $\Gamma^{f(i+1)}$.

[9]The square root of a positive semidefinite matrix A is taken to be $PD^{\frac{1}{2}}$, where $A = PDP'$ is the eigendecomposition of A. We take the square root instead of the Cholesky factor to preserve consistency between the cases $q = r$ and $q < r$.

When $q < r$, we take

$$H^{(i+1)} = W^{(i+1)} \left(M^{(i+1)}\right)^{\frac{1}{2}},$$

where $W^{(i+1)}$ is an $r \times q$ matrix whose columns are orthonormal eigenvectors of $\Gamma^{f(i+1)}$ corresponding to the q largest eigenvalues, and $M^{(i+1)}$ is a diagonal $q \times q$ matrix collecting these eigenvalues. Note that $H^{(i+1)}$) is exactly the square root of $\Gamma^{f(i+1)}$, so that we may take $H^{(i+1)}$ to be equal to the above quantity regardless of whether $q = r$ or $q < r$.

To obtain a more tractable expression for the values obtained so far, define

$$\begin{aligned}
\mathbf{Z} &= \sum_{t=1}^{T} x_t f_{t|T}^{(i)} \\
\mathbf{E} &= \sum_{t=1}^{T} \left[C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right] \\
\mathbf{F} &= \sum_{t=1}^{T} \left[f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)} \right] \\
\mathbf{F}_{-1} &= \sum_{t=1}^{T} \left[f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} + P_{t-1|T}^{(i)} \right].
\end{aligned}$$

Then, we have

$$\begin{aligned}
\Lambda^{(i+1)} &= \mathbf{Z}\mathbf{F}^{-1} \\
G^{(i+1)} &= \mathbf{E}\mathbf{F}_{-1}^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\Gamma^{x(i+1)} &= \frac{1}{T}\left(X'X - \Lambda^{(i+1)}\mathbf{Z}' - \mathbf{Z}\Lambda^{(i+1)\prime} + \Lambda^{(i+1)}\mathbf{F}\Lambda^{(i+1)\prime}\right) \\
&= \frac{1}{T}\left(X'X - \mathbf{Z}\mathbf{F}^{-1}\mathbf{Z}'\right) \\
\Gamma^{f(i+1)} &= \frac{1}{T}\left(\mathbf{F} - \mathbf{E}G^{(i+1)\prime} - G^{(i+1)}\mathbf{E}' + G^{(i+1)}\mathbf{F}_{-1}G^{(i+1)\prime}\right) \\
&= \frac{1}{T}\left(\mathbf{F} - \mathbf{E}F_{-1}^{-1}\mathbf{E}'\right) \\
&= \frac{1}{T}\sum_{t=1}^{T}\left[f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)} - \left(C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime} \right) G^{(i+1)\prime} \right].
\end{aligned}$$

Summarizing, at the end of the M-step, we are left with the following updated parameter values:

$$\Lambda^{(i+1)} = \mathbf{Z}\mathbf{F}^{-1} \tag{3.17}$$

$$G^{(i+1)} = \mathbf{E}\mathbf{F}_{-1}^{-1} \tag{3.18}$$

$$\Gamma^{x(i+1)} = \frac{1}{T}\left(X'X - \mathbf{Z}\mathbf{F}^{-1}\mathbf{Z}\right) \tag{3.19}$$

$$\Gamma^{f(i+1)} = \frac{1}{T}\left(\mathbf{F} - \mathbf{E}\mathbf{F}_{-1}^{-1}\mathbf{E}'\right) \tag{3.20}$$

where

$$\mathbf{Z} = \sum_{t=1}^{T} x_t f_{t|T}^{(i)} \tag{3.21}$$

$$\mathbf{E} = \sum_{t=1}^{T} \left[C_{t,t-1|T}^{(i)} + f_{t|T}^{(i)} f_{t-1|T}^{(i)\prime}\right] \tag{3.22}$$

$$\mathbf{F} = \sum_{t=1}^{T} \left[f_{t|T}^{(i)} f_{t|T}^{(i)\prime} + P_{t|T}^{(i)}\right] \tag{3.23}$$

$$\mathbf{F}_{-1} = \sum_{t=1}^{T} \left[f_{t-1|T}^{(i)} f_{t-1|T}^{(i)\prime} + P_{t-1|T}^{(i)}\right]. \tag{3.24}$$

3.4.2 Estimating Large Dynamic Factor Models

For the sake of completeness, we re-state the basic state-space form of our DFM:

$$\begin{aligned} x_t &= \Lambda f_t + \Sigma e_t \\ f_t &= c + G f_{t-1} + H u_t. \end{aligned}$$

The QMLE method is often used to estimate small DFMS, or those with a small cross-sectional dimension N. If the cross-sectional dimension N is large, on the other hand, we have a **large dynamic factor model (LDFM)**, and QMLE method above cannot be used without modification. Most notably, since the parameters

$$\Lambda, \Sigma, c, G, H$$

must be estimated in the above model, there are a total of $Nr + \frac{N(N+1)}{2} + r + r^2 + rq = O(N^2)$ parameters that must be estimated. However, since both T and N go to infinity

in LDFMs, the number of parameters to be estimated can easily to go infinity faster than the number of observations NT, which makes estimation unreliable. For this reason, two alternative methods of estimating LDFMs have been proposed.

The first, proposed in Doz, Giannone, and Reichlin (2011), is a two-step estimator based on the non-parametric PC estimator of static factor models. Recall that assumptions on the cross-sectional and time series dependence on the idiosyncratic errors Σe_t, as well as the boundedness of the true factors f_t and factor loadings Λ are sufficient to guarantee that the PC estimators are well-behaved. Therefore, consistent estimators of the factors and factor loadings in the measurement equation can be obtained even without taking into consideration the dynamics of the factors contained in the transition equation.

Using the consistent estimators of the factors as proxies for the true factors, an OLS regression now yields estimates of the factor intercept c, mean reversion parameter G and the innovation matrix HH'. It remains to estimate the covariance of the idiosyncratic errors e_t, and it is here that we introduce an approximation; specifically, we use an approximating model where the idiosyncratic errors are uncorrelated and homoskedastic, so that their covariance matrix is given as $\sigma^2 I_N$. σ^2 can then be estimated as a function of the PC estimators and data. Finally, once these parameter estimates have been obtained, we use them to estimate the factors using the Kalman smoother. Formally, we proceed as follows:

Step 1: Initial Estimates of the Parameters

The factor loadings Λ, the time t factors f_t, and the idiosyncratic variance σ^2 are estimated via PCA as follows:

$$\overline{\Lambda} = \sqrt{N} \times r \text{ orthonormal eigenvectors of } \frac{1}{NT} X'X \text{ corresponding to the } r \text{ largest eigenvalues } \mu_1 \geq \cdots \geq \mu_r > 0$$

$$\overline{f}_t = \left(\overline{\Lambda}'\overline{\Lambda}\right)^{-1} \overline{\Lambda}' x_t = \frac{1}{N} \overline{\Lambda}' x_t$$

$$\overline{F} = \begin{pmatrix} \overline{f}_1' \\ \vdots \\ \overline{f}_T' \end{pmatrix} = \frac{1}{N} X \overline{\Lambda}$$

$$\overline{\sigma}^2 = \frac{1}{NT} \operatorname{tr}\left(X'X\right) - \frac{1}{NT} \sum_{i=1}^{r} \mu_i.$$

Using the estimated factors $\overline{f}_t$, we estimate the transition equation parameters c, G

and $\Gamma^f = HH'$ as follows:

$$\begin{pmatrix} \overline{c}' \\ \overline{G}' \end{pmatrix} = \left(\sum_{t=2}^{T} \overline{z}_{t-1} \overline{z}'_{t-1} \right)^{-1} \left(\sum_{t=2}^{T} \overline{z}_{t-1} \overline{f}'_t \right),$$

$$\overline{\Gamma^f} = \frac{1}{T} \sum_{t=2}^{T} \left(\overline{f}_t - \overline{c} - \overline{G f}_{t-1} \right) \left(\overline{f}_t - \overline{c} - \overline{G f}_{t-1} \right)',$$

where

$$\overline{z}_{t-1} = \begin{pmatrix} 1 \\ \overline{f}_{t-1} \end{pmatrix}.$$

Step 2: Smoothed Estimates of the Factors

Given the estimated parameters

$$\overline{\theta} = \{\overline{\Lambda}, \overline{\sigma}^2 I_N, \overline{c}, \overline{G}, \overline{\Gamma^f}\},$$

we can obtain the smoothed version of the factors and their variances,

$$f_{t|T}(\overline{\theta}) \quad \text{and} \quad P_{t|T}(\overline{\theta}).$$

The smoothers can be computed using either method introduced in the previous section, depending on whether H is a square matrix ($q = r$) or not ($q < r$).

The smoothing process serves to eliminate any noise that may have been present in the original factor estimates $\overline{F}$.

It is shown in Doz, Giannone, and Reichlin (2011), among others, that the estimates $\overline{\Lambda}, \overline{c}, \overline{G}$ and $\overline{\Gamma^f}$ are consistent as $N, T \to \infty$ regardless of the approximation introduced into the model through the covariance matrix of the idiosyncratic errors. Likewise, the smoothed factors $f_{t|T}(\overline{\theta})$ are shown to be consistent for a rotation of the true factors f_t as $N, T \to \infty$ despite the additional approximation of Gaussian errors, under which we derived the Kalman smoother.

An alternative approach, studied in Barigozzi and Luciani (2019), is the usual QMLE method, with the added approximation that the idiosyncratic errors are uncorrelated and (possibly) heteroskedastic. This means that the parameters to be estimated are

$$\Lambda, \sigma_1^2, \cdots, \sigma_N^2, c, G, H,$$

where σ_i^2 is the variance of the ith idiosyncratic error. In other words, we need only estimate a total of $Nr+N+r+r^2+rq=O(N)$ parameters, so that the number of parameters will be smaller than the number of obserations NT as $N,T\to\infty$. Estimation of the model is done through the EM algorithm; the updated parameter values are the same as above, except that now $\Gamma^{x(i+1)}$ is given as a diagonal matrix whose diagonal entries are those of

$$\frac{1}{T}\sum_{t=1}^{T}\left[\left(x_t-\Lambda^{(i+1)}f_{t|T}^{(i)}\right)\left(x_t-\Lambda^{(i+1)}f_{t|T}^{(i)}\right)'+\Lambda^{(i+1)}P_{t|T}^{(i)}\Lambda^{(i+1)\prime}\right].$$

In other words, we are estimating the off-diagonal terms of $\Gamma^x=\Sigma\Sigma'$ as 0. It is pointed out in Barigozzi and Luciani (2019) that the two-step estimator is a special case of the QMLE method with the number of iterations equal to 1. Below are summarized all the approximations made when estimating the model by QMLE:

- The idiosyncratic errors e_t and factor innovations u_t are approximated by i.i.d. normally distributed processes.
- The global maximum of the log-likelihood is approximated by the local maximum to which the EM algorithm converges.
- Values like $\mathbb{E}[f_tf_t'\mid\mathcal{F}_T]$ are approximated by their Kalman smoother values, which are equal only when the system is Gaussian.
- (**For large DFMs**) The cross-sectional dependence of the idiosyncratic errors e_t are approximated by 0.

In spite of these approximations made in the estimation process, Barigozzi and Luciani (2019) show that, under weak assumptions on the cross-sectional and temporal dependence of the error terms, the QMLE estimators of the model parameters and factors are consistent.

3.4.3 Estimating the Dynamic Nelson-Siegel Model

We now apply the techniques studied above to estimate the DN-S model. To this end, we opt for the two-step method in Doz, Giannone, and Reichlin (2011) for its relative computational simplicity. Below we adapt the two-step estimation procedure for the DN-S model, which is made simple because we already derived the estimates for the N-S model earlier:

Step 1: Initial Estimates of the Parameters

The decay parameter κ, the time t factors f_t, and the idiosyncratic variance σ^2 are estimated via least squares as follows:

$$\overline{\kappa} = \underset{\kappa\in[\epsilon,1-\epsilon]}{\operatorname{argmax}} \quad \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}'\mathcal{Y}\right) - \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\mathcal{Y}'\right)$$

$$\overline{f}_t = \left[\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right]^{-1}\Lambda(\overline{\kappa})'\mathcal{Y}_t$$

$$\overline{F} = \begin{pmatrix}\overline{f}_1' \\ \vdots \\ \overline{f}_T'\end{pmatrix} = \mathcal{Y}\Lambda(\overline{\kappa})\left[\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right]^{-1}$$

$$\overline{\sigma}^2 = \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}'\mathcal{Y}\right) - \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\mathcal{Y}'\right)$$

Using the estimated factors $\overline{f}_t$, we estimate the transition equation parameters c, G and $\Gamma^f = HH'$ as follows:

$$\begin{pmatrix}\overline{c}' \\ \overline{G}'\end{pmatrix} = \left(\sum_{t=2}^{T}\overline{z}_{t-1}\overline{z}_{t-1}'\right)^{-1}\left(\sum_{t=2}^{T}\overline{z}_{t-1}\overline{f}_t'\right),$$

$$\overline{\Gamma^f} = \frac{1}{T}\sum_{t=2}^{T}\left(\overline{f}_t - \overline{c} - \overline{G}\overline{f}_{t-1}\right)\left(\overline{f}_t - \overline{c} - \overline{G}\overline{f}_{t-1}\right)',$$

where

$$\overline{z}_{t-1} = \begin{pmatrix}1 \\ \overline{f}_{t-1}\end{pmatrix}.$$

Step 2: Smoothed Estimates of the Factors

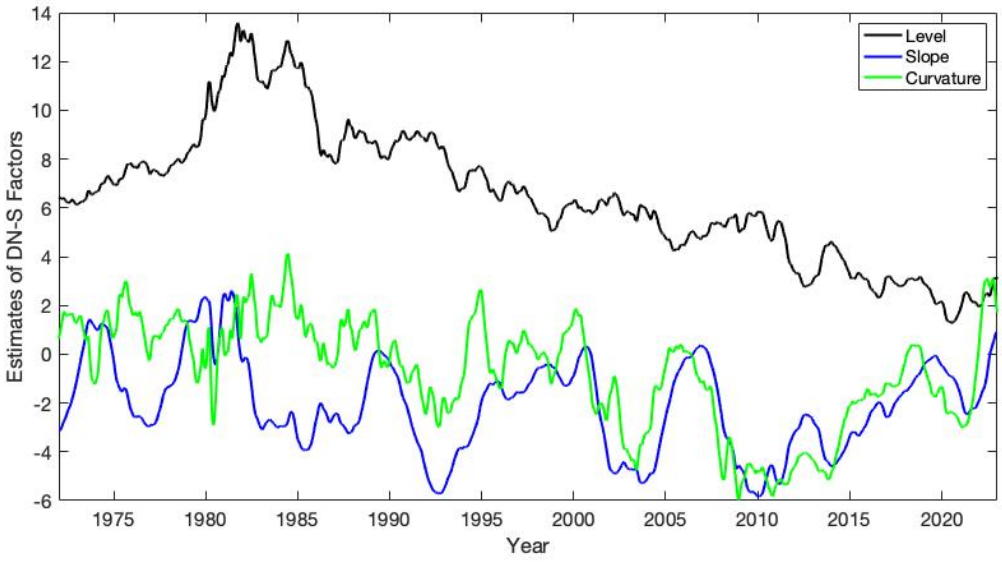
Given the estimated parameters

$$\overline{\theta} = \{\Lambda(\overline{\kappa}), \overline{\sigma}^2 I_N, \overline{c}, \overline{G}, \overline{\Gamma^f}\},$$

we can obtain the smoothed version of the factors and their variances,

$$f_{t|T}(\overline{\theta}) \quad \text{and} \quad P_{t|T}(\overline{\theta}).$$

Figure 3.8: **Estimated Dynamic Nelson-Siegel Factors** The figure below maps the smoothed N-S factor estimates obtained via the two-step estimation procedure.



It is shown in the appendix that the estimates of the transition equation parameters c, G, H obtained using the first step estimates of the factors are consistent for their true values. The consistency of $\overline{\sigma}^2$ for the true measurement error variance σ^2 follows easily from the proof in appendix A, and the fact that

$$\frac{1}{\sqrt{m}}\left\|\Lambda^0 - \Lambda(\overline{\kappa})\right\| = o_p(1)$$

and

$$\left\|\left(\frac{\Lambda^{0\prime}\Lambda^0}{m}\right)^{-1} - \left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1}\right\| = o_p(1).$$

The smoothed estimates of the dynamic N-S factors are presented in Figure 3.8. Compared to the N-S factors estimated via least squares, the smoothed factors are literally more smooth. In general, they follow the same trends as the N-S factors estimated earlier, which is a testament to the accuracy of the least squares estimators.

Chapter 4

Affine Term Structure Models

In this chapter, we study the basic affine term structure model as formulated in Duffie and Kan (1996)[1] and developed in subsequent works such as Dai and Singleton (2000), Dai and Singleton (2002) and Joslin, Singleton, and Zhu (2011).

In term structure models, it is typically assumed that there exists a sequence of latent factors $\{f_t\}_{t\in\mathbb{N}}$ such that time t bond prices and yields are determined as functions of the time t factors f_t. Suppose there are n factors, so that each f_t is an n-dimensional random vector.

We first make preliminary assumptions. Throughout, we assume that there exists an SDF process $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ with $\mathcal{M}_0 = 1$ such that the time t price of an asset with time $t+1$ payoff X_{t+1} is given by

$$\mathbb{E}_t\left[\mathcal{M}_{t+1}X_{t+1}\right].$$

The SDF process is assumed to be the empirical SDF process defined as

$$\mathcal{M}_{t+1} = \exp\Big(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}^{\mathbb{P}}\Big),$$

where λ_t is the n-dimensional market prices of risk and $v_{t+1}^{\mathbb{P}}$ an n-dimensional random vector that is standard normally distributed under the physical measure given the information up to time t. Owing to Girsanov's theorem, we put $v_{t+1}^{\mathbb{Q}}$, the n-dimensional random vector that is standard normally distributed under the risk-neutral measure, as

$$v_{t+1}^{\mathbb{Q}} = \lambda_t + v_{t+1}^{\mathbb{P}}.$$

Unless stated otherwise, all expectations at time t are taken conditional on the information up to time t [2].

[1] There, the model was formulated in continuous time, but we opt to exposit in discrete time.

[2] Formally, we assume there is a filtration $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ such that the σ-algebra $\mathcal{F}_t$ represents the information up to time t. The conditional expectation $\mathbb{E}_t[\cdot]$ is equivalent to the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}_t]$.

Term structure models are comprised of four components:

1) **The Short Rate Dynamics**
This dictates how the short rate is determined as a function of f_t. Usually, the short rate dynamics are written as

$$r_t = g(f_t)$$

for some function $g(\cdot)$.

2) **The $\mathbb{Q}$-(risk-neutral) Dynamics**
This dictates the dynamics of the factors f_t under the risk-neutral measure. Usually, the risk-neutral dynamics are written as

$$f_{t+1} = \mu_t^{\mathbb{Q}} + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}.$$

Here, $\mu_t^{\mathbb{Q}}$ is the conditional mean of f_{t+1} under the risk-neutral measure and $\Sigma_t\Sigma_t'$ the conditional variance.

3) **The $\mathbb{P}$-(physical) Dynamics**
This dictates the dynamics of the factors f_t under the physical measure. Usually, the physical dynamics are written as

$$f_{t+1} = \mu_t^{\mathbb{P}} + \Sigma_t \cdot v_{t+1}^{\mathbb{P}}.$$

Note that $\mu_t^{\mathbb{P}}$ is the conditional mean of f_{t+1} unde the physical measure and, since $v_{t+1}^{\mathbb{P}}$ and $v_{t+1}^{\mathbb{Q}}$ differ only by location, $\Sigma_t\Sigma_t'$ is still the conditional variance of f_{t+1}.

4) **The Specification for the Market Prices of Risk λ_t**
This dictates how the market prices of risk are determined as functions of the factors f_t. Formally, we assume there exists a function $\Lambda_t(\cdot)$ such that

$$\lambda_t = \Lambda_t(f_t).$$

In practice, we need only specify the short rate dynamics and two of the latter three

I have stated this for the sake of completeness; unless you have a good grasp on measure theory at this stage, I advise you to ignore this technical detail.

components. To see why, note that the following relationship holds:

$$\begin{aligned}\mathbb{E}_t\left[f_{t+1}\right] &= \mu_t^{\mathbb{P}} \\ &= \mathbb{E}_t\left[\mu_t^{\mathbb{Q}} + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}\right] = \mu_t^{\mathbb{Q}} + \Sigma_t \cdot \lambda_t.\end{aligned}$$

Therefore, the market prices of risk are given as functions of Σ_t, $\mu_t^{\mathbb{Q}}$ and $\mu_t^{\mathbb{P}}$ as follows:

$$\lambda_t = \Sigma_t^{-1}\left(\mu_t^{\mathbb{P}} - \mu_t^{\mathbb{Q}}\right).$$

If we specify the risk-neutral and physical dynamics, which means that we specify Σ_t, $\mu_t^{\mathbb{Q}}$ and $\mu_t^{\mathbb{P}}$, then we automatically obtain an expression for the market prices of risk λ_t. On the other hand, if we specify the market prices of risk and the risk-neutral dynamics, then because $\mu_t^{\mathbb{P}}$ is given as

$$\mu_t^{\mathbb{P}} = \mu_t^{\mathbb{Q}} + \Sigma_t \cdot \lambda_t,$$

and the variance of the factors is the same under either measure, we automatically obtain the physical factor dynamics. The same goes for the case where we specify the physical dynamics and the market prices of risk.

In most cases, we choose to specify the risk-neutral and physical dynamics. However, in some cases, most notably when we do not want to involve the risk-neutral measure, we can choose to specify the market prices of risk and the physical dynaimcs instead.

4.1 Defintion of Affine Term Structure Models

Affine term structure models (ATSM) are term structure models in which the short rate r_t, the risk-neutral conditional mean $\mu_t^{\mathbb{Q}}$, and the conditional variance $\Sigma_t\Sigma_t'$ are affine functions of the factors f_t. Specifically, in general affine term structure models the short rate, risk-netural dynamics and the conditional variance are typically given as follows[3]:

$$\begin{aligned}r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}\end{aligned}$$

[3]This is the specification chosen in Dai and Singleton (2000), who characterize affine term structure models depending on the number of factors n and the rank m of $\mathcal{B} = (\beta_1, \cdots, \beta_n)$, or the dependence of the conditional variance on the factors. This complicated topic is not broached here.

$$\Sigma_t = \Sigma \underbrace{\begin{pmatrix} \sqrt{\alpha_1 + \beta_1' f_t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\alpha_n + \beta_n' f_t}. \end{pmatrix}}_{\sqrt{S_t}}$$

We will see below that, under the above specification, yields are also given as affine functions of the factors, hence the name "affine" term structure model.

The number of lags included in the model may, of course, be larger than 1. Note that the $n \times n$ matrix Σ governs the dependence of the factors on each other, and that the conditional variance of the ith normalized factors is given as an affine function $\alpha_i + \beta_i' f_t$ of the factors. It follows that, for the term structure model to make sense, each $\alpha_i + \beta_i' f_t$ must be non-negative and Σ must be non-singular. An affine term structure model that satisfies these conditions is called **admissible**.

An affine term structure model in which the conditional variance $\Sigma_t \Sigma_t'$ is time-invariant is called a **Gaussian affine term structure model (GATSM)**. In light of the setting above, GATSMs are special cases of ATSMs in which $\beta_i = O_{n\times 1}$ and $\alpha_i = 1$ for any $1 \leq i \leq n$. Thus, a GATSM is admissible if and only if Σ is nonsingular. Most of the time we take Σ to be the Cholesky factor of the conditional variance, which we assume to be positive definite.

The objective of a term structure model is to recover bond prices $P_t(\tau)$, yields $Y_t(\tau)$, and various forward rates and risk premia related to zero-coupon bonds, as functions of the underlying factors f_t. In the sections that follow, we first study how to solve for bond prices in affine term structure models. This step involves only the risk-neutral measure. Afterward, we incorporate the physical measure to solve for bond risk premia, and briefly discuss possible specifications for market prices ofirisk. Finally, we study the various identification issues that arise from the presence of latent factors, and possible methods to uniquely identify the model.

4.2 Solving for Bond Prices

Consider an ATSM with short rate dynamics and risk-neutral dynamics given by

$$r_t = \delta + \beta' f_t \tag{4.1}$$

and

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}. \tag{4.2}$$

The conditional variance is given as

$$\Sigma_t \Sigma_t' = \Sigma \begin{pmatrix} \alpha_1 + \beta_1' f_t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_n + \beta_n' f_t, \end{pmatrix} \Sigma',$$

and we assume that the ATSM in question is admissible.

In Duffie and Kan (1996), it is shown that, in an affine term structure model, bond prices $P_t(\tau)$ must be given as an exponential-affine function of the factors f_t. That is, $P_t(\tau)$ must assume the form

$$P_t(\tau) = \exp\Big(-a(\tau) - b(\tau)' f_t\Big), \tag{4.3}$$

where $a(\cdot)$ and $b(\cdot)$ are functions of the time to maturity τ. To solve for bond prices means to find $a(\cdot)$ and $b(\cdot)$. Note that we already have the initial values for these functions; since $P_t(0) = 1$, it must be the case that

$$a(0) = 0 \quad \text{and} \quad b(0) = O_{n\times 1}.$$

The values of $a(\cdot)$ and $b(\cdot)$ for maturities greater than 0 can be found by means of the no-arbitrage equation. Specifically, recall that an asset's prices is equal to its expected discounted payoff. Since a zero-coupon bond's time $t+1$ payoff is simply its price at time $t+1$, the no-arbitrage equation becomes

$$P_t(\tau) = \mathbb{E}_t^{\mathbb{Q}} \left[\exp(-r_t) \cdot P_{t+1}(\tau - 1)\right] \tag{4.4}$$

for a bond with τ period left to maturity at time t. The fact that the complex SDF is replaced by the much simpler expression $\exp(-r_t)$ is one of the reasons we use the risk-neutral version of the no-arbitrage equation.

Substituting the bond price formula (4.3) into the equation (4.4) gives us

$$\exp\Big(-a(\tau) - b(\tau)' f_t\Big) = \cdot \mathbb{E}_t^{\mathbb{Q}} \left[\exp\Big(-r_t - a(\tau - 1) - b(\tau - 1)' f_{t+1}\Big)\right].$$

Using the short rate and risk-neutral dynamics now gives us

$$\begin{aligned} \exp\Big(-a(\tau) - b(\tau)' f_t\Big) = \exp\Big(-\delta - \beta' f_t - a(\tau - 1) - b(\tau - 1)' K^{\mathbb{Q}} - b(\tau - 1)' G^{\mathbb{Q}} f_t\Big) \\ \times \mathbb{E}_t^{\mathbb{Q}} \left[\exp\Big(-b(\tau - 1)' \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}\Big)\right]. \end{aligned}$$

The formula for the MGF of normally distributed variables tells us that

$$\mathbb{E}_t^{\mathbb{Q}} \left[\exp\Big(-b(\tau - 1)' \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}\Big)\right] = \exp\Big(\frac{1}{2} b(\tau - 1)' \Sigma_t \Sigma_t' b(\tau - 1)\Big),$$

so we have the equation

$$\begin{aligned}&\exp\Big(-a(\tau)-b(\tau)'f_t\Big)\\&=\exp\Big(-\delta-\beta'f_t-a(\tau-1)-b(\tau-1)'K^{\mathbb{Q}}-b(\tau-1)'G^{\mathbb{Q}}f_t+\frac{1}{2}b(\tau-1)'\Sigma_t\Sigma_t'b(\tau-1)\Big).\end{aligned}$$

Taking logs on both sides yields

$$\begin{aligned}&-a(\tau)-b(\tau)'f_t\\&\quad=-\delta-\beta'f_t-a(\tau-1)-b(\tau-1)'K^{\mathbb{Q}}-b(\tau-1)'G^{\mathbb{Q}}f_t+\frac{1}{2}b(\tau-1)'\Sigma_t\Sigma_t'b(\tau-1)\\&\quad=-\delta-\beta'f_t-a(\tau-1)-b(\tau-1)'K^{\mathbb{Q}}-b(\tau-1)'G^{\mathbb{Q}}f_t\\&\qquad+\frac{1}{2}\sum_{i=1}^{n}(b(\tau-1)'\Sigma)_i^2\cdot\alpha_i+\frac{1}{2}\sum_{i=1}^{n}(b(\tau-1)'\Sigma)_i^2\cdot\beta_i'f_t,\end{aligned}$$

where $(b(\tau-1)'\Sigma)_i$ denotes the ith element of the n-dimensional row vector $b(\tau-1)'\Sigma$.

We now match intercepts terms with intercept terms and coefficient terms with coefficient terms to obtain the **Ricatti equations**

$$a(\tau)=\delta+a(\tau-1)+b(\tau-1)'K^{\mathbb{Q}}-\frac{1}{2}\sum_{i=1}^{n}(b(\tau-1)'\Sigma)_i^2\cdot\alpha_i \tag{4.5}$$

$$b(\tau)=\beta+G^{\mathbb{Q}\prime}b(\tau-1)-\frac{1}{2}\sum_{i=1}^{n}(b(\tau-1)'\Sigma)_i^2\cdot\beta_i. \tag{4.6}$$

These, together with the initial conditions $a(0)=0$ and $b(0)=O_{n\times1}$, allow us to recursively solve for bond prices.

Note that, in the GATSM case with $\beta_i=O_{n\times1}$ and $\alpha_i=1$ for any $1\le i\le n$, the Ricatti equations are reduced to

$$a(\tau)=\delta+a(\tau-1)+b(\tau-1)'K^{\mathbb{Q}}-\frac{1}{2}b(\tau-1)'\Sigma\Sigma'b(\tau-1) \tag{4.7}$$

$$b(\tau)=\beta+G^{\mathbb{Q}\prime}b(\tau-1). \tag{4.8}$$

The simplicity of the Ricatti equations in the GATSM case also means that we can obtain closed-form solutions for $a(\cdot)$ and $b(\cdot)$ through the use of the initial conditions $a(0)=0$ and $b(0)=O_{n\times1}$; these solutions can also be found in Hamilton and Wu (2012):

$$b(\tau)=\left[\sum_{j=0}^{\tau-1}\left(G^{\mathbb{Q}\prime}\right)^j\right]\beta$$

$$a(\tau) = \tau\delta + \left(\sum_{s=1}^{\tau-1} b(s)\right)' K^{\mathbb{Q}} - \frac{1}{2}\sum_{s=1}^{\tau-1} b(s)'\Sigma\Sigma' b(s).$$

Given $a(\cdot)$ and $b(\cdot)$, we can now see that, by definition, the yield on a zero-coupon bond with τ periods to maturity at time t is

$$Y_t(\tau) = -\frac{1}{\tau}\log(P_t(\tau)) = \underbrace{\frac{a(\tau)}{\tau}}_{\alpha(\tau)} + \underbrace{\frac{b(\tau)'}{\tau}}_{\beta(\tau)'} f_t.$$

In other words, in our model yields are affine functions of the factors f_t.

So far, we have relied only on the short rate dynamics and the risk-neutral dynamics to derive bond prices and yields as functions of the factors f_t. This tells us that, when it comes to pricing bonds, we can always work under a risk-neutral environment; the risk aversion of investors plays no part in the derivation of bond prices. However, once we start talking about risk premia, or the compensation investors demand for taking on risk, the physical dynamics, and by extension the market price of risk, become indispensable. This is because, when we talk of expected rates of return, the expectation is with respect to the physical measure, not the risk-neutral measure.

4.3 Bond Risk Premia

We now express bond risk premia in terms of the market price of risk λ_t and the model parameters. The core relationship exploited here is that of $v_{t+1}^{\mathbb{Q}}$ and $v_{t+1}^{\mathbb{P}}$, which are related by a location change

$$v_{t+1}^{\mathbb{P}} + \lambda_t = v_{t+1}^{\mathbb{Q}}.$$

4.3.1 One-Period Ahead Risk Premium

Recall that the one-period ahead risk premium is defined as

$$\begin{aligned}\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] &= \mathbb{E}_t\left[\log(P_{t+1}(\tau-1))\right] - \log(P_t(\tau)) - r_t \\ &= -a(\tau-1) - b(\tau-1)'\mathbb{E}_t\left[f_{t+1}\right] + a(\tau) + b(\tau)' f_t - \delta - \beta' f_t.\end{aligned}$$

The risk-neutral dynamics tell us that

$$\begin{aligned}\mathbb{E}_t\left[f_{t+1}\right] &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot \mathbb{E}_t\left[v_{t+1}^{\mathbb{Q}}\right] \\ &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot \lambda_t,\end{aligned}$$

while the recursive solutions to $a(\cdot)$ and $b(\cdot)$ reveal

$$\begin{aligned}a(\tau) - a(\tau-1) &= \delta + b(\tau-1)' K^{\mathbb{Q}} - \frac{1}{2}\sum_{i=1}^{n}\left(b(\tau-1)'\Sigma\right)_i^2 \cdot \alpha_i \\ b(\tau) - b(\tau-1)' G^{\mathbb{Q}} - \beta' &= -\frac{1}{2}\sum_{i=1}^{n}\left(b(\tau-1)'\Sigma\right)_i^2 \cdot \beta_i'.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] &= a(\tau) - a(\tau-1) + b(\tau)' f_t - b(\tau-1)'\mathbb{E}_t\left[f_{t+1}\right] - \delta - \beta' f_t \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left(b(\tau-1)'\Sigma\right)_i^2 \cdot \left(\alpha_i + \beta_i' f_t\right) - b(\tau-1)'\Sigma_t \cdot \lambda_t \\ &= -\frac{1}{2} b(\tau-1)'\Sigma_t\Sigma_t' b(\tau-1) - b(\tau-1)'\Sigma_t \cdot \lambda_t.\end{aligned}$$

The term

$$-\frac{1}{2} b(\tau-1)'\Sigma_t\Sigma_t' b(\tau-1)$$

is referred to as the **Jensen's Inequality term**, since it originates from the concavity of the exponential function. As in the case of log-normal returns in the C-CAPM, we tend to ignore the effect of this term and define the adjusted one-period ahead risk premium as

$$RP_{t,adj}^{(\tau)} := \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] + \frac{1}{2} b(\tau-1)'\Sigma_t\Sigma_t' b(\tau-1) = -b(\tau-1)'\Sigma_t \cdot \lambda_t.$$

The time t covariance of excess returns and the risk factors $v_{t+1}^{\mathbb{P}}$ is given as

$$\begin{aligned}\mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}^{\mathbb{P}}\right) &= \mathrm{Cov}_t\left(-b(\tau-1)' f_{t+1}, v_{t+1}^{\mathbb{P}}\right) \\ &= -b(\tau-1)'\Sigma_t \cdot \mathrm{Var}_t\left(v_{t+1}^{\mathbb{P}}\right),\end{aligned}$$

where we used the fact that, under the physical measure, f_{t+1} follows a VAR(1) process with conditional variance $\Sigma_t\Sigma_t'$ and innovation process $v_{t+1}^{\mathbb{P}}$. Now, the risk premium becomes

$$RP_{t,adj}^{(\tau)} = \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}^{\mathbb{P}}\right)\left(\mathrm{Var}_t\left(v_{t+1}^{\mathbb{P}}\right)\right)^{-1} \cdot \lambda_t.$$

As before, this adjusted risk premium is given as the product of two components: the beta term

$$\text{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}^{\mathbb{P}}\right)\left(\text{Var}_t\left(v_{t+1}^{\mathbb{P}}\right)\right)^{-1}$$

and the market price of risk λ_t.

4.3.2 The Term Premium

Here, we show that the term premium can be expressed as the time average of one-period expected excess returns across the life of a bond, which enables us to derive a (near) closed form solution for the term premium in terms of model parameters.

The term premium for a τ-maturity bond at time t is defined as

$$TP_t(\tau) = Y_t(\tau) - \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[r_{t+h}\right]$$

It is shown in Cochrane and Piazzesi (2008) that the term premium can also be expressed as the average time t expected excess return over the life of the bond, unadjusted for the Jensen's inequality term. To see this, note that

$$\begin{aligned} Y_t(\tau) = -\frac{1}{\tau}\log(P_t(\tau)) &= \frac{1}{\tau}\left[\log(P_{t+\tau}(0)) - \log(P_t(\tau))\right] \\ &= \frac{1}{\tau}\mathbb{E}_t\left[\log(P_{t+\tau}(0)) - \log(P_t(\tau))\right]. \end{aligned}$$

This can then be expressed as a telescoping sum as follows:

$$Y_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[\log(P_{t+h+1}(\tau-h-1)) - \log(P_{t+h}(\tau-h))\right],$$

so that the term premium is written as

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[\log(P_{t+h+1}(\tau-h-1)) - \log(P_{t+h}(\tau-h)) - r_{t+h}\right].$$

Note that each expression within the brackets on the right is the excess return from time $t+h$ to $t+h+1$ for a $\tau-h$-period bond, so that

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[exr_{t+h+1}^{(\tau-h)}\right] \tag{4.9}$$

Note that, by the law of iterated expectations,

$$\mathbb{E}_t\left[exr_{t+h+1}^{(\tau-h)}\right] = \mathbb{E}_t\left[\mathbb{E}_{t+h}\left[exr_{t+h+1}^{(\tau-h)}\right]\right]$$

for any $0 \leq h \leq \tau - 1$, so that the term premium has the equivalent expression

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[RP_{t+h}^{(\tau-h)}\right],$$

where each $RP_{t+h}^{(\tau-h)}$ is the risk premium unadjusted for the Jensen's inequality term, derived above as

$$RP_{t+h}^{(\tau-h)} = -b(\tau-h-1)'\Sigma_{t+h}\cdot\lambda_{t+h} - \frac{1}{2}b(\tau-h-1)'\Sigma_{t+h}\Sigma_{t+h}'b(\tau-h-1).$$

4.3.3 The Forward Risk Premium

We can likewise furnish an expression for the forward risk premium in terms of one-period ahead expected excess returns, using which it is expressed as a function of model parameters.

The h-period ahead forward risk premium at time t is defined as

$$FRP_t(h) = f_t^{(h)} - \mathbb{E}_t\left[r_{t+h}\right].$$

Cochrane and Piazzesi (2008) also show that the forward risk premium can be expressed as the sum of the differences in one-period ahead expected excess returns. To derive this expression, we use the fact that

$$f_t^{(h)} = r_{t+h} + r_{t,t+h}^{(h+1)} - r_{t,t+h}^{(h)},$$

and since $f_t^{(h)}$ is known at time t, taking time t expectations on both sides yields

$$FRP_t(h) = f_t^{(h)} - \mathbb{E}_t\left[r_{t+h}\right] = \mathbb{E}_t\left[r_{t,t+h}^{(h+1)} - r_{t,t+h}^{(h)}\right].$$

Using the definition of holding period returns, we can express each as a telescoping sum:

$$\begin{aligned}
FRP_t(h) &= \mathbb{E}_t\left[r_{t,t+h}^{(h+1)}\right] - \mathbb{E}_t\left[r_{t,t+h}^{(h)}\right] \\
&= \mathbb{E}_t\left[\log(P_{t+h}(1)) - \log(P_t(h+1))\right] - \mathbb{E}_t\left[\log(P_{t+h}(0)) - \log(P_t(h))\right] \\
&= \sum_{i=0}^{h-1}\mathbb{E}_t\left[\log(P_{t+i+1}(h-i)) - \log(P_{t+i}(h-i+1))\right]
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=0}^{h-1} \mathbb{E}_t \left[\log(P_{t+i+1}(h-i-1)) - \log(P_{t+i}(h-i)) \right] \\
&= \sum_{i=0}^{h-1} \mathbb{E}_t \left[r_{t+i+1}^{(h-i+1)} - r_{t+i+1}^{(h-i)} \right] \\
&= \sum_{i=0}^{h-1} \mathbb{E}_t \left[exr_{t+i+1}^{(h-i+1)} - exr_{t+i+1}^{(h-i)} \right],
\end{aligned}
$$

where the last equality follows from adding and subtracting r_{t+i+1} for each $0 \leq i \leq h-1$. We have thus obtained the expression

$$
FRP_t(h) = \sum_{i=0}^{h-1} \mathbb{E}_t \left[exr_{t+i+1}^{(h-i+1)} - exr_{t+i+1}^{(h-i)} \right]. \tag{4.10}
$$

As with the term premium, the law of iterated expectations tells us that

$$
\begin{aligned}
\mathbb{E}_t \left[exr_{t+i+1}^{(h-i+1)} - exr_{t+i+1}^{(h-i)} \right] &= \mathbb{E}_t \left[\mathbb{E}_{t+i} \left[exr_{t+i+1}^{(h-i+1)} - exr_{t+i+1}^{(h-i)} \right] \right] \\
&= \mathbb{E}_t \left[RP_{t+i}^{(h-i+1)} - RP_{t+i}^{(h-i)} \right].
\end{aligned}
$$

Therefore, the forward term premium is in terms of the risk premium and Jensen's inequality terms as

$$
FRP_t(h) = \sum_{i=0}^{h-1} \mathbb{E}_t \left[RP_{t+i}^{(h-i+1)} - RP_{t+i}^{(h-i)} \right].
$$

4.3.4 Equivalence of Expectation Hypotheses

The expressions

$$
\begin{aligned}
TP_t(\tau) &= \frac{1}{\tau} \sum_{h=0}^{\tau-1} \mathbb{E}_t \left[RP_{t+h}^{(\tau-h)} \right] \\
FRP_t(h) &= \sum_{i=0}^{h-1} \mathbb{E}_t \left[RP_{t+i}^{(h-i+1)} - RP_{t+i}^{(h-i)} \right],
\end{aligned}
$$

also tell us that the three exectations hypotheses studied in section 3.1.2 are all equivalent. Recall that the EH assumes three forms:

$$
RP_t(\tau) = 0, \quad TP_t(\tau) = 0, \quad \text{and} \quad FRP_t(h) = 0
$$

for any t and τ. We can now show that the RP EH and TP EH imply the others:

- **RP EH implies TP EH and FRP EH**

Suppose that the RP EH, namely $RP_t(\tau) = 0$, holds. Then, since $TP_t(\tau)$ and $FP_t(h)$ are functions of the one-period ahead risk premium, both the term premium and forward risk premium are also 0, so that the TP EH and the FRP EH hold.

- **TP EH implies RP EH and FRP EH**
  Suppose that the TP EH, namely $TP_t(\tau) = 0$, holds. Putting $\tau = 1$, we can see that

$$TP_t(1) = \mathbb{E}_t\left[RP_t^{(1)}\right] = RP_t^{(1)} = 0.$$

  Suppose that $RP_t^{(\tau)} = 0$ for any $\tau \geq 1$ and t. Then, since

$$TP_t(\tau+1) = \frac{1}{\tau+1}\sum_{h=0}^{\tau}\mathbb{E}_t\left[RP_{t+h}^{(\tau+1-h)}\right] = \frac{1}{\tau+1}RP_t(\tau+1) = 0,$$

  implying that $RP_t(\tau+1) = 0$. Thus, by induction, the RP EH holds, and by the preceding result, the FRP EH holds as well.

- **FRP EH implies RP EH and TP EH**
  Suppose that the FRP EH, namely $FRP_t(h) = 0$, holds. Putting $h = 1$, and using the fact that $RP_t(1) = 0$ (the risk-premium of a risk-free asset is 0), we can see that

$$FRP_t(1) = RP_t(2) - RP_t(1) = RP_t(2) = 0.$$

  Continuing by induction as above shows us that the RP EH holds [4], so that the TP EH also holds.

By implication, if any one of the expectation hypotheses does not hold, then neither do the other two. In the presence of risk aversion, it is easy to show that the RP EH does not hold, so the TP EH and the FRP EH must also be violated.

4.4 Various Specifications for Market Prices of Risk

So far, we have only specified the short rate dynamics and risk-neutral dynamics of the model. To close out the model, we must either provide a specification for the physical dynamics of the model, or the market price of risk. Here, we study three popular forms of specifying λ_t; the first two specify λ_t directly, while the third first specifies the physical dynamics and then derives λ_t using the risk-neutral and physical dynamics. The last two

[4] Try this as an exercise.

approaches will be seen to be equivalent in Gaussian models. The exposition is heavily based on Duffee (2002) and Cheridito, Filipovic, and Kimmel (2007).

4.4.1 Completely Affine Models

The **completely affine** model of λ_t is the preferred specification in many of the earliest models of the term structure, including the models of Vasicek (1977) and the renowned CIR model (Cox, Ingersoll, and Ross (1985)), and was formalized in Dai and Singleton (2000). To understand the implications of this specification, we must first study Dai and Singleton's model, a generalization of the CIR model, in more depth.

Consider the short rate and risk-neutral factor dynamics of an admissible ATSM that belongs to the $\mathbb{A}_m(n)$ class, which indicates that there are n factors in the model whose volatility depends on $0 \leq m \leq n$ factors:

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}, \\ \Sigma_t &= \Sigma \cdot \operatorname{diag}\left(\sqrt{\alpha_1 + \beta_1' f_t}, \cdots, \sqrt{\alpha_n + \beta_n' f_t}\right). \end{aligned}$$

We say that the volatility of the factors depends on m factors in the sense that the rank of the matrix

$$\mathcal{B} = \begin{pmatrix} \beta_1 & \cdots & \beta_n \end{pmatrix}$$

is equal to m.

In a completely affine $\mathbb{A}_m(n)$ model, the market prices of risk are given as

$$\lambda_t = \Sigma_t \cdot \tilde{\lambda} = \begin{pmatrix} \sqrt{\alpha_1 + \beta_1' f_t} \cdot \tilde{\lambda}_1 \\ \vdots \\ \sqrt{\alpha_n + \beta_n' f_t} \cdot \tilde{\lambda}_n \end{pmatrix},$$

where $\tilde{\lambda}$ is an n-dimensional nonrandom vector. We can easily see that, from the following identity,

$$\mu_t^{\mathbb{P}} = \Sigma_t \cdot \lambda_t + \mu_t^{\mathbb{Q}}.$$

the physical factor dynamics are also affine in the factors f_t:

$$f_{t+1} = \underbrace{\left[K^{\mathbb{Q}} + \Sigma \cdot \begin{pmatrix} \alpha_1 \cdot \tilde{\lambda}_1 \\ \vdots \\ \alpha_n \cdot \tilde{\lambda}_n \end{pmatrix}\right]}_{K^{\mathbb{P}}} + \underbrace{\left[G^{\mathbb{Q}} + \Sigma \cdot \begin{pmatrix} \beta_1' \\ \vdots \\ \beta_n' \end{pmatrix}\right]}_{G^{\mathbb{P}}} \cdot f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{P}}.$$

Furthermore, we can easily see that the product $\lambda_t \lambda_t'$ is also affine in the factors. This is why this model is called "completely affine".

Affine physical factor dynamics, in particular, are useful because this means that the factors follow a VAR process under both the risk-netural and physical measures. As we primarily focus on the parameters pertaining to the physical dynamics of the factors during estimation, this considerably simplifies the estimation process. Dai and Singleton impose the following identification restrictions on the short rate and factor dynamics:

i) The elements of β are all non-negative.

ii) The first m factors $f_{1t}, \cdots, f_{mt}$ are non-negative.

iii) We have

$$\alpha_i = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ 1 & \text{if } m+1 \leq i \leq n \end{cases}.$$

iv) For any $1 \leq i \leq n$, β_i is the ith standard basis of $\mathbb{R}^n$, so that $\beta_i' f_t = f_{it}$, the ith factor.

For any $m+1 \leq i \leq n$, the last $n-m$ elements of each β_i are equal to 0, and its first m elements are all non-negative. β_i can be denoted

$$\beta_i = \begin{pmatrix} \tilde{\beta}_i \\ O_{(n-m)\times 1} \end{pmatrix}.$$

v) $G^{\mathbb{P}}$ is conformably partitioned as

$$G^{\mathbb{P}} = \begin{pmatrix} G_{11}^{\mathbb{P}} & O_{m\times(n-m)} \\ G_{21}^{\mathbb{P}} & G_{22}^{\mathbb{P}} \end{pmatrix}$$

when $m > 0$, and the elements of $G_{11}^{\mathbb{P}}$ are non-negative. When $m = 0$, $G^{\mathbb{P}}$ is lower triangular.

vi) The matrix Σ governing the cross correlations of the factors is given as

$$\Sigma = I_n.$$

vii) The last $n - m$ elements of $K^{\mathbb{P}}$ are equal to 0.

First, note that, under this normalization, the factor innovations are independent. Heuristically, we are assuming that the first m factors Granger cause the latter $n - m$ factors, but that the converse does not hold. This is one of two ways to impose restrictions on the correlations between the factors; we will discuss this matter in more detail in the next section on identification.

Since the first m factors are all non-negative and each $\tilde{\beta}_i$ is comprised wholly of non-negative elements, the model is admissible. Furthermore, as the first m factors approach 0, so does their conditional variance (the conditional variance of the last $n - m$ factors approach 1). The non-negativity restriction on $G_{11}^{\mathbb{P}}$ and the zero restriction on $G_{12}^{\mathbb{P}}$ is also imposed to keep the first m factors from becoming negative [5]. Dai and Singleton claim that this canonical model is maximal in the sense that they are the minimal possible restrictions that ensures admissibility (at least in the continuous time case) and econometric identification. In the next section, we show that the canonical model is identified against invariant affine transformations, but underidentified in the econometric sense.

Focusing now on the market price of risk, Duffee (2002) points out that λ_t has the following advantages and disadvantages:

i) **Advantage: Affine Physical Factor Dynamics**
The physical factor dynamics are affine in a completely affine model. The advantage that this confers was mentioned above.

ii) **Advantage: Continuity at 0**
In the completely affine model, compensation for risk goes to 0 as risk (=variance of the factors) goes to 0. This means that, in this model, bond prices are continuous at 0, which we saw earlier was one of the core assumptions needed for an SDF to

[5] In fact, in a model of continuous time, the restrictions on $G^{\mathbb{P}}$ are sufficient to ensure the non-negativity of the first m factors. In discrete time, we no longer have this luxury.

exist.

iii) **Disadvantage: Compensation Depends only on Factor Variance**
In a completely affine model, since λ_t is a function only of the factor variances, it means that the second moments of the factors contain all the necessary information on risk, which is an unrealistic assumption.

More concretely, Duffee shows that one of the main stylized facts concerning bond excess returns is that they are low on average but exhibit high volatility. In a completely affine model, bond excess returns can be kept low (=λ_t is kept low) if and only if factor variances are small, which means that bond excess returns must exhibit low volatility. This represents a failure of the completely affine model to replicate stylized facts of the yield curve.

iv) **Disadvantage: Each Price of Risk has a Fixed Sign**
Since the factor standard deviations are always non-negative, the sign of the ith market price of risk λ_{it} depends entirely on the sign of $\tilde{\lambda}_i$. This means that a positive price of risk must remain positive, and that a negative price of risk must remain negative, leading to a failure to replicate the stylized fact that bond excess returns often change signs (due to their low level but high volatility).

4.4.2 Essentially Affine Models

To address the failures of the completely affine model, Duffee proposes in his 2002 paper the **essentially affine** model of λ_t. The specification of the short rate dynamics and the risk-neutral dynamics are identical to the completely affine model; the only change is with the specification for the market prices of risk:

$$\lambda_t = \Sigma_t \tilde{\lambda} + \Sigma_t^- \Lambda \cdot f_t,$$

where $\tilde{\lambda}$ is an n-dimensional nonrandom vector, Λ is an $n \times n$ nonrandom matrix and Σ_t^- is an $n \times n$ diagonal random matrix such that its ith diagonal element $\Sigma_{t,ii}^-$ is given as

$$\Sigma_{t,ii}^- = \begin{cases} 0 & \text{if } 1 \leq i \leq m \\ \frac{1}{\sqrt{1+\beta_i' f_t}} & \text{if } m+1 \leq i \leq n \end{cases}.$$

The first part of this specification is identical to the completely affine model, but under the essentially affine specification the market prices of risk also depend on the factors f_t

independent of the conditional variances $\alpha_i + \beta_i' f_t$. This added flexibility means that the volatility of λ_t can be maintained at a high enough level while still keeping its level low, making up for the disadvantages of the completely affine model.

In addition, the essentially affine model retains the advantages of the completely affine specification. First, as the volatilities of the factors go to 0 the volatility of λ_t also goes to 0. The physical factor dynamics are also affine under the essentially affine model, which motivates the name "essentially" affine: to see this, we once again utilize the identity

$$\mu_t^{\mathbb{P}} = \mu_t^{\mathbb{Q}} + \Sigma_t \lambda_t.$$

Here,

$$\Sigma_t \lambda_t = \Sigma_t^2 \tilde{\lambda} + \begin{pmatrix} O_{m\times m} & O_{m\times(n-m)} \\ O_{(n-m)\times m} & I_{n-m} \end{pmatrix} \Lambda f_t,$$

which is an affine function of the factors, so $\mu_t^{\mathbb{P}}$ is also affine in f_t. Specifically, the physical factor dynamics are now given as

$$f_{t+1} = \left[K^{\mathbb{Q}} + \begin{pmatrix} O_{m\times 1} \\ \tilde{\lambda}_{m+1} \\ \vdots \\ \tilde{\lambda}_n \end{pmatrix} \right] + \left[G^{\mathbb{Q}} + \begin{pmatrix} \operatorname{diag}\left(\tilde{\lambda}_1, \cdots, \tilde{\lambda}_m\right) & O_{m\times(n-m)} \\ \begin{pmatrix} \tilde{\beta}_{m+1}' \cdot \tilde{\lambda}_{m+1} \\ \vdots \\ \tilde{\beta}_n' \cdot \tilde{\lambda}_n \end{pmatrix} & \Lambda_{22} \end{pmatrix} \right] f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{P}},$$

where Λ_{22} collects the $(n-m)\times(n-m)$ block matrix in the (2,2) position of Λ.

4.4.3 Extended Affine Models

The completely affine and essentially affine models choose to specify the market prices of risk first, and then derive affine physical factor dynamics based on the market price of risk specification. In contrast, the **extended affine** model introduced in Cheridito, Filipovic, and Kimmel (2007) first specifies affine physical factor dynamics, and then derives market prices of risk as a consequence of the risk-neutral and physical dynamics.

The extended affine model with n factors whose volatility depends on the first $0 \leq m \leq n$ factors is given as

$$r_t = \delta + \beta' f_t$$

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}$$

$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{Q}}$$

$$\Sigma_t = \Sigma \cdot \text{diag}\left(\sqrt{f_{1t}}, \cdots, \sqrt{f_{mt}}, \sqrt{\alpha_{m+1} + \beta'_{m+1} f_t}, \cdots, \sqrt{\alpha_n + \beta'_n f_t}\right),$$

where once again the elements of α_i and β_i are all non-negative, and the last $n-m$ elements of each β_i are equal to 0, to render the model admissible. In this model, the market prices of risk are given as

$$\begin{aligned}\lambda_t &= \Sigma_t^{-1}\left(\mu_t^{\mathbb{P}} - \mu_t^{\mathbb{Q}}\right) \\ &= \Sigma_t^{-1}\left(K^{\mathbb{P}} - K^{\mathbb{Q}}\right) + \Sigma_t^{-1}\left(G^{\mathbb{P}} - G^{\mathbb{Q}}\right) f_t.\end{aligned}$$

When $m = 0$, the essentially affine and extended affine models are identical, since Σ_t does not depend on the time subscript t and thus λ_t is an affine function of the factors f_t under both models. Since the case $m = 0$ corresponds to Gaussian ATSMs, in Gaussian models specifying affine market prices of risk and affine physical dynamics leads to the same model. Due to this equivalence, we often choose the essentially/extended affine specification for market prices of risk when working with GATSMs.

4.5 The Identification Problem

In the models investigated so far, bond prices and yields are determined by latent factors f_t that evolve dynamically according to a VAR specification. We are then faced with the problem: how do we identify the factors f_t? For an illustration, consider a Gaussian ATSM model with short rate dynamics and factor dynamics given as

$$r_t = \delta + \beta' f_t \tag{4.11}$$

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}} \tag{4.12}$$

$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}.. \tag{4.13}$$

Recall that the first two equations are sufficient to derive bond prices and yields, and that the third equation tells us that the factors follow a VAR(1) process; specifically, the time t yield of a τ-period bond is given as

$$Y_t(\tau) = \alpha(\tau) + \beta(\tau)' f_t \tag{4.14}$$

Now consider an **invariant affine transformation** of the factors f_t, that is, factors X_t defined as

$$X_t = A + B \cdot f_t$$

for some $A \in \mathbb{R}^n$ and a non-singular matrix $B \in \mathbb{R}^{n\times n}$. In terms of X_t, the short rate and risk-neutral dynamics are given as

$$r_t = \delta_X + \beta_X' f_t \tag{4.15}$$
$$X_{t+1} = K_X^{\mathbb{Q}} + G_X^{\mathbb{Q}} X_t + \Sigma_X \cdot v_{t+1}^{\mathbb{Q}} \tag{4.16}$$
$$X_{t+1} = K_X^{\mathbb{P}} + G_X^{\mathbb{P}} X_t + \Sigma_X \cdot v_{t+1}^{\mathbb{P}}, \tag{4.17}$$

where the new parameters are given as

$$\begin{aligned}
\delta_X &= \delta - \beta' B^{-1} A \\
\beta_X &= B^{-1\prime}\beta \\
K_X^i &= B(I_n - G^i)B^{-1}\cdot A + BK^i \quad \text{for any } i = \mathbb{P},\mathbb{Q} \\
G_X^i &= BG^iB^{-1} \quad \text{for any } i = \mathbb{P},\mathbb{Q} \\
\Sigma_X &= B\cdot\Sigma
\end{aligned}$$

for any $1 \le i \le n$. The bond pricing formula then tells us that there exist functions $\alpha_X(\cdot)$ and $\beta_X(\cdot)$ such that

$$Y_t(\tau) = \alpha_X(\tau) + \beta_X(\tau)' X_t. \tag{4.18}$$

Therefore, given the data on the yields $Y_t(\tau)$, we have no way to know whether the yields were generated given the latent factors f_t under equations (4.11), (4.12), (4.13) and (4.14), or if they were generated given the latent factors X_t under the equations (4.15), (4.16), (4.17) and (4.18). In the presence of such ambiguity, we say that the model is **underidentified**.

Our goal is to identify the model against invariant affine transformations, that is, to impose restrictions on the model parameters so that there are only one set of factors that generate the yields and simultaneously satisfy these restrictions. Mathematically, we want to impose restrictions on the model paramters so that, if

$$X_t = A + Bf_t$$

is another set of factors that satisfies the restrictions, then $A = O_{n\times 1}$ and $B = I_n$. Heuristically, identification is the process of reducing the number of free parameters as much as possible. In general, affine term structure models are identified in two steps:

1) **Equivalence of True Model to Canonical Form**
   Suppose the true short rate and factor dynamics are given as

$$r_t = \delta + \beta' f_t$$

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}$$

$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}.$$

Then, there exists an invariant affine transformation

$$X_t = A + B \cdot f_t$$

such that the short rate and factor dynamics formulated in terms of X_t satisfies the given identification restrictions. This form of the model is called the **canonical form** of the model.

2) **Uniqueness of Canonical Form**
We now show that the canonical form is identified against invariant affine transformations. Specifically, let X_t be factors under which the short rate and risk-neutral dynamics satisfy the identification restrictions. Then, if

$$Z_t = \mathcal{A} + \mathcal{B} \cdot X_t$$

is an affine transformations of X_t under which the short rate and risk-neutral dynamics also satisfy the identification restrictions, then $Z_t = X_t$.

In other words, even though the true factors may be f_t, we can only consistently estimate X_t, an affine rotation of the true factors; the fact that the X_t is a rotation of the true factors follows from the first step, and the consistent estimation is made possible by the second step. This is similar to how factor models are identified in works such as Bai (2003), where the PC estimators of the factors are shown to be consistent only for a rotation of the true factors.

Below we study two popular methods of identifying Gaussian ATSMs. Since the parameters $\delta, \beta, K^{\mathbb{Q}}, G^{\mathbb{Q}}$ and Σ govern how the yields are determined as functions of the factors[6]. Thus,during the identification process we impose restrictions on $\delta, \beta, K^{\mathbb{Q}}, G^{\mathbb{Q}}$ and Σ. Most practitioners choose to either impose the JSZ restrictions, named as such after the seminal work by Joslin, Singleton, and Zhu (2011), or the arbitrage-free Nelson-Siegel

[6]Alternatively, we can impose restrictions on the short rate and $\mathbb{P}$-dynamics of the model. If we impmose identification restrictions on the risk-neutral dynamics instead of the physical dynamics, then we can use our factor estimates to estimate $G^{\mathbb{P}}$ via unrestricted OLS. However, if we impose these restrictions on $G^{\mathbb{P}}$ instead, we must estimate a restricted version of the VAR equation

$$f_t = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma_t \cdot v_{t+1}^{\mathbb{P}},$$

which is decidedly more difficult than unrestricted OLS. This is one reason Joslin, Singleton, and Zhu (2011), and indeed many others, choose to impose restrictions on the $\mathbb{Q}$-dynamics over the $\mathbb{P}$-dynamics.

(AFNS) restrictions, first introduced in Christensen, Diebold, and Rudebusch (2011).

4.5.1 The Dai-Singleton Canonical Model

Before studying the JSZ and AFNS models, we first introduce the Dai-Singleton canonical model, introduced in Dai and Singleton (2000), which is one of the first models that attempted to impose identification restrictions on ATSMs. While the version of the model in Dai and Singleton (2000) imposes restrictions on the physical factor dynamics instead of the risk-neutral dynamics, to maintain consistency with the JSZ model that follows we instead study an equivalent version of the model that imposes restrictions on the risk-netural dynamics, which is studied in Singleton (2006).

Consider a Gaussian ATSM with short rate and risk-neutral dynamics[7] given by

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

As studied briefly above, Dai and Singleton impose the identification restrictions

i) The elements of β are all non-negative.

ii) $\Sigma = I_n$.

iii) $G^{\mathbb{Q}}$ is lower triangular with no eigenvalues equal to 1, that is, no unit roots. The diagonal entries of $G^{\mathbb{Q}}$ are also distinct and ordered in decreasing order.

iv) $K^{\mathbb{Q}} = O_{n\times 1}$.

It may come to your attention that the conditions that $G^{\mathbb{Q}}$ has no eigenvalues equal to 1 and that the diagonal entries of $G^{\mathbb{Q}}$ are distinct and ordered in decreasing order are not part of the original identification restrictions proposed in Singleton (2006). In Hamilton and Wu (2012), it is pointed out that, without these restrictions, we end up with an unidentified model. The specific way in which these additional restrictions help identify the model is shown in the appendix. It is also shown there that any Gaussian ATSM where $G^{\mathbb{Q}}$ has real and distinct eigenvalues withint the unit circle can undergo invariant affine transformations in a manner that satisfies the above constraints, and that a model that satisfies the above restrictions is identified against invariant affine transformations.

If the risk-neutral factor dynamics include a unit root, one way to identify the model may be to impose restrictions on the physical factor dynamics instead of the risk-neutral factor dynamics, that is, to restrict $K^{\mathbb{P}} = O_{n\times 1}$ and $G^{\mathbb{P}}$ lower triangular with no unit

[7]We only state the risk-neutral dynamics because the physical factor dynamics are left unrestricted and thus do not play a role in model identification.

roots and decreasing diagonal entries. However, if both the risk-neutral and physical factor dynamics contain unit roots, then we must impose additional zero restrictions on δ or β in order to identify the model. This is what we opt for in the FS-ZLB model.

Another problem that this identification scheme suffers from, also pointed out in Hamilton and Wu (2012), is that does not encompass ATSMs where $G^{\mathbb{Q}}$ or $G^{\mathbb{P}}$ have complex eigenvalues, since in this case there may not exist a decomposition $G^{\mathbb{Q}} = ULU'$ with real lower triangular L and real orthogonal U. We therefore study the JSZ model, which provides an alternative means of identifying Gaussian ATSMs that is robust under both the presence of a (single) unit root and complex eigenvalues in the mean reversion parameters.

4.5.2 The JSZ Model

Joslin, Singleton, and Zhu (2011) (henceforth JSZ) introduced a seminal identification scheme for Gaussian ATSMs. The generic GATSM framework consists of the following short rate and risk neutral dynamics[8]:

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where $\Sigma\Sigma'$ is the (positive definite and nonrandom) conditional variance of the factors. JSZ introduced the following restrictions to the model:

i) Σ is lower triangular; usually, it is taken to be the Cholesky factor of the conditional variance matrix.

ii) $\delta = 0$ and $\beta = \iota$, the n-dimensional vector consisting of 1s.

iii) $G^{\mathbb{Q}}$ is in ordered Jordan form, that is, it is a block diagonal matrix

$$G^{\mathbb{Q}} = \operatorname{diag}\left(J_1^{\mathbb{Q}}, \cdots, J_m^{\mathbb{Q}}\right)$$

where each block $J_i^{\mathbb{Q}}$ is a Jordan block

$$J_i^{\mathbb{Q}} = \begin{pmatrix} \lambda_i^{\mathbb{Q}} & 1 & \cdots & 0 & 0 \\ 0 & \lambda_i^{\mathbb{Q}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_i^{\mathbb{Q}} & 1 \\ 0 & 0 & \cdots & 0 & \lambda_i^{\mathbb{Q}} \end{pmatrix}$$

[8]We again only state the risk-neutral dynamics because the physical factor dynamics are left unrestricted and thus do not play a role in model identification.

for some real $\lambda_i^{\mathbb{Q}}$ such that $\left|\lambda_i^{\mathbb{Q}}\right| \leq 1$. The blocks are ordered so that $1 \geq \lambda_1^{\mathbb{Q}} > \cdots > \lambda_m^{\mathbb{Q}}$. We allow for the existence of at most a single unit root, or an eigenvalue equal to 1, and if a unit root exists, it is ordered as the first eigenvalue.

iv) The last $n-1$ entries of $K^{\mathbb{Q}}$ are equal to 0, so that

$$K^{\mathbb{Q}} = \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ O_{(n-1)\times 1} \end{pmatrix}.$$

JSZ show that, given any Gaussian ATSM with short rate and risk-neutral dynamics

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where $G^{\mathbb{Q}}$ contains eigenvalues on or within the unit circle and at most one unit root, there exists a unique observationally equivalent ATSM that satisfies the JSZ restrictions above.

In the appendix, we show that any ATSM where $G^{\mathbb{Q}}$ has real and distinct eigenvalues is observationally equivalent to an ATSM with short rate and risk-netural dynamics

$$r_t = \iota' X_t \tag{4.19}$$

$$X_{t+1} = \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \lambda_1^{\mathbb{Q}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{\mathbb{Q}} \end{pmatrix} \cdot X_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}, \tag{4.20}$$

where Σ is lower triangular with positive diagonal entries, and that this model is identified against invariant affine transformations. In other words, the risk-neutral dynamics can be summarized in terms of the $n+1+\frac{n(n+1)}{2}$ parameters

$$\theta^{\mathbb{Q}} = \{k_\infty^{\mathbb{Q}}, \lambda_1^{\mathbb{Q}}, \cdots, \lambda_n^{\mathbb{Q}}, \Sigma\}.$$

Solving for bond prices under the above canonical JSZ form allows us to formulate yields as affine functions of the factors X_t:

$$\mathcal{Y}_t = \mathcal{A}(\theta^{\mathbb{Q}}) + \mathcal{B}(\theta^{\mathbb{Q}}) \cdot X_t. \tag{4.21}$$

Specifically,

$$\mathcal{B}(\theta) = \begin{pmatrix} \frac{b(\tau_1)'}{\tau_1} \\ \vdots \\ \frac{b(\tau_m)'}{\tau_m} \end{pmatrix} = \begin{pmatrix} \frac{1}{\tau_1}\frac{1-(\lambda_1^{\mathbb{Q}})^{\tau_1}}{1-\lambda_1^{\mathbb{Q}}} & \cdots & \frac{1}{\tau_1}\frac{1-(\lambda_n^{\mathbb{Q}})^{\tau_1}}{1-\lambda_n^{\mathbb{Q}}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\tau_m}\frac{1-(\lambda_1^{\mathbb{Q}})^{\tau_m}}{1-\lambda_1^{\mathbb{Q}}} & \cdots & \frac{1}{\tau_m}\frac{1-(\lambda_n^{\mathbb{Q}})^{\tau_m}}{1-\lambda_n^{\mathbb{Q}}} \end{pmatrix} \tag{4.22}$$

and

$$\mathcal{A}(\theta) = \begin{pmatrix} \frac{a(\tau_1)}{\tau_1} \\ \vdots \\ \frac{a(\tau_m)}{\tau_m} \end{pmatrix} = \begin{pmatrix} \sum_{s=1}^{\tau_1-1}\left[k_\infty^{\mathbb{Q}}\frac{1-(\lambda_1^{\mathbb{Q}})^s}{1-\lambda_1^{\mathbb{Q}}} - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\Sigma\Sigma')_{ij}\frac{1-(\lambda_i^{\mathbb{Q}})^s}{1-\lambda_i^{\mathbb{Q}}}\frac{1-(\lambda_j^{\mathbb{Q}})^s}{1-\lambda_j^{\mathbb{Q}}}\right] \\ \vdots \\ \sum_{s=1}^{\tau_m-1}\left[k_\infty^{\mathbb{Q}}\frac{1-(\lambda_1^{\mathbb{Q}})^s}{1-\lambda_1^{\mathbb{Q}}} - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\Sigma\Sigma')_{ij}\frac{1-(\lambda_i^{\mathbb{Q}})^s}{1-\lambda_i^{\mathbb{Q}}}\frac{1-(\lambda_j^{\mathbb{Q}})^s}{1-\lambda_j^{\mathbb{Q}}}\right] \end{pmatrix} \tag{4.23}$$

where we define

$$\frac{1-\left(\lambda_i^{\mathbb{Q}}\right)^s}{1-\lambda_i^{\mathbb{Q}}} = s$$

for any $s \geq 1$ if $\lambda_i^{\mathbb{Q}} = 1$. Below we continue to work with this simpler model.

A Model with Observable Factors

The JSZ canonical factors X_t in the above model are latent factors. One of the main contributions of the JSZ model is that it allows us to derive an observationally equivalent model with observable factors. Suppose that the sample consists of yields of m maturities, collected into the m-dimensional random vector $\mathcal{Y}_t$, and that there are $n < m$ latent factors. In general, we assume the existence of measurement errors e_t that cause the observed yields, $\mathcal{Y}_t^o$, to deviate from the theoretical yields $\mathcal{Y}_t$:

$$\mathcal{Y}_t^o = \mathcal{Y}_t + e_t.$$

This allows us to write the model, together with the physical factor dynamics, as a state-space model and facilitate estimation. Instead of working with this general setup, suppose that there exists n portfolios of yields that are observed without error, that is, assume the existence of an $n \times m$ random matrix W and an n-dimensional random vector μ such that

$$\mathcal{P}_t = \mu + W\mathcal{Y}_t$$

is observed without error, or

$$W\mathcal{Y}_t^o = W\mathcal{Y}_t.$$

Below are two popular choices for this affine transformation:

1) **Principal Components**
One way to choose μ, W is in a manner that makes $\mathcal{P}_t$ the principal components of the demeaned yields. Formally, we would have

$$\mathcal{P}_t = W\left(\mathcal{Y}_t^o - \overline{\mathcal{Y}}\right),$$

where $\overline{\mathcal{Y}}$ is the sample mean of the yields and the rows of W are orthonormal eigenvectors corresponding to the n largest eigenvalues of the sample covariance matrix

$$\frac{1}{mT}\sum_{t=1}^{T}\left(\mathcal{Y}_t^o - \overline{\mathcal{Y}}\right)\left(\mathcal{Y}_t^o - \overline{\mathcal{Y}}\right)'.$$

Note that this requires the consistency of the sample covariance matrix for the true covariance matrix, or the weak stationarity and variance ergodicity of the yield process $\{\mathcal{Y}_t\}_{t\in\mathbb{Z}}$. Furthermore, the parameters μ, W do not depend on the model parameters in this case.

2) **Perfectly Priced Yields**
An alternative approach is to assume that some yields are perfectly priced, while others are observed with error. A popular choice is to arrange the yields so that the first n yields, collected in $\mathcal{Y}_t^1$, are priced without error, while the remaining $m-n$ yields, collected in $\mathcal{Y}_t^2$, are not. In this case, we would write

$$\begin{pmatrix}\mathcal{Y}_t^{1o}\\ \mathcal{Y}_t^{2o}\end{pmatrix} = \begin{pmatrix}\mathcal{Y}_t^{1}\\ \mathcal{Y}_t^{2}\end{pmatrix} + \begin{pmatrix}O_{n\times 1}\\ e_t\end{pmatrix}.$$

Letting $\mathcal{A}(\theta^{\mathbb{Q}})_1$ and $\mathcal{B}(\theta^{\mathbb{Q}})_1$ collect the first n rows of $\mathcal{A}(\theta^{\mathbb{Q}})$ and $\mathcal{B}(\theta^{\mathbb{Q}})$, it follows that

$$\mathcal{Y}_t^{1o} = \mathcal{Y}_t^1 = \mathcal{A}(\theta^{\mathbb{Q}})_1 + \mathcal{B}(\theta^{\mathbb{Q}})_1 \cdot X_t,$$

so that

$$X_t = \underbrace{-\mathcal{B}(\theta^{\mathbb{Q}})_1^{-1}\mathcal{A}(\theta^{\mathbb{Q}})_1}_{\mu} + \underbrace{\mathcal{B}(\theta^{\mathbb{Q}})_1^{-1}}_{W}\cdot\mathcal{Y}_t^{1o}.$$

The fact that

$$\mathcal{Y}_t^{1o} = \mathcal{Y}_t^1$$

in this setup means that the n portfolios X_t are observed without error, so that $\mathcal{P}_t = X_t$. Note that here, μ and W are functions of the parameters $\theta^{\mathbb{Q}}$.

This specification may be more restrictive than the PC one, since it requires some yields themselves, instead of a portfolio of yields, to be priced without error. On the other hand, it confers on the model added generality compared to the PC specification because the case of non-stationary yields can also be accomodated.

In either case, since

$$\mathcal{P}_t = \mu + W\mathcal{Y}_t = \left(\mu + W\mathcal{A}(\theta^{\mathbb{Q}})\right) + W\mathcal{B}(\theta^{\mathbb{Q}}) \cdot X_t, \tag{4.24}$$

$\mathcal{P}_t$ is an invariant affine transformation of the JSZ canonical factors X_t (provided that $W\mathcal{B}(\theta^{\mathbb{Q}})$ is nonsingular). It follows that an observationally equivalent Gaussian ATSM with factors $\mathcal{P}_t$ has short rate and factor dynamics

$$r_t = \delta_{\mathcal{P}} + \beta_{\mathcal{P}} \cdot \mathcal{P}_t \tag{4.25}$$

$$\mathcal{P}_{t+1} = K_{\mathcal{P}}^{\mathbb{Q}} + G_{\mathcal{P}}^{\mathbb{Q}} \cdot \mathcal{P}_t + \Sigma_{\mathcal{P}} \cdot v_{t+1}^{\mathbb{Q}}. \tag{4.26}$$

$$\mathcal{P}_{t+1} = K_{\mathcal{P}}^{\mathbb{P}} + G_{\mathcal{P}}^{\mathbb{P}} \cdot \mathcal{P}_t + \Sigma_{\mathcal{P}} \cdot v_{t+1}^{\mathbb{P}}. \tag{4.27}$$

where

$$\delta_{\mathcal{P}} = -\iota' \left(\mu + W\mathcal{A}(\theta^{\mathbb{Q}})\right) \tag{4.28}$$

$$\beta_{\mathcal{P}} = \left(\mathcal{B}(\theta^{\mathbb{Q}})'W'\right)^{-1} \iota \tag{4.29}$$

$$G_{\mathcal{P}}^{\mathbb{Q}} = \left(W\mathcal{B}(\theta^{\mathbb{Q}})\right) \begin{pmatrix} \lambda_1^{\mathbb{Q}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{\mathbb{Q}} \end{pmatrix} \left(W\mathcal{B}(\theta^{\mathbb{Q}})\right)^{-1} \tag{4.30}$$

$$K_{\mathcal{P}}^{\mathbb{Q}} = \left[I_n - G_{\mathcal{P}}^{\mathbb{Q}}\right] \left(\mu + W\mathcal{A}(\theta^{\mathbb{Q}})\right) + k_{\infty}^{\mathbb{Q}} \cdot W\mathcal{B}(\theta^{\mathbb{Q}})_1 \tag{4.31}$$

$$\Sigma_{\mathcal{P}} = \left(W\mathcal{B}(\theta^{\mathbb{Q}})\right)\Sigma \tag{4.32}$$

and $\mathcal{B}(\theta^{\mathbb{Q}})_1$ is the first column of $\mathcal{B}(\theta^{\mathbb{Q}})$, while no restrictions are imposed on the parameters governing the physical factor dynamics.

In terms of $\mathcal{P}_t$, yields are given as

$$\mathcal{Y}_t = \underbrace{\mathcal{A}(\theta^{\mathbb{Q}}) - \mathcal{B}(\theta^{\mathbb{Q}})\left[W\mathcal{B}(\theta^{\mathbb{Q}})\right]^{-1}\left(\mu + \mathcal{A}(\theta^{\mathbb{Q}})\right)}_{\text{Intercept}} \tag{4.33}$$

$$+ \underbrace{\mathcal{B}(\theta^{\mathbb{Q}})\left[W\mathcal{B}(\theta^{\mathbb{Q}})\right]^{-1}}_{\text{Factor Loadings}} \cdot \mathcal{P}_t$$

Thus, under the assumption that the portfolios $\mathcal{P}_t$ are observed without error, JSZ shows us how we can formulate a Gaussian ATSM with factors $\mathcal{P}_t$ that depends on as few parameters as possible.

4.5.3 The AFNS Model

The **arbitrage-free Nelson-Siegel (AFNS)** model, presented in continuous time by Christensen, Diebold, and Rudebusch (2011) and adapted for discrete time by Niu and Zeng (2012), is a special case of the JSZ model that introduces no-arbitrage into the Nelson-Siegel model. Although the N-S model fits the yield curve very well, its main shortcoming is that it is an empirical model that does not impose the no-arbitrage condition. In its base form, the N-S model allows for arbitrage opportunities to arise, as we will see below. The AFNS model is an attempt to retain the advantages conferred by the N-S model, namely its excellent yield curve fit, while establishing a theoretical basis that the base model sorely lacks.

The AFNS model is a Gaussian ATSM first and foremost. It short rate and factor dynamics are given, as usual, as

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v^{\mathbb{Q}}_{t+1} \\ f_{t+1} &= K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v^{\mathbb{P}}_{t+1}. \end{aligned}$$

Under this specification, we saw earlier that bond prices are given as exponential-affine functions of the factors

$$P_t(\tau) = \exp\Big(-a(\tau) - b(\tau)' f_t\Big)$$

and as such that the yields are affine functions of the factors:

$$Y_t(\tau) = \frac{a(\tau)}{\tau} + \frac{b(\tau)'}{\tau} f_t. \tag{4.34}$$

Like its namesake, in the AFNS model the factor loadings $\frac{b(\tau)}{\tau}$ are given as the N-S loadings, that is,

$$\frac{b(\tau)'}{\tau} = \begin{pmatrix} 1 & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} & \frac{1-\exp(-\tau\kappa)}{\tau\kappa} - \exp(-\tau\kappa) \end{pmatrix}$$

for some decay parameter κ. This allows us to identify the three factors in f_t as the level L_t, slope S_t, and curvature C_t, as we did in the base N-S model. Since

$$\frac{b(\tau)}{\tau} = \frac{1}{\tau}\left[\sum_{j=0}^{\tau-1}\left(G^{\mathbb{Q}\prime}\right)^j\right]\beta,$$

the question now is whether $G^{\mathbb{Q}}$ and β can be chosen so that $\frac{b(\tau)}{\tau}$ actually assumes the N-S factor loading form above. Fortunately, it is shown in Niu and Zeng (2012) that specifying

$$\beta' = \begin{pmatrix} 1 & \frac{1-\exp(-\kappa)}{\kappa} & \frac{1-\exp(-\kappa)}{\kappa} - \exp(-\kappa) \end{pmatrix}$$

and

$$G^{\mathbb{Q}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \exp(-\kappa) & 1 \\ 0 & 0 & \exp(-\kappa) \end{pmatrix}$$

actually does lead to $\frac{b(\tau)}{\tau}$ being given as the N-S factor loadings.

Choosing to impose no restrictions on the physical factor dynamics on the model means that $K^{\mathbb{P}}$ is not restricted. In other words, given that $G^{\mathbb{P}}$ has eigenvalues within the unit circle (or in other words, that the factors are stationary), the factors f_t are allowed to have non-zero means. This means that we can restrict the intercept term in both the short rate and risk-neutral dynamics to be zero[9]. Thus, the short rate and risk-neutral factor dynamics of the AFNS model are given as

$$r_t = \begin{pmatrix} 1 & \frac{1-\exp(-\kappa)}{\kappa} & \frac{1-\exp(-\kappa)}{\kappa} - \exp(-\kappa) \end{pmatrix} f_t \tag{4.35}$$

$$f_{t+1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \exp(-\kappa) & 1 \\ 0 & 0 & \exp(-\kappa) \end{pmatrix} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}} \tag{4.36}$$

In light of the JSZ model above, if Σ is lower triangular, then the AFNS model can be

[9]In practice, we often choose to instead restrict the intercept $K^{\mathbb{P}}$ of the factors under the physical measure instead of $K^{\mathbb{Q}}$, especially when the factors are non-stationary under the physical measure. This matter will be dealt in more detail in the next section.

taken to be a restricted version of the JSZ canonical model with

$$k_\infty^{\mathbb{Q}} = 0$$
$$\lambda^{\mathbb{Q}} = (1, \exp(-\kappa), \exp(-\kappa)).$$

The AFNS model boasts a different specification for how the factors load on the short rate; nevertheless, the risk-neutral dynamics depend on fewer parameters compared to the canonical JSZ model. By implication, since the JSZ canonical model is identified against invariant affine transformations, so must the AFNS model.

Note how, in equation (4.34), there is an intercept term $\frac{a(\tau)}{\tau}$ in addition to the factor loadings and factors found in the base N-S model. This intercept term is a by-product of introducing no-arbitrage into the mix, and suggests that the base N-S model, which lacks an intercept, does not satisfy the no-arbitrage condition. Heuristically, the intercept term can be interpreted as a correction for yields of various maturities that prevents arbitrage opportunities from arising.

Compared to the identification schemes we studied earlier, the AFNS model takes a slightly different approach. Instead of starting with an arbitrary Gaussian ATSM and showing that the true latent factors can be appropriately transformed to obtain an observationally equivalent but identified ATSM, the AFNS model assumes from the outset that the true model is the one given by (4.35) and (4.36). In this context, the fact that the AFNS model is identified against invariant affine transformations indicates that we can consistently estimate the actual level, slope and curvature factors, instead of just a rotation of them.

This fact was actually seen in the earlier section on estimating the N-S model. Instead of the consistency results in works such as Bai and Ng (2002) in which the estimated factors are consistent only for a rotation of the true factors, it was shown in that section that the least squares estimator of the N-S factors is consistent for the actual N-S factors. This ease of identification, as well as the clear role that each N-S factor plays, is one of the many reasons practitioners use the AFNS model over the more robust JSZ canonical model.

4.6 Estimating Gaussian ATSMs

Here we introduce some notable approaches to estimating Gaussian ATSMs using linear regressions. Consider a general Gaussian ATSM with short rate and factor dynamics given as

$$r_t = \delta + \beta' f_t$$

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}$$
$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}.$$

As we will no doubt be familiar with by now, the affine short rate and risk-neutral dynamics imply that the yields are given as an affine function of the factors:

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B} \cdot f_t,$$

where $\mathcal{A}$ and $\mathcal{B}$ are functions of the $\mathbb{Q}$-parameters

$$\theta^{\mathbb{Q}} = \{\delta, \beta, K^{\mathbb{Q}}, G^{\mathbb{Q}}, \Sigma\}.$$

In general, we assume that there exists a measurement error e_t for the yields so that the model can be written in state space form as:

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B} \cdot f_t + \Sigma_e \cdot e_t \tag{4.37}$$
$$f_t = K^{\mathbb{P}} + G^{\mathbb{P}} \cdot f_{t-1} + \Sigma \cdot v_t^{\mathbb{P}}, \tag{4.38}$$

where $\Sigma_e \Sigma_e'$ is the covariance matrix of the measurement errors. A fully rigorous estimation of the model parameters would involve Gaussian MLE or Bayesian estimation via the Metropolis-Hastings algorithm, which uses the Kalman filter and smoother, or Carter and Kohn's backward recursion, to recover the factors. Due to the high irregularity of the likelihood function, as well as the large number of parameters present in the model, this proves difficult and overly complicated in practice.

Therefore, in practice we take inspirations from estimation methods such as Doz, Giannone, and Reichlin (2011) and assume that the factors are observable functions of the data. Often, they are given as the principal components of the yields, or as linear combinations of the yields based on the absence of measurement errors for yields of some maturities. Below we introduce three popular methods of estimating Gaussian ATSMs under the assumption of observable factors. In all three methods, the factors are taken to be affine transformations of the yields.

4.6.1 Joslin, Singleton and Zhu (JSZ)

We start by studying how the likelihood is derived in the JSZ model, and how it allows for a clear separation between parameters related to the risk-neutral and physical dynamics of the model. In particular, it allows for the estimation of some parameters via OLS, which greatly alleviates the computational burden of maximizing a likelihood function or sampling via the Metropolis-Hastings algorithm. For this reason, this is the preferred

method of estimation in works such as Bauer and Rudebusch (2016) and Bauer and Rudebusch (2020).

Suppose we have a Gaussian ATSM with observable factors $\mathcal{P}_t$ that are given as an affine transformation of $\mathcal{Y}_t$ as

$$\mathcal{P}_t = \mu + W\mathcal{Y}_t.$$

We assume these are observed without error in the sense that $W\mathcal{Y}_t = W\mathcal{Y}_t^o$. Below, we omit the superscript o and let all yields that appear be the observed values.

Let the short rate and factor dynamics be given as in equations (4.25) to (4.27), under which the yields are given as an affine function of the factors:

$$\mathcal{Y}_t = A\left(\theta^{\mathbb{Q}}\right) + B\left(\theta^{\mathbb{Q}}\right) \cdot \mathcal{P}_t, \tag{4.39}$$

where $A(\cdot)$ and $B(\cdot)$ are functions of the $\mathbb{Q}$-parameters

$$\theta^{\mathbb{Q}} = \{k_\infty^{\mathbb{Q}}, \lambda_1^{\mathbb{Q}}, \cdots, \lambda_n^{\mathbb{Q}}, \Sigma\}$$

in the manner specified in equations (4.23), (4.29) and (4.21).

Consider the simplest case, where the state-space form of the model is given as

$$\begin{aligned}\mathcal{Y}_t &= A\left(\theta^{\mathbb{Q}}\right) + B\left(\theta^{\mathbb{Q}}\right) \cdot f_t + \Sigma_e \cdot e_t \\ f_t &= K^{\mathbb{P}} + G^{\mathbb{P}} f_{t-1} + \Sigma \cdot v_t^{\mathbb{P}},\end{aligned}$$

with yield measurement errors e_t, which have covariance matrix $\Sigma_e \Sigma_e'$. In addition, assume the processes $\{e_t\}_{t\in\mathbb{Z}}$ and $\{v_t^{\mathbb{P}}\}_{t\in\mathbb{Z}}$ are i.i.d. standard normally distributed. In addition, suppose that e_t is independent of all leads and lags of f_t, while $v_t^{\mathbb{P}}$ is independent of all lags of f_t. Letting θ collect all the model parameters, and denoting the yield data by $\mathcal{Y}$, the log-likelihood function can then be decomposed as[10]

$$l(\mathcal{Y} \mid \theta) = \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{F}_{t-1}, \theta),$$

where $\mathcal{F}_{t-1}$ is the information up to time $t-1$ and $f(\cdot)$ denotes densities. Note that the factors $\mathcal{P}_t$ are assumed to be an observable affine transformation of the yields $\mathcal{Y}_t$; it follows that the density of $\mathcal{P}_t$ given $\mathcal{Y}_t$ is just the point mass at $\mathcal{P}_t$, so that

$$f(\mathcal{P}_t \mid \mathcal{Y}_t, \mathcal{F}_{t-1}, \theta) = 1.$$

[10]We are assuming that the initial values of the data and factors are given, so that this log likelihood function is technically a conditional likelihood function.

By Bayes' rule, we can then see that

$$1 = f(\mathcal{P}_t \mid \mathcal{Y}_t, \mathcal{F}_{t-1}, \theta) = \frac{f(\mathcal{Y}_t \mid \mathcal{P}_t, \mathcal{F}_{t-1}, \theta) \cdot f(\mathcal{P}_t \mid \mathcal{F}_{t-1}, \theta)}{f(\mathcal{Y}_t \mid \mathcal{F}_{t-1}, \theta)}.$$

Furthermore, $\mathcal{P}_{t-1}$ is contained in the information set $\mathcal{F}_{t-1}$ and the conditional distribution of $\mathcal{P}_t$ given $\mathcal{F}_{t-1}$ depends only on $\mathcal{P}_{t-1}$. Similarly, the conditional distribution of $\mathcal{Y}_t$ given $\mathcal{P}_t, \mathcal{F}_{t-1}$ depends only on $\mathcal{P}_t$. As such, we can see that

$$f(\mathcal{Y}_t \mid \mathcal{F}_{t-1}, \theta) = f(\mathcal{Y}_t \mid \mathcal{P}_t, \theta) \cdot f(\mathcal{P}_t \mid \mathcal{P}_{t-1}, \theta).$$

Here, the conditional distribution of $\mathcal{P}_t$ given $\mathcal{P}_{t-1}$ is governed by the $\mathbb{P}$-dynamic parameters

$$K_{\mathcal{P}}^{\mathbb{P}}, G_{\mathcal{P}}^{\mathbb{P}}, \Sigma,$$

while the conditional distribution of $\mathcal{Y}_t$ given $\mathcal{P}_t$ depends only on the $\mathbb{Q}$-dynamic parameters

$$k_{\infty}^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma.$$

The log-likelihood can thus be written as

$$\begin{aligned} l(\mathcal{Y} \mid \theta) = &\sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{P}_t; k_{\infty}^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma) \\ &+ \sum_{t=1}^{T} \log f(\mathcal{P}_t \mid \mathcal{P}_{t-1}; K_{\mathcal{P}}^{\mathbb{P}}, G_{\mathcal{P}}^{\mathbb{P}}, \Sigma) \end{aligned} \tag{4.40}$$

In other words, if we specify the risk-neutral and factor dynamics, then there is a clean break in the log-likelihood between the $\mathbb{Q}$-parameters and $\mathbb{P}$-parameters. In particular, note that $K_{\mathcal{P}}^{\mathbb{P}}$ and $G_{\mathcal{P}}^{\mathbb{P}}$ only appear in the transition equation log-likelihood, which is quadratic in these parameters. This suggests that the MLEs of these parameters can be recovered via OLS, indicating that we need only maximize the log-likelihood with respect to the parameters $k_{\infty}^{\mathbb{Q}}, \lambda^{\mathbb{Q}}$ and Σ once we concentrate out $K_{\mathcal{P}}^{\mathbb{P}}$ and $G_{\mathcal{P}}^{\mathbb{P}}$. JSZ also suggest using the OLS estimator of $\Sigma\Sigma'$ recovered from the transition equation as an initial value for numerical optimization.

The model can also be easily estimated via Bayesian methods using the decomposed likelihood derived above. Since $K_{\mathcal{P}}^{\mathbb{P}}$ and $G_{\mathcal{P}}^{\mathbb{P}}$ only appear in the log-likelihood of the transition equation, we can Gibbs sample these parameters as we would a seemingly unrelated regression (SUR) model. Furthermore, because the log-likelihood of the measurement equation is linear in $k_{\infty}^{\mathbb{Q}}$, it can also be Gibbs sampled. The remaining parameters $\lambda^{\mathbb{Q}}$

and Σ can now be sampled via the Metropolis-Hastings algorithm; in some cases, we even Gibbs sample Σ from the log-likelihood of the transition equation under the assumption that the dependence of the measurement equation on Σ is not too high.

Since the AFNS model is also a special case of the JSZ model with a unit root in the risk-neutral dynamics and two identical eigenvalues, its estimation can also proceed based on the log-likelihood derived above.

4.6.2 Hamilton and Wu (HW)

Hamilton and Wu (2012) exploits the fact that Gaussian ATSMs are observationally equivalent to a restricted Gaussian VAR when some yields are observed without error. This leads them to propose a two-step method where the reduced form restricted Gaussian VAR parameters are first estimated, and afterward the structural parameters are backed out from the reduced form parameters via a GMM-type method.

Suppose that the yields are ordered so that the first n yields are observed without error. In this case, the model can be written in the state-space form

$$\begin{aligned} \mathcal{Y}_t^o &= \mathcal{A} + \mathcal{B} \cdot f_t + \begin{pmatrix} O_{n\times 1} \\ \Sigma_e \cdot e_t \end{pmatrix} \\ f_{t+1} &= K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}, \end{aligned}$$

where $\Sigma_e \Sigma_e'$ is the covariance matrix associated with the measurement errors for the last $m-n$ yields. Below, we omit the superscript o and let all yields that appear be the observed values.

Conformably partition $\mathcal{Y}_t$, $\mathcal{A}$ and $\mathcal{B}$ as

$$\mathcal{Y}_t = \begin{pmatrix} \mathcal{Y}_t^1 \\ \mathcal{Y}_t^2 \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} \underbrace{\mathcal{A}_1}_{n\times 1} \\ \underbrace{\mathcal{A}_2}_{(m-n)\times 1} \end{pmatrix}, \quad \begin{pmatrix} \underbrace{\mathcal{B}_1}_{n\times n} \\ \underbrace{\mathcal{B}_2}_{(m-n)\times n} \end{pmatrix}.$$

Then, the factors are affine functions of the observed yields:

$$f_t = \mathcal{B}_1^{-1} \left(\mathcal{Y}_t - \mathcal{A}_1 \right),$$

so that, using the physical factor dynamics, we are able to recover the dynamics of the first n yields $\mathcal{Y}_t^1$:

$$\begin{aligned} \mathcal{Y}_t^1 &= \mathcal{A}_1 + \mathcal{B}_1 \cdot f_t \\ &= \mathcal{A}_1 + \mathcal{B}_1 K^{\mathbb{P}} + \mathcal{B}_1 G^{\mathbb{P}} \cdot f_{t-1} + \mathcal{B}_1 \Sigma \cdot v_t^{\mathbb{P}} \end{aligned}$$

$$= \left[\left(I_n - \mathcal{B}_1 G^{\mathbb{P}} \mathcal{B}_1^{-1}\right)\mathcal{A}_1 + \mathcal{B}_1 K^{\mathbb{P}}\right] + \mathcal{B}_1 G^{\mathbb{P}} \mathcal{B}_1^{-1} \cdot \mathcal{Y}_{t-1}^1 + \mathcal{B}_1 \Sigma \cdot v_t^{\mathbb{P}}.$$

Furthermore, the measurement equation tells us that the remaining $(m-n)$ yields are determined as an affine function of the first n yields:

$$\begin{aligned}\mathcal{Y}_t^2 &= \mathcal{A}_2 + \mathcal{B}_2 \cdot f_t + \Sigma_e \cdot e_t \\ &= \left[\mathcal{A}_2 - \mathcal{B}_2 \mathcal{B}_1^{-1} \mathcal{A}_1\right] + \mathcal{B}_2 \mathcal{B}_1^{-1} \cdot \mathcal{Y}_t^1 + \Sigma_e \cdot e_t.\end{aligned}$$

The log-likelihood function is given as

$$\begin{aligned}l(\mathcal{Y} \mid \theta) &= \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{F}_{t-1}, \theta) \\ &= \sum_{t=1}^{T} \log f(\mathcal{Y}_t^2 \mid \mathcal{Y}_t^1, \mathcal{F}_{t-1}, \theta) + \sum_{t=1}^{T} \log f(\mathcal{Y}_t^1 \mid \mathcal{F}_{t-1}, \theta).\end{aligned}$$

where $\mathcal{F}_{t-1}$ is the information contained in the yields up to time $t-1$. Under the usual assumption of i.i.d. Gaussian errors and their backward looking exogeneity, the conditional distribution of $\mathcal{Y}_t^2$ given $\mathcal{Y}_t^1$ and $\mathcal{F}_{t-1}$ depends only on $\mathcal{Y}_t^1$, and likewise, the conditional distribution of $\mathcal{Y}_t^1$ given the past yields depends only on $\mathcal{Y}_{t-1}^1$. Thus,

$$l(\mathcal{Y} \mid \theta) = \sum_{t=1}^{T} \log f(\mathcal{Y}_t^2 \mid \mathcal{Y}_t^1, \theta) + \sum_{t=1}^{T} \log f(\mathcal{Y}_t^1 \mid \mathcal{Y}_{t-1}, \theta). \tag{4.41}$$

Hamilton and Wu make the observation that, because the conditional distributions above are Gaussian, the ATSM is observationally equivalent to the reduced form restricted Gaussian VAR

$$\mathcal{Y}_t^1 = A_1^* + \Phi_1^* \cdot \mathcal{Y}_{t-1} + H_1^* \cdot v_t^{\mathbb{P}} \tag{4.42}$$

$$\mathcal{Y}_t^2 = A_2^* + \Phi_2^* \cdot \mathcal{Y}_t^1 + \Sigma_e \cdot e_t, \tag{4.43}$$

where

$$A_1^* = \left(I_n - \mathcal{B}_1 G^{\mathbb{P}} \mathcal{B}_1^{-1}\right)\mathcal{A}_1 + \mathcal{B}_1 K^{\mathbb{P}} \tag{4.44}$$

$$\Phi_1^* = \mathcal{B}_1 G^{\mathbb{P}} \mathcal{B}_1^{-1} \tag{4.45}$$

$$\Omega_1^* := H_1^* H_1^{*\prime} = \mathcal{B}_1 \Sigma \Sigma' \mathcal{B}_1' \tag{4.46}$$

$$A_2^* = \mathcal{A}_2 - \mathcal{B}_2 \mathcal{B}_1^{-1} \mathcal{A}_1 \tag{4.47}$$

$$\Phi_2^* = \mathcal{B}_2 \mathcal{B}_1^{-1}. \tag{4.48}$$

The model is identified in the econometric sense if and only if we can recover the structural

parameters

$$\delta, \beta, K^{\mathbb{Q}}, G^{\mathbb{Q}}, \Sigma, K^{\mathbb{P}}, G^{\mathbb{P}}$$

as unique functions of the reduced form parameters

$$A_1^*, \Phi_1^*, A_2^*, \Phi_2^*, H_1^* H_1^{*\prime}.$$

To this end, we may impose identifying restrictions on the model; candidates include any of the identification schemes introduced above (Dai-Singleton, JSZ, AFNS).

Once we have imposed the identification restrictions, Hamilton and Wu propose the following minimum chi-square estimation (MCSE) approach to estimate the structural parameters:

Step 1: Estimation of the Reduced-Form Parameters

We first estimate the reduced form parameters $A_1^*, \Phi_1^*, \Omega_1^*, A_2^*, \Phi_2^*, \Sigma_e \Sigma_e'$ via OLS estimation of the restricted VAR system.

Step 2: Recovering the Structural Parameters

Suppose we impose identification constraints that involve $K^{\mathbb{P}} = O_{n\times 1}$. We solve for the parameters in $G^{\mathbb{Q}}$, β, δ and $K^{\mathbb{Q}}$ by minimizing the distance

$$\left\|\Phi_2^* - \mathcal{B}_2\mathcal{B}_1^{-1}\right\|^2 + \left\|\Omega_1^* - \mathcal{B}_1\Sigma\Sigma'\mathcal{B}_1'\right\|^2 + \left|\mathcal{A}_1 - (I_n - \Phi_1^*)\, A_1^*\right|^2 + \left|\mathcal{A}_2 - A_2^* - \mathcal{B}_2\mathcal{B}_1^{-1}\mathcal{A}_1\right|^2.$$

Then, we obtain $G^{\mathbb{P}}$ as

$$G^{\mathbb{P}} = \mathcal{B}_1^{-1}\Phi_1^*\mathcal{B}_1,$$

where we evaluate $\mathcal{B}_1$ using the estimates obtained in the preceding stage.

Hamilton and Wu claim that this two-step approach to estimation yields consistent estimates of the parameters while alleviating the computational burden of directly maximizing the log-likelihood, required in, say, the JSZ model.

The method that appears in the second step of the estimation procedure is referred to as minimum chi-square estimation. In the appendix, its namesake, as well as the reason it yields consistent estimates of the parameters, is explained. Hamilton and Wu (2012) also show that under the appropriate choices of weights, the MCSEs of the structural parameters are as efficient as their full-information MLEs, that is, the MLEs obtained from maximizing the log-likelihood directly.

4.6.3 Adrian, Crump and Moench (ACM)

In Adrian, Crump, and Moench (2013) (henceforth ACM), an alternative means of estimating Gaussian ATSMs is proposed, which requires only linear regressions. So far, we have studied ATSMs in which the risk-neutral dynamics and form of the market price of risk were first specified (e.g. Dai and Singleton (2000)), and those where the risk-neutral and physical dynamics were first specified (e.g. Joslin, Singleton, and Zhu (2011)). In constrast, ACM first specify the phyiscal dynamics and the form of the market price of risk. Subsequently, they derive linear relationships between bond excess returns on the one hand and the market price of risk and risk factors on the other. This allows us to consistently recover the parameters related to the $\mathbb{P}$-dynamics and the market price of risk via OLS. Finally, the short rate dynamics are estimated via OLS and the bond price formula is calculated. Below we study how ACM formulate their Gaussian ATSM and how they estimate the model paramters through a simple three-step estimation procedure.

The starting point is the physical factor dynamics, given as

$$f_{t+1} = K + Gf_t + \Sigma \cdot v_{t+1}, \tag{4.49}$$

where v_{t+1} follows an n-dimensional standard normal distribution conditional on past information under the physical measure. The short rate dynamics are given as

$$r_t = \delta_0 + \delta_1' f_t + u_t^{(1)} \tag{4.50}$$

where $\delta_0 + \delta_1' f_t$ is the orthogonal projection of r_t on the space spanned by $\{1, f_{1t}, \cdots, f_{nt}\}$ with respect to the L^2 norm. By the definition of projections,

$$\begin{aligned} \mathbb{E}\left[u_t^{(1)}\right] &= 0, \\ \mathbb{E}\left[f_t \cdot u_t^{(1)}\right] &= O_{n\times 1}. \end{aligned}$$

In contrast to the usual short rate dynamics, the one in the ACM model explicitly includes a measurement error $u_t^{(1)}$ that is uncorrelated with the factors f_t.

Under the assumption of no-arbitrage there exists an SDF process $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ with $\mathcal{M}_0 = 1$ such that

$$P_t(\tau) = \mathbb{E}_t\left[\mathcal{M}_{t+1} P_{t+1}(\tau - 1)\right].$$

As usual, we assume the form of the SDF if given as

$$\mathcal{M}_{t+1} = \exp\left(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}\right)$$

for an n-dimensional vector λ_t of market prices of risk. Following the essentially affine and extended affine models, ACM assume that market prices of risk are affine functions of the factors:

$$\lambda_t = \Sigma^{-1}(\lambda + \Lambda f_t). \tag{4.51}$$

By definition, the one-period excess bond return for a τ-maturity bond from time t to time $t+1$ is given as

$$exr_{t+1}^{(\tau)} = \log P_{t+1}(\tau-1) - \log P_t(\tau) - r_t = \log\left(\frac{P_{t+1}(\tau-1)}{P_t(\tau)}\exp(-r_t)\right),$$

so that

$$\exp\left(exr_{t+1}^{(\tau)}\right) = \frac{P_{t+1}(\tau-1)}{P_t(\tau)}\exp(-r_t).$$

By the no-arbitrage condition,

$$\begin{aligned}
1 &= \mathbb{E}_t\left[\frac{P_{t+1}(\tau-1)}{P_t(\tau)}\mathcal{M}_{t+1}\right] \\
&= \mathbb{E}_t\left[\frac{P_{t+1}(\tau-1)}{P_t(\tau)}\exp(-r_t)\exp\left(-\frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}\right)\right] \\
&= \mathbb{E}_t\left[\exp\left(exr_{t+1}^{(\tau)} - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}\right)\right] \\
&= \exp\left(-\frac{1}{2}\lambda_t'\lambda_t\right)\cdot\mathbb{E}_t\left[\exp\left(\begin{pmatrix}1 & -\lambda_t'\end{pmatrix}\begin{pmatrix}exr_{t+1}^{(\tau)} \\ v_{t+1}\end{pmatrix}\right)\right].
\end{aligned}$$

Suppose that $exr_{t+1}^{(\tau)}$ and v_{t+1} are jointly normally distributed given the information up to time t. Then, the formula for the MGF of normally distributed random vectors tells us that

$$1 = \exp\left(-\frac{1}{2}\lambda_t'\lambda_t + \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] + \frac{1}{2}\begin{pmatrix}1 & -\lambda_t'\end{pmatrix}\begin{pmatrix}\mathrm{Var}_t\left(exr_{t+1}^{(\tau)}\right) & \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right) \\ \mathrm{Cov}_t\left(v_{t+1}, exr_{t+1}^{(\tau)}\right) & I_n\end{pmatrix}\begin{pmatrix}1 \\ -\lambda_t\end{pmatrix}\right)$$

$$= \exp\left(\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] + \frac{1}{2}\mathrm{Var}_t\left(exr_{t+1}^{(\tau)}\right) - \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\cdot\lambda_t\right).$$

The affine specification for the market prices of risk tells us that

$$\begin{aligned}
\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] &= \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\cdot\lambda_t - \frac{1}{2}\mathrm{Var}_t\left(exr_{t+1}^{(\tau)}\right) \\
&= \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\Sigma^{-1}(\lambda + \Lambda f_t) - \frac{1}{2}\mathrm{Var}_t\left(exr_{t+1}^{(\tau)}\right).
\end{aligned}$$

In other words, the expected part of the one period ahead excess bond returns is an affine function of the factors at time t, with an additional convexity term. Defining

$$\beta_t^{(\tau)\prime} = \text{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\Sigma^{-1},$$

we can write

$$\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] = \beta_t^{(\tau)\prime}(\lambda + \Lambda f_t) - \frac{1}{2}\text{Var}_t\left(exr_{t+1}^{(\tau)}\right).$$

Now note that we can decompose the forecasting error as follows:

$$exr_{t+1}^{(\tau)} - \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] = \beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1} + e_{t+1}^{(\tau)},$$

where $e_{t+1}^{(\tau)}$ is defined as the difference between the forecasting error and $\beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1}$. Its time t conditional mean is 0, and since

$$\begin{aligned}
\mathbb{E}_t\left[e_{t+1}^{(\tau)} v_{t+1}'\Sigma'\beta_t^{(\tau)}\right] &= \mathbb{E}_t\left[\left(exr_{t+1}^{(\tau)} - \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] - \beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1}\right) v_{t+1}'\right] \cdot \Sigma'\beta_t^{(\tau)} \\
&= \text{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\Sigma'\beta_t^{(\tau)} - \beta_t^{(\tau)\prime}\Sigma\Sigma' \cdot \beta_t^{(\tau)} \\
&= \beta_t^{(\tau)\prime}\Sigma\Sigma'\beta_t^{(\tau)} - \beta_t^{(\tau)\prime}\Sigma\Sigma' \cdot \beta_t^{(\tau)} = 0,
\end{aligned}$$

$e_{t+1}^{(\tau)}$ and $\beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1}$ are uncorrelated conditional on time t information.

So far, we have shown that bond excess returns can be decomposed as follows:

$$\begin{aligned}
exr_{t+1}^{(\tau)} = \quad & \underbrace{\beta_t^{(\tau)\prime}(\lambda + \Lambda f_t)}_{\text{Expected Excess Return}} \quad \underbrace{- \frac{1}{2}\text{Var}_t\left(exr_{t+1}^{(\tau)}\right)}_{\text{Convexity Term}} \\
& + \quad \underbrace{\beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1}}_{\text{Rate of Return Innovation}} \quad + \quad \underbrace{e_{t+1}^{(\tau)}}_{\text{Return Pricing Error}}.
\end{aligned}$$

The rate of return innovation is the part of the forecasting error that is explained by the risk factors v_{t+1}, while the return pricing error is the part of the forecasting error that cannot be explained by v_{t+1}. In particular, since $e_{t+1}^{(\tau)}$ is uncorrelated with the part explained by v_{t+1}, it can be treated as an idiosyncratic measurement error.

Suppose $\{e_t^{(\tau)}\}_{t\in\mathbb{Z}}$ has constant variance σ^2. Under this assumption, we can see that

$$\begin{aligned}
\text{Var}_t\left(exr_{t+1}^{(\tau)}\right) &= \text{Var}_t\left(\beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1} + e_{t+1}^{(\tau)}\right) \\
&= \beta_t^{(\tau)\prime}\Sigma\Sigma'\beta_t^{(\tau)} + \sigma^2
\end{aligned}$$

because v_{t+1} and $e_{t+1}^{(\tau)}$ are uncorrelated. Excess bond returns are given as

$$\begin{aligned} exr_{t+1}^{(\tau)} = \beta_t^{(\tau)\prime}(\lambda + \Lambda f_t) - \frac{1}{2}\left(\beta_t^{(\tau)\prime}\Sigma\Sigma'\beta_t^{(\tau)} + \sigma^2\right) \\ + \beta_t^{(\tau)\prime}\Sigma \cdot v_{t+1} + e_{t+1}^{(\tau)}. \end{aligned}$$

A final assumption we make is that the beta term $\beta_t^{(\tau)}$ is constant over time; in other words, bond excess returns respond in the same manner to changes in market prices of risk regardless of the time. This can be seen as part of an attempt to identify the role of the beta term and the market price of risk, where the former serves as the factor loading and latter as the factors that explain bond excess returns. The beta term is also time-invariant in the usual Gaussian ATSM, which indicates that this assumption is not too restrictive. Under this additional simplification, bond excess returns are finally given as

$$exr_{t+1}^{(\tau)} = \beta^{(\tau)\prime}(\lambda + \Lambda f_t) - \frac{1}{2}\left(\beta^{(\tau)\prime}\Sigma\Sigma'\beta^{(\tau)} + \sigma^2\right) + \beta^{(\tau)\prime}\Sigma \cdot v_{t+1} + e_{t+1}^{(\tau)}. \tag{4.52}$$

Suppose the sample consists of yields of m maturities. For any time t, stacking equation (4.52) for each of these maturities shows us that

$$\begin{aligned} exr_{t+1} = \begin{pmatrix} exr_{t+1}^{(\tau_1)} \\ \vdots \\ exr_{t+1}^{(\tau_m)} \end{pmatrix} = \begin{pmatrix} \beta^{(\tau_1)\prime} \\ \vdots \\ \beta^{(\tau_m)\prime} \end{pmatrix}(\lambda + \Lambda f_t) - \frac{1}{2}\begin{pmatrix} \operatorname{vec}\left(\beta^{(\tau_1)}\beta^{(\tau_1)\prime}\right)' \\ \vdots \\ \operatorname{vec}\left(\beta^{(\tau_m)}\beta^{(\tau_m)\prime}\right)' \end{pmatrix} \cdot \operatorname{vec}\left(\Sigma\Sigma'\right) \\ - \frac{1}{2}\sigma^2 \cdot \iota_m + \begin{pmatrix} \beta^{(\tau_1)\prime} \\ \vdots \\ \beta^{(\tau_m)\prime} \end{pmatrix}\Sigma \cdot v_{t+1} + \underbrace{\begin{pmatrix} e_{t+1}^{(\tau_1)} \\ \vdots \\ e_{t+1}^{(\tau_m)} \end{pmatrix}}_{e_{t+1}}, \end{aligned}$$

where we used the fact that

$$\begin{aligned} \beta^{(\tau)\prime}\Sigma\Sigma'\beta^{(\tau)} &= \operatorname{vec}\left(\beta^{(\tau)\prime}\Sigma\Sigma'\beta^{(\tau)}\right) \\ &= \left(\beta^{(\tau)} \bigotimes \beta^{(\tau)}\right)' \cdot \operatorname{vec}\left(\Sigma\Sigma'\right) \\ &= \operatorname{vec}\left(\beta^{(\tau)}\beta^{(\tau)\prime}\right)' \cdot \operatorname{vec}\left(\Sigma\Sigma'\right). \end{aligned}$$

Define

$$\beta = \begin{pmatrix} \beta^{(\tau_1)\prime} \\ \vdots \\ \beta^{(\tau_m)\prime} \end{pmatrix} \in \mathbb{R}^{m \times n} \quad \text{and} \quad B = \begin{pmatrix} \operatorname{vec}\left(\beta^{(\tau_1)}\beta^{(\tau_1)\prime}\right)' \\ \vdots \\ \operatorname{vec}\left(\beta^{(\tau_m)}\beta^{(\tau_m)\prime}\right)' \end{pmatrix} \in \mathbb{R}^{m \times n^2}.$$

Then, we can see that

$$exr_{t+1} = \left[\beta \cdot \lambda - \frac{1}{2}\left(B \cdot \text{vec}\left(\Sigma\Sigma'\right) + \sigma^2 \cdot \iota_m\right)\right] + \beta\Lambda \cdot f_t + \beta\Sigma \cdot v_{t+1} + e_{t+1},$$

and collecting all the observations into the matrix

$$exr = \begin{pmatrix} exr_1' \\ \vdots \\ exr_T' \end{pmatrix} = \iota_T \cdot \underbrace{\left[\lambda' \cdot \beta' - \frac{1}{2}\left(\text{vec}\left(\Sigma\Sigma'\right)' B' + \sigma^2 \iota_m'\right)\right]}_{\mathbf{a}'} + F_{-1} \cdot \underbrace{\Lambda'\beta'}_{\mathbf{b}'} + V \cdot \underbrace{\Sigma'\beta'}_{\mathbf{c}'} + E$$

$$= \begin{pmatrix} \iota_T & F_{-1} & V \end{pmatrix} \begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \\ \mathbf{c}' \end{pmatrix} + E,$$

where

$$F = \begin{pmatrix} f_1' \\ \vdots \\ f_T' \end{pmatrix}, \quad F_{-1} = \begin{pmatrix} f_0' \\ \vdots \\ f_{T-1}' \end{pmatrix}, \quad V = \begin{pmatrix} v_1' \\ \vdots \\ v_T' \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} e_1' \\ \vdots \\ e_T' \end{pmatrix}.$$

The excess bond return regression

$$exr_{t+1}^{(\tau)} = \beta^{(\tau)\prime}(\lambda + \Lambda f_t) - \frac{1}{2}\left(\beta^{(\tau)\prime}\Sigma\Sigma'\beta^{(\tau)} + \sigma^2\right) + \beta^{(\tau)\prime}\Sigma \cdot v_{t+1} + e_{t+1}^{(\tau)}$$

allows us to price bonds in the ACM model. Since the $\mathbb{Q}$-dynamics are affine under affine market prices of risk and VAR(1) physical dynamics, and the short rate dynamics are also affine with an added pricing error term $u_t^{(1)}$, bond prices are determined as

$$P_t(\tau) = \exp\Big(a(\tau) + b(\tau)' f_t + \tau \cdot u_t^{(\tau)}\Big),$$

which is exactly the exponential-affine form found in usual ATSMs, except for the added error term $u_t^{(\tau)}$. Thus, in the ACM model, the yield pricing errors are baked into the foundations of the model.

From the definition of excess bond returns, we can see that

$$exr_{t+1}^{(\tau)} = a(\tau-1) + b(\tau-1)' f_{t+1} + (\tau-1)u_{t+1}^{(\tau-1)} - a(\tau) - b(\tau)' f_t - \tau \cdot u_t^{(\tau)} - \delta_0 - \delta_1' f_t - u_t^{(1)}$$

$$= \left(a(\tau-1) - a(\tau) + b(\tau-1)'K - \delta_0\right) + \left(b(\tau-1)'G - b(\tau)' - \delta_1'\right) f_t$$
$$+ b(\tau-1)'\Sigma \cdot v_{t+1} + \left((\tau-1)u_{t+1}^{(\tau-1)} - \tau \cdot u_t^{(\tau)} - u_t^{(1)}\right).$$

Matching terms between the above equation and equation (4.52) shows us that $\beta^{(\tau)} = b(\tau-1)$, and thus that

$$a(\tau) = a(\tau-1) + b(\tau-1)'(K-\lambda) + \frac{1}{2}\left(b(\tau-1)'\Sigma\Sigma' b(\tau-1) + \sigma^2\right) - \delta_0 \tag{4.53}$$

$$b(\tau)' = b(\tau-1)'(G-\Lambda) - \delta_1' \tag{4.54}$$

with initial values $a(0) = 0$ and $b(0) = O_{n\times 1}$. The additional term σ^2 contained in the first equation comes from the presence of the yield pricing errors in the bond pricing formula.

In addition, the following relationship must hold between yield pricing errors and return pricing errors:

$$(\tau-1)u_{t+1}^{(\tau-1)} - \tau \cdot u_t^{(\tau)} - u_t^{(1)} = e_{t+1}^{(\tau)}.$$

In other words, if the yield pricing errors are serially uncorrelated, then the return pricing errors must be serially correlated. ACM point out that this is an undesriable implication, so they assume serially uncorrelated return pricing errors instead.

As in the usual Gaussian ATSM, yields are given as affine functions of the factors:

$$Y_t(\tau) = -\frac{a(\tau)}{\tau} - \frac{b(\tau)'}{\tau} f_t + u_t^{(\tau)}, \tag{4.55}$$

and our measurement equation can be derived by stacking these for various maturities.

The excess bond formula (4.52) also allows for the following three-step estimation procedure. Recall that the parameters we must estimate are

$$\underbrace{\{\delta_0, \delta_1\}}_{\text{Short Rate Dynamics}} \qquad \underbrace{\{K, G, \Sigma\Sigma'\}}_{\text{Physical Dynamics}} \qquad \underbrace{\{\lambda, \Lambda\}}_{\text{Market Prices of Risk}}$$

Step 1: **Estimating $\mathbb{P}$-Dynamic and Short Rate Parameters**

As in JSZ or HW, the ACM model assumes observable factors f_t, constructed either through principal components or the use of macroeconomic variables. Given these observable factors, the first step of the ACM method involves estimating the parameters K, G and $\Omega := \Sigma\Sigma'$ via OLS. Specifically, the estimates are given as

$$\begin{pmatrix} \hat{K} & \hat{G} \end{pmatrix} = F' \begin{pmatrix} \iota_T & F_{-1} \end{pmatrix} \left[\begin{pmatrix} \iota_T' \\ F_{-1}' \end{pmatrix} \begin{pmatrix} \iota_T & F_{-1} \end{pmatrix} \right]^{-1} \tag{4.56}$$

and

$$\hat{\Omega} = \frac{1}{T} \left(F - \iota_T \cdot \hat{K}' - F_{-1}\hat{G}' \right)' \left(F - \iota_T \cdot \hat{K}' - F_{-1}\hat{G}' \right). \tag{4.57}$$

with $\hat{\Sigma}$ being the Cholesky factor of $\hat{\Omega}$.

In addition, we can estimate the short rate parameters δ_0, δ_1 via OLS:

$$\begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \end{pmatrix} = \left[\begin{pmatrix} \iota_T' \\ F' \end{pmatrix} \begin{pmatrix} \iota_T & F \end{pmatrix} \right]^{-1} \begin{pmatrix} \iota_T' \\ F' \end{pmatrix} r, \tag{4.58}$$

where we define $r = (r_1, \cdots, r_T)'$.

Step 2: Estimating the Excess Bond Return Regression

Using the estimates $\hat{K}$ and $\hat{G}$ procured above, we can estimate the factor innovations as

$$\hat{v}_t = \hat{\Sigma}^{-1} \left(f_t - \hat{K} - \hat{G} f_{t-1} \right),$$

and thus

$$\hat{V} = \begin{pmatrix} \hat{v}_1' \\ \vdots \\ \hat{v}_T' \end{pmatrix} = \left(F - \iota_T \cdot \hat{K}' - F_{-1}\hat{G}' \right) \hat{\Sigma}^{-1\prime}. \tag{4.59}$$

Now we turn to the excess bond return regression

$$exr = \begin{pmatrix} \iota_T & F_{-1} & V \end{pmatrix} \begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \\ \mathbf{c}' \end{pmatrix} + E.$$

While the regressors in V are unobservable, we can use the generated regressors $\hat{V}$ in their stead. Therefore, we can estimate the parameters $\mathbf{a}, \mathbf{b}, \mathbf{c}$ via OLS by regressing

excess bond returns on a constant, f_{t-1} and $\hat{v}_t$:

$$\begin{pmatrix} \hat{\mathbf{a}} & \hat{\mathbf{b}} & \hat{\mathbf{c}} \end{pmatrix} = exr' \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \left[\begin{pmatrix} \iota_T' \\ F_{-1}' \\ \hat{V}' \end{pmatrix} \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \right]^{-1}. \tag{4.60}$$

The return pricing error variance is then estimated as

$$\hat{\sigma}^2 = \frac{1}{mT} \operatorname{tr}\left(\hat{E}'\hat{E}\right), \tag{4.61}$$

where

$$\hat{E} = exr - \iota_T \cdot \hat{\mathbf{a}}' - F_{-1}\hat{\mathbf{b}}' + \hat{V} \cdot \hat{\mathbf{c}}'.$$

Step 3: Estimating the Market Price of Risk Parameters

Using the estimates obtained so far, we now back out the market price of risk parameters. Note that $\mathbf{a}$ and $\mathbf{b}$ are related to the market price of risk parameters λ and Λ as

$$\begin{aligned} \lambda &= \left(\beta'\beta\right)^{-1}\beta'\left[\mathbf{a} + \frac{1}{2}\left(B \cdot \operatorname{vec}(\Omega) + \sigma^2 \cdot \iota_m\right)\right] \\ \Lambda &= \left(\beta'\beta\right)^{-1}\beta'\mathbf{b}, \end{aligned}$$

where $\beta'\beta$ is nonsingular because it has full rank with $n < m$. Furthermore, β can be recovered from $\mathbf{c}$ via the relationship

$$\beta = \mathbf{c}\Sigma^{-1}.$$

Therefore, we estimate β as

$$\hat{\beta} = \hat{\mathbf{c}}\hat{\Sigma}^{-1}, \tag{4.62}$$

using which we can procure our estimate of B as

$$\hat{B} = \begin{pmatrix} \operatorname{vec}\left(\hat{\beta}_1\hat{\beta}_1\right)' \\ \vdots \\ \operatorname{vec}\left(\hat{\beta}_m\hat{\beta}_m\right)' \end{pmatrix}, \tag{4.63}$$

where $\hat{\beta}_i'$ is the ith row of $\hat{\beta}$. Then, we estimate the market price of risk parameters

as

$$\hat{\lambda} = \left(\hat{\beta}'\hat{\beta}\right)^{-1}\hat{\beta}'\left[\hat{\mathbf{a}} + \frac{1}{2}\left(\hat{B}\cdot\mathrm{vec}\left(\hat{\Omega}\right) + \hat{\sigma}^2\cdot\iota_m\right)\right] \tag{4.64}$$

$$\hat{\Lambda} = \left(\hat{\beta}'\hat{\beta}\right)^{-1}\hat{\beta}'\hat{\mathbf{b}}. \tag{4.65}$$

ACM derive the joint asymptotic distribution of $\hat{\beta}$, $\hat{\lambda}$ and $\hat{\Lambda}$ under assumptions A1 to A3 above, along with the ergodicity of the factors and the independence of the short rate pricing errors $\{u_t^{(1)}\}_{t\in\mathbb{Z}}$ and factor innovations $\{v_t\}_{t\in\mathbb{Z}}$. The consistency of the estimators, which is much easier to show, is proved in the appendix.

The difference between the ACM model on the one hand and the JSZ and HW models on the other is that the ACM model uses the relationship between excess bond returns and factors as the basic building block of the estimation process, while the latter models use the relationship between yields and factors. Therefore, while consistent estimation in the ACM model requires independent return pricing errors, consistent estimation in the JSZ and HW models require independent yield pricing errors. The independence of each type of pricing error is mutually exclusive, so that the choice of which model to use in estimation boils down to the assumption one makes about the properties of the pricing errors.

4.6.4 Simple Self-Consistent (SSC) Estimator

While the ACM estimator studied in the previous section allows us to estimate ATSMs via OLS by taking advantage of excess return equations, a recent finding in Goliński and Spencer (2021) casts doubt as to the self-consistency of the estimator. Recall that the ACM model starts by specifying the short rate dynamics, physical factor dynamics, and market prices of risk:

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma\cdot v_{t+1}^{\mathbb{P}} \\ \lambda_t &= \lambda + \Lambda f_t, \end{aligned}$$

where $v_{t+1}^{\mathbb{P}}$ follows the standard normal distribution under the physical measure. By implication, the risk-neutral dynamics are given in the VAR(1) form

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma\cdot v_{t+1}^{\mathbb{Q}},$$

where $v_{t+1}^{\mathbb{Q}} = v_{t+1}^{\mathbb{P}} + \lambda_t$ follows the standard normal distribution under the risk-neutral measure.

ACM procures the estimates $\hat{K}^{\mathbb{P}}, \hat{G}^{\mathbb{P}}$ and $\hat{\Sigma}$ via OLS estimation of the physical factor dynamics and, using the factor structure of excess bond returns, uses excess bond return equations to obtain estimators $\hat{\lambda}$ and $\hat{\Lambda}$ of the market price of risk parameters, again by OLS. The estimators of the $\mathbb{Q}$-dynamic parameters can then given as

$$\hat{K}^{\mathbb{Q}} = \hat{K}^{\mathbb{P}} + \hat{\Sigma}\hat{\lambda}$$
$$\hat{G}^{\mathbb{Q}} = \hat{G}^{\mathbb{P}} + \hat{\Sigma}\hat{\Lambda},$$

using the well-known relationship between market prices of risk and the factor dynamics. During the estimation process, they assume that a certain portfolio $\mathcal{P}_t = W\mathcal{Y}_t$ of yields is observed without error, using $\mathcal{P}_t$ as the observed yield factors.

Goliński and Spencer (2021) point out that, while the ACM estimators above may be consistent in the statistical sense, they are not self consistent in the following sense. Recall that, in the ACM model, yields are given as affine functions of the factors, as in usual ATSMs; together with measurement errors, we can express the observed yields $\mathcal{Y}_t^o$ at time t as affine functions of the observed factors $\mathcal{P}_t$:

$$\mathcal{Y}_t^o = \mathcal{A} + \mathcal{B} \cdot \mathcal{P}_t + \Sigma_e \cdot e_t,$$

where $\mathcal{A}$ and $\mathcal{B}$ are functions of the $\mathbb{Q}$-parameters $K^{\mathbb{Q}}, G^{\mathbb{Q}}, \Sigma$ and the short-rate parameters δ, β. That $\mathcal{P}_t = W\mathcal{Y}_t$ is observed without error means that $W\Sigma_e \cdot e_t = O_{n\times 1}$, so that

$$\mathcal{P}_t = W\mathcal{Y}_t^o = W\mathcal{A} + W\mathcal{B} \cdot \mathcal{P}_t.$$

By implication, the following identities must hold:

$$W\mathcal{A} = O_{n\times 1}, \quad W\mathcal{B} = I_n.$$

These identities are referred to in Goliński and Spencer (2021) as the self-consistency conditions. For the ACM estimators of the model parameters, the above identities need only hold at the limit due to the consistency of the estimators; this is indicative that the ACM model is not self-consistent. An example of a self-consistent model is the JSZ model, as we will see below.

Taking inspiration from the JSZ identification scheme, Goliński and Spencer (2021) propose a refinement of the ACM estimators of the model parameters. In other words, they introduce a method of estimating the JSZ model via the ACM estimator; the resulting estimators of the model parameters are referred to as the **simple self-consistent (SSC)** estimators. The main advanatage of SSC estimation is that it allows for the JSZ model to be estimated without having to rely on numerical maximization of the log likelihood function.

Specifically, consider the basic JSZ model, in which the short rate dynamics and factor dynamics are given in the canonical form

$$r_t = \iota' f_t$$

$$f_{t+1} = \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ O_{(n-1)\times 1} \end{pmatrix} + J^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}$$

$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}$$

in terms of the n latent factors f_t. We know that yields are then affine in the factors; letting $\mathcal{Y}_t$ collect the m sample yields of maturities $\tau_1 < \cdots < \tau_m$,

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B} f_t,$$

where

$$\mathcal{A} = \begin{pmatrix} \frac{a(\tau_1)}{\tau_1} \\ \vdots \\ \frac{a(\tau_m)}{\tau_m} \end{pmatrix} \quad \text{and} \quad \mathcal{B} = \begin{pmatrix} \frac{b(\tau_1)'}{\tau_1} \\ \vdots \\ \frac{b(\tau_m)'}{\tau_m} \end{pmatrix}.$$

The functions $a(\cdot)$ and $b(\cdot)$ satisfy the usual Ricatti equations

$$a(\tau) = a(\tau-1) + b(\tau-1)' \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ O_{(n-1)\times 1} \end{pmatrix} - \frac{1}{2} b(\tau-1)' \Sigma \Sigma' b(\tau-1)$$

$$b(\tau)' = b(\tau-1)' J^{\mathbb{Q}} + \iota$$

with initial conditions $a(0) = 0$, $b(0) = O_{n\times 1}$. We saw above that solving these equations yields

$$b(\tau) = \left[\sum_{s=0}^{\tau-1} \left(J^{\mathbb{Q}\prime} \right)^s \right] \iota \tag{4.66}$$

$$a(\tau) = \left(\sum_{s=1}^{\tau-1} b_1(s) \right) k_\infty^{\mathbb{Q}} - \frac{1}{2} \sum_{s=1}^{\tau-1} b(s)' \Sigma \Sigma' b(s) \tag{4.67}$$

for any $\tau > 0$, where $b_1(s)$ is the first element of any $b(s)$. By implication, $\mathcal{B}$ depends only on $J^{\mathbb{Q}}$.

As in JSZ, we also assume that the n-dimensional portfolio $\mathcal{P}_t = W \mathcal{Y}_t$ of sample yields is observed without error. This allows us to express $\mathcal{P}_t$ as an affine function of the latent

factors:

$$\mathcal{P}_t = W\mathcal{A} + W\mathcal{B} \cdot f_t,$$

and since this makes $\mathcal{P}_t$ an invariant affine transformation of f_t, the model can be expressed in terms of $\mathcal{P}_t$ as the observed factors:

$$r_t = \delta_{\mathcal{P}} + \beta_{\mathcal{P}}' \mathcal{P}_t$$

$$\mathcal{P}_{t+1} = K_{\mathcal{P}}^i + G_{\mathcal{P}}^i \mathcal{P}_t + \Sigma_{\mathcal{P}} \cdot v_{t+1}^i \quad \text{for any } i = \mathbb{P}, \mathbb{Q},$$

so that the parameters are related to one another as

$$\delta_{\mathcal{P}} = -\iota' \left[W\mathcal{B}\right] \tag{4.68}$$

$$\beta_{\mathcal{P}} = \left[W\mathcal{B}\right]^{-1\prime} \iota \tag{4.69}$$

$$K_{\mathcal{P}}^{\mathbb{Q}} = \left[W\mathcal{B}\right] \left(I_n - J^{\mathbb{Q}}\right) \left[W\mathcal{B}\right]^{-1} \cdot W\mathcal{A} + \left[W\mathcal{B}\right]^{-1} \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ O_{(n-1)\times 1} \end{pmatrix} \tag{4.70}$$

$$G_{\mathcal{P}}^{\mathbb{Q}} = \left[W\mathcal{B}\right] J^{\mathbb{Q}} \left[W\mathcal{B}\right]^{-1} \tag{4.71}$$

$$\Sigma_{\mathcal{P}} = \left[W\mathcal{B}\right] \Sigma. \tag{4.72}$$

It follows that the sample yields are also affine functions of the observed factors $\mathcal{P}_t$:

$$\mathcal{Y}_t = \mathcal{A}_{\mathcal{P}} + \mathcal{B}_{\mathcal{P}} \cdot \mathcal{P}_t,$$

where

$$\mathcal{A}_{\mathcal{P}} = \mathcal{A} - \mathcal{B} \left[W\mathcal{B}\right]^{-1} W\mathcal{A} \tag{4.73}$$

$$\mathcal{B}_{\mathcal{P}} = \mathcal{B} \left[W\mathcal{B}\right]^{-1}. \tag{4.74}$$

Clearly, the JSZ model satisfies the self-consistency conditions

$$W\mathcal{A}_{\mathcal{P}} = O_{n\times 1} \quad \text{and} \quad W\mathcal{B}_{\mathcal{P}} = I_n.$$

Our goal is to estimate the parameters of the model formulated in terms of the observed factors, that is,

$$\theta = \{\delta_{\mathcal{P}},\ \beta_{\mathcal{P}},\ K_{\mathcal{P}}^i,\ G_{\mathcal{P}}^i,\ \Sigma_{\mathcal{P}}\}$$

for $i = \mathbb{P}, \mathbb{Q}$. The SSC estimator exploits equation (4.71), which shows us that $J^{\mathbb{Q}}$ is the ordered Jordan form of $G^{\mathbb{Q}}$, and thus is determined by the eigenvalues of $G^{\mathbb{Q}}$. In light of this discovery, SSC estimation proceeds stepwise as follows:

Step 1: Consistent Estimation of Model Parameters via ACM

First, we procure (statistically) consistent estimators of the model parameters via ACM, which are denoted

$$\delta_{\mathcal{P}}^{ACM},\ \beta_{\mathcal{P}}^{ACM},\ K_{\mathcal{P}}^{i,ACM},\ G_{\mathcal{P}}^{i,ACM},\ \Sigma_{\mathcal{P}}^{ACM}$$

for $i = \mathbb{P}, \mathbb{Q}$. We retain $K_{\mathcal{P}}^{\mathbb{P},ACM}$, $G_{\mathcal{P}}^{\mathbb{P},ACM}$, $\Sigma_{\mathcal{P}}^{ACM}$, which are the OLS estimators of the VAR(1) system

$$\mathcal{P}_{t+1} = K_{\mathcal{P}}^{\mathbb{P}} + G_{\mathcal{P}}^{\mathbb{P}}\mathcal{P}_t + \Sigma_{\mathcal{P}} \cdot v_{t+1}^{\mathbb{P}},$$

as our estimators of the $\mathbb{P}$-parameters of the model.

Using these estimators, we are further able to consistently estimate the factor loadings and intercept $\mathcal{B}_{\mathcal{P}}$ and $\mathcal{A}_{\mathcal{P}}$ of the yields on the observed factors. We denote these estimators by $\mathcal{B}_{\mathcal{P}}^{ACM}$ and $\mathcal{A}_{\mathcal{P}}^{ACM}$.

Step 2: Recovering Parameters of the Latent Factor Model

We now use the estimators above to recover consistent estimators of the parameters of the latent factor model, that is,

$$k_{\infty}^{\mathbb{Q}},\ J^{\mathbb{Q}},\ \Sigma.$$

First, we let the estimator $\hat{J}^{\mathbb{Q}}$ of $J^{\mathbb{Q}}$ be given as the ordered Jordan form of $G_{\mathcal{P}}^{\mathbb{Q},ACM}$, which is consistent because of the consistency of the ordered eigenvalues of $G_{\mathcal{P}}^{\mathbb{Q},ACM}$ for the ordered eigenvalues of $G_{\mathcal{P}}^{\mathbb{Q}}$.

Given $\hat{J}^{\mathbb{Q}}$, we can consistently estimate the loadings $\mathcal{B}$ of the yields on the latent factors, since it is a continuous function of the eigenvalues contained in $J^{\mathbb{Q}}$. Denoting this consistent estimator by $\hat{\mathcal{B}}$, we are able to consistently estimate Σ as

$$\hat{\Sigma} = \left[W\hat{\mathcal{B}}\right]^{-1}\Sigma_{\mathcal{P}}^{ACM}.$$

Finally, equation (4.67) shows us that

$$\mathcal{A} = \underbrace{\begin{pmatrix} \frac{1}{\tau_1}\sum_{s=1}^{\tau_1-1} b_1(s) \\ \vdots \\ \frac{1}{\tau_m}\sum_{s=1}^{\tau_m-1} b_1(s) \end{pmatrix}}_{c_0} k_\infty^{\mathbb{Q}} - \frac{1}{2}\underbrace{\begin{pmatrix} \frac{1}{\tau_1}\sum_{s=1}^{\tau_1-1} b(s)'\Sigma\Sigma' b(s) \\ \vdots \\ \frac{1}{\tau_m}\sum_{s=1}^{\tau_m-1} b(s)'\Sigma\Sigma' b(s) \end{pmatrix}}_{c_1}. \tag{4.75}$$

Substituting equation (4.75) into equation (4.73) shows us that

$$\begin{aligned} \mathcal{A}_{\mathcal{P}} &= \left(I_m - \mathcal{B}\left[W\mathcal{B}\right]^{-1} W\right)\left(c_0 \cdot k_\infty^{\mathbb{Q}} - c_1\right) \\ &= Hc_0 \cdot k_\infty^{\mathbb{Q}} - Hc_1, \end{aligned}$$

where $H = I_m - \mathcal{B}\left[W\mathcal{B}\right]^{-1} W$. If Hc_0 is nonzero, then

$$k_\infty^{\mathbb{Q}} = \left(c_0' H' H c_0\right)^{-1}\left(\mathcal{A}_{\mathcal{P}} + Hc_1\right).$$

Since c_0 and c_1 are continuous functions of the parameters contained in $\mathcal{B}$ and Σ, we can obtain consistent estimators $\hat{c}_0$ and $\hat{c}_1$ of c_0 and c_1 by using $\hat{\mathcal{B}}$ and $\hat{\Sigma}$. The equation above suggests that we can estimate $k_\infty^{\mathbb{Q}}$ as

$$\hat{k}_\infty^{\mathbb{Q}} = \left(\hat{c}_0'\hat{H}'\hat{H}\hat{c}_0\right)^{-1}\left(\mathcal{A}_{\mathcal{P}}^{ACM} + \hat{H}\hat{c}_1\right),$$

where $\hat{H} = I_m - \hat{\mathcal{B}}\left[W\hat{\mathcal{B}}\right]^{-1} W$. Using the consistent estimators $\hat{k}_\infty^{\mathbb{Q}}$, $\hat{J}^{\mathbb{Q}}$ and $\hat{\Sigma}$ of the $\mathbb{Q}$-parameters under the latent factors, we can now procure a consistent estimator $\hat{\mathcal{A}}$ of $\mathcal{A}$.

Step 3: Refining ACM Estimators

Now that we have obtained consistent estimators of $\mathcal{A}, \mathcal{B}, k_\infty^{\mathbb{Q}}, J^{\mathbb{Q}}$ and Σ, the model parameters $\delta_{\mathcal{P}}, \beta_{\mathcal{P}}, K_{\mathcal{P}}^{\mathbb{Q}}$ and $G_{\mathcal{P}}^{\mathbb{Q}}$ can be estimated as continuous functions of these estimators, where the functional forms are given in equations (4.68) to (4.71).

In summary, the SSC estimators of the model parameters are given as follows, where e_1 is the first standard basis vector in $\mathbb{R}^n$:

$$\hat{K}_{\mathcal{P}}^{\mathbb{P}} = K_{\mathcal{P}}^{\mathbb{P},ACM} \tag{4.76}$$

$$\hat{G}_{\mathcal{P}}^{\mathbb{P}} = G_{\mathcal{P}}^{\mathbb{P},ACM} \tag{4.77}$$

$$\hat{\Sigma}_{\mathcal{P}} = \Sigma_{\mathcal{P}}^{ACM} \tag{4.78}$$

$$\hat{J}^{\mathbb{Q}} = \text{Ordered Jordan Form of } G_{\mathcal{P}}^{\mathbb{Q},ACM}$$

$$\hat{b}(\tau) = \left[\sum_{s=0}^{\tau-1}\left(\hat{J}^{\mathbb{Q}}\right)^{s}\right]\iota \quad \text{for any } \tau > 0$$

$$\hat{\mathcal{B}} = \begin{pmatrix}\frac{\hat{b}(\tau_1)}{\tau_1} & \dots & \frac{\hat{b}(\tau_m)}{\tau_m}\end{pmatrix}'$$

$$\hat{\Sigma} = \left[W\hat{\mathcal{B}}\right]^{-1}\Sigma_{\mathcal{P}}^{ACM}$$

$$\hat{H} = I_m - \hat{B}\left[W\hat{B}\right]^{-1}W$$

$$\hat{c}_0 = \begin{pmatrix}\frac{1}{\tau_1}\sum_{s=1}^{\tau_1-1}\hat{b}(s) \\ \vdots \\ \frac{1}{\tau_m}\sum_{s=1}^{\tau_m-1}\hat{b}(s)\end{pmatrix}e_1$$

$$\hat{c}_1 = \frac{1}{2}\begin{pmatrix}\frac{1}{\tau_1}\sum_{s=1}^{\tau_1-1}\hat{b}(s)'\hat{\Sigma}\hat{\Sigma}'\hat{b}(s) \\ \vdots \\ \frac{1}{\tau_m}\sum_{s=1}^{\tau_m-1}\hat{b}(s)'\hat{\Sigma}\hat{\Sigma}'\hat{b}(s)\end{pmatrix}$$

$$\hat{k}_{\infty}^{\mathbb{Q}} = \left(\hat{c}_0'\hat{H}'\hat{H}\hat{c}_0\right)^{-1}\left(\mathcal{A}_{\mathcal{P}}^{ACM} + \hat{H}\hat{c}_1\right)$$

$$\hat{a}(\tau) = \left(\sum_{s=1}^{\tau-1}\hat{b}(s)\right)e_1\cdot\hat{k}_{\infty}^{\mathbb{Q}} - \frac{1}{2}\sum_{s=1}^{\tau-1}\hat{b}(s)'\hat{\Sigma}\hat{\Sigma}'\hat{b}(s) \quad \text{for any } \tau > 0$$

$$\hat{\mathcal{A}} = \begin{pmatrix}\frac{\hat{a}(\tau_1)}{\tau_1} & \dots & \frac{\hat{a}(\tau_m)}{\tau_m}\end{pmatrix}'$$

$$\hat{\delta}_{\mathcal{P}} = -\iota'\left[W\hat{\mathcal{B}}\right] \tag{4.79}$$

$$\hat{\beta}_{\mathcal{P}} = \left[W\hat{\mathcal{B}}\right]^{-1\prime}\iota \tag{4.80}$$

$$\hat{K}_{\mathcal{P}}^{\mathbb{Q}} = \left[W\hat{\mathcal{B}}\right]\left(I_n - \hat{J}^{\mathbb{Q}}\right)\left[W\hat{\mathcal{B}}\right]^{-1}\cdot W\hat{\mathcal{A}} + \left[W\hat{\mathcal{B}}\right]^{-1}\begin{pmatrix}\hat{k}_{\infty}^{\mathbb{Q}} \\ O_{(n-1)\times 1}\end{pmatrix} \tag{4.81}$$

$$\hat{G}_{\mathcal{P}}^{\mathbb{Q}} = \left[W\hat{\mathcal{B}}\right]\hat{J}^{\mathbb{Q}}\left[W\hat{\mathcal{B}}\right]^{-1} \tag{4.82}$$

Chapter 5

Special Topics in the Term Structure Literature

Having introduced and studied the standard ATSM in the previous chapter, here we turn to some of its extensions that have gained popularity in the literature. First, we study macro-finance ATSMs, in which macroeconomic variables or factors are included alongside latent factors when formulating the model. This adds a level of realism to the model, since macroeconomic variables are likely to be closely related to the term structure of interest rates, It also confers on the model desirable features such as better yield forecasts and the ability to derive impulse responses of yields to macro shocks, or vice versa.

Next, we move onto ATSMs with regime-switching parameters. This is a tractable way of accounting for structural breaks in the data, most notably during crises such as the Great Financial Crisis (GFC) and the COVID pandemic, or alternatively between recessions and expansions in general. The regime-switching approach has also proven to be a tractable way of accounting for the zero lower bound (ZLB).

The third topic we will discuss stems from Bauer and Rudebusch (2020), where it is shown that including macroeconomic trends, or falling stars, into ATSMs greatly improves its forecasts, yields more plausible bond risk premia, and helps alleviate small sample bias during estimation. This type of model represents a departure from the usual stationary factor dynamics found in Gaussian ATSMs, and we will discuss some of the complications that arise from the assumption of non-stationary factors.

Finally, we conclude by studying the shadow-rate model for ZLB modeling. This type of model, which imposes the ZLB restriction on the usual ATSM with minimal alterations, has proven to be a powerful means of accounting for the ZLB, and it also allows practitioners to obtain a measure, in the shadow rate, for the stance of monetary policy during ZLB episodes. Imposing the ZLB restriction leads to a non-linear relationship between the factors and yields, however, so we discuss how this relationship is derived and how it affects estimation.

5.1 Macro-Finance Term Structure Models

Here, we first study the baseline macro-finance ATSM introduced in Ang and Piazzesi (2003). Then, we move onto the spanning hypothesis, which is an ongoing debate as to whether macro variables contain information about bond returns unspanned by, or not contained in, the yield curve. Whether macro variables are spanned or unspanned has significant implications for the formulation of macro-finance ATSMs, so we conclude by briefly discussing these implications.

5.1.1 The Baseline Macro-Finance ATSM

Our exposition on macro-finance ATSMs starts from Ang and Piazzesi (2003). The inclusion of macroeconomic factors in that model is motivated by the Taylor rule, which posits that how the central bank determines the short rate r_t is dependent on macroeconomic variables. Specifically, under the basic Taylor rule r_t is determined as

$$r_t = \delta + \beta_m' M_t + u_t, \tag{5.1}$$

where M_t contains measures of the output gap and inflation, and u_t is an interest rate shock orthogonal to M_t. In contrast, in the classical ATSM the short rate is determined as an affine function of latent factors f_t as

$$r_t = \delta + \beta' f_t. \tag{5.2}$$

The core idea of Ang and Piazzesi (2003) is to combine these two expressions, so that the short rate is determined as

$$r_t = \delta + \beta_m' M_t + \beta_f' f_t, \tag{5.3}$$

where f_t are latent factors that are orthogonal to the macro variables contained in M_t. This suggests that the factors in ATSMs should contain not only latent factors but also macroeconomic variables.

As such, Ang and Piazzesi (2003) propose a Gaussian ATSM with the following short

rate and factor dynamics[1]:

$$r_t = \delta + \underbrace{\begin{pmatrix}\beta_f' & \beta_m'\end{pmatrix}}_{\beta} \underbrace{\begin{pmatrix} f_t \\ M_t \end{pmatrix}}_{X_t}, \tag{5.4}$$

$$\begin{pmatrix} f_{t+1} \\ M_{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} K_f^i \\ K_m^i \end{pmatrix}}_{K^i} + \underbrace{\begin{pmatrix} G_{ff}^i & G_{fm}^i \\ G_{mf}^i & G_{mm}^i \end{pmatrix}}_{G^i} \begin{pmatrix} f_t \\ M_t \end{pmatrix} + \underbrace{\begin{pmatrix} \Sigma_{ff} & O_{n\times k} \\ \Sigma_{mf} & \Sigma_{mm} \end{pmatrix}}_{\Sigma} \cdot \underbrace{\begin{pmatrix} v_{f,t+1}^i \\ v_{m,t+1}^i \end{pmatrix}}_{v_{t+1}^i}, \quad \text{for } i = \mathbb{P}, \mathbb{Q} \tag{5.5}$$

where there are n latent factors f_t and k observable macro variables M_t included in the model. $v_{t+1}^{\mathbb{Q}}$ is conditionally standard normal under the risk-neutral measure, and $v_{t+1}^{\mathbb{P}}$ under the physical measure. As usual, the SDF process $\{\mathcal{M}_t\}_{t\in\mathbb{N}}$ is given by $\mathcal{M}_0 = 1$ and

$$\mathcal{M}_{t+1} = \exp\Big(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}^{\mathbb{P}}\Big), \tag{5.6}$$

where $\lambda_t = (\lambda_{ft}', \lambda_{mt}')'$ is an $n+k$-dimensional random vectors representing the market price of risk. Note that, in contrast to the classic ATSM, the innovations associated with the latent factors $v_{f,t+1}^{\mathbb{Q}}$, as well as those associated to the macro variables $v_{m,t+1}^{\mathbb{P}}$, affect the SDF. This indicates that, in this model, macro uncertainty is explicitly considered as a source of systematic risk. Finally, recall that, under the above empirical SDF specification, $v_{t+1}^{\mathbb{Q}}$ and $v_{t+1}^{\mathbb{P}}$ are related by

$$v_{t+1}^{\mathbb{Q}} = \lambda_t + v_{t+1}^{\mathbb{P}}.$$

It will already be clear that this model is just the usual Gaussian ATSM with factors X_t, since its short rate and risk-neutral dynamics can be written as

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ X_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} X_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

Therefore, bond prices are given in the usual exponential affine form

$$P_t(\tau) = \exp\Big(-a(\tau) - b(\tau)' X_t\Big)$$

[1] In the original paper, the short rate, physical dynamics and market price of risk are specified first. The affine specification for the market prices of risk imply the that risk-neutral dynamics are also affine, so specifying the short rate and factor dynamics first is an equivalent way to specify the ATSM.

where the functions $a(\cdot)$ and $b(\cdot)$ satisfy the Ricatti equations

$$a(\tau) = \delta + a(\tau-1) + b(\tau-1)'K^{\mathbb{Q}} - \frac{1}{2}b(\tau-1)'\Sigma\Sigma'b(\tau-1)$$
$$b(\tau) = \beta + G^{\mathbb{Q}\prime}b(\tau-1)$$

and the initial conditions $a(0) = 0$, $b(0) = O_{(n+k)\times 1}$. Yields are then determined as affine functions of X_t, that is, of both macro variables and latent factors:

$$Y_t(\tau) = \underbrace{\frac{a(\tau)}{\tau}}_{\alpha(\tau)} + \underbrace{\frac{b_f(\tau)'}{\tau}}_{\beta_f(\tau)'} \cdot f_t + \underbrace{\frac{b_m(\tau)'}{\tau}}_{\beta_m(\tau)'} \cdot M_t, \tag{5.7}$$

where $b_f(\tau)$ and $b_m(\tau)$ collect the first n and last k entries of $b(\tau)$, respectively. Ang and Piazzesi (2003) identify the latent factors by imposing the restrictions of Dai and Singleton (2000) on the physical factor dynamics. It is shown in Hamilton and Wu (2012) that this does not fully identify the model, similarly to the Dai and Singleton (2000) model.

To estimate the model, we can formulate it in state space form

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B}_f \cdot f_t + \mathcal{B}_m \cdot M_t + \Sigma_e \cdot e_t$$
$$X_t = K^{\mathbb{P}} + G^{\mathbb{P}} X_{t-1} + \Sigma \cdot v_t^{\mathbb{P}},$$

where e_t are a set of standard normal yield pricing errors. Usually, we take the first three (or four) principal components $\hat{f}_t$ of the yields as proxies for the latent factors f_t, and then estimate the above model with $\hat{X}_t = (\hat{f}_t', M_t')'$ in place of $X_t = (f_t', M_t')'$. This is similar to the two-step procedure used to estimate factor-augmented VARs (FAVARs) in Bernanke, Boivin, and Eliasz (2005), and various works in the term structure literature have also shown that it is an econometrically sound approach to estimation.

5.1.2 The Spanning Hypothesis

The model in Ang and Piazzesi (2003) is a simple and intuitive way to incorporate macro variables into an ATSM, but in some ways it yields counterintuitive conclusions. Of them, the hypothesis of spanned macro variables is the most hotly debated. Stacking equation (5.7) for a sample of $n+k$ yields with maturity $\tau_1, \cdots, \tau_{n+k}$ shows us that

$$\underbrace{\begin{pmatrix} Y_t(\tau_1) \\ \vdots \\ Y_t(\tau_{n+k}) \end{pmatrix}}_{\mathcal{Y}_t} = \underbrace{\begin{pmatrix} \alpha(\tau_1) \\ \vdots \\ \alpha(\tau_{n+k}) \end{pmatrix}}_{\mathcal{A}} + \underbrace{\begin{pmatrix} \beta_f(\tau_1)' \\ \vdots \\ \beta_f(\tau_{n+k})' \end{pmatrix}}_{\mathcal{B}_f} \cdot f_t + \underbrace{\begin{pmatrix} \beta_m(\tau_1)' \\ \vdots \\ \beta_m(\tau_{n+k})' \end{pmatrix}}_{\mathcal{B}_m} \cdot M_t.$$

If the $(n+k)\times(n+k)$ matrix $\mathcal{B}=(\mathcal{B}_f,\mathcal{B}_m)$ is nonsingular, then the latent factors and macro variables can be written as

$$\begin{pmatrix} f_t \\ M_t \end{pmatrix} = \mathcal{B}^{-1}\left(\mathcal{Y}_t - \mathcal{A}\right),$$

or as an affine function of the cross-section of yields. Thsi shows us that the macro variables M_t that are relevant for bond pricing and determination of bond risk premia can be recovered as an affine function of the yields. In other words, there is no information contained in macro variables concerning the yield curve that is not also contained in the yields themselves. We say that the Ang and Piazzesi (2003) model implies that the information in macro variables are spanned by the yield curve.

However, there is empirical evidence that rejects the spanning hypothesis; these works show that macro variables actually do help predict bond excess returns even when the yield curve has been controlled for. That is, they show that some macro variables contain information on bond risk premia beyond that contained in the yield curve, which is the direct opposite of what is claimed by the spanning hypothesis. To understand how such conclusions are reached, we first give an overview of the expansive literature on bond return predictability.

Some of the earliest efforts made to identify the factors that helps predict excess bond returns are found in Fama and Bliss (1987) and Campbell and Shiller (1991). In Fama and Bliss (1987), it is found that one-period ahead excess bond returns are predicted by the forward rate spread, that is, the difference in the forward rate and the short rate. Specifically, they run a regression of the form

$$exr_{t+1}^{(\tau)} = a + b\left(f_t^{(\tau)} - r_t\right) + u_{t+1}, \tag{5.8}$$

where $f_t^{(\tau)}$ is the τ-period forward rate, that is, an estimate of the short rate at tiem $t+\tau$, and find that b is significantly larger than 0. Likewise, Campbell and Shiller (1991) finds that the yield spread, or the difference between long and short yields, help predict excess bond returns by running regressions of the form

$$exr_{t+1}^{(\tau)} = a + b\left(Y_t(\tau) - r_t\right) + u_{t+1} \tag{5.9}$$

and conducting significance tests on b [2] [3].

[2] In the original paper, the h-period ahead excess return for a τ-period bond, $exr_{t,t+h}^{(\tau)}$, is regressed on the corresponding yield spread $Y_t(\tau) - h\cdot Y_t(h)$. For notational simplicity, we introduce only the version with $h=1$.

[3] These excess bond return regressions can actually be used to compare different ATSMs. In Dai and Singleton (2002), the excess bond return predictability criteria that ATSMs must satisfy are referred to

Based on these findings, Cochrane and Piazzesi (2005) regress bond excess returns on all available forward spreads, using the forward rate data provided in Fama and Bliss (1987). They find that the coefficients on these forward spreads maintain a hump-shaped pattern for excess returns of bonds of different maturities, which leads them to hypothesize that a single linear combination of these forward spreads might be able to predict excess returns of bonds of all maturities. This is indeed what they find; this linear combination is referred to as the Cochrane-Piazzesi (CP) factor, and it is shown that it is distinct from the first three yield curve PCs – the level, slope and curvature factors. Since the CP factor does not correspond to any of the traditional yield factors in the literature, in Cochrane and Piazzesi (2008) a four-factor Gaussian ATSM is estimated, where the CP factor is included as the fourth factor alongside the usual level, slope and curvature factors.

It is using precisely this CP factor that Ludvigson and Ng (2009) tests whether macro variables contain information on excess bond returns beyond that contained in the yield curve. The finding in Cochrane and Piazzesi (2005) suggests the CP factor is the distillation of all the information on excess bond returns contained in the yield curve. As such, Ludvigson and Ng (2009) run the regression

$$exr_{t+1}^{(\tau)} = a + b' \cdot F_t + \gamma' CP_t + u_{t+1}, \tag{5.10}$$

where F_t contains the principal components extracted from a large panel of macro variables. They find that the macro factors F_t have predictive power for excess bond returns even when the CP factor has been controlled for via CP_t, and as such that macro variables contain information about bond returns not contained in the yield curve. In this sense, their finding supports the hypothesis that macro variables are unspanned, or contain information unspanned by the yield curve.

Many other works support the finding in Ludvigson and Ng (2009) that macro variables have predictive power for excess returns even when the information in the yield curve has been controlled for. However, Bauer and Hamilton (2018) warn against hastily concluding that macro factors contain unspanned information. Specifically, they point out that traditional tests for the null $H_0 : b = 0$ based on equation (5.10) can be misleading in the following ways:

- Since the regressor controlling for the yield curve, CP_t, is necessarily correlated with the time t forecast error u_t, strict exogeneity does not hold. This may lead to significant bias in parameter estimates under small samples (similarly to how the estimates from an autoregressive model are consistent but biased).

collectively as LPY, and they evaluate different ATSMs on the basis of LPY. The Dai-Singleton canonical model with square root processes fail to pass the LPY test; this is one reason why Gaussian ATSMs, which handily pass the LPY test, are preferred.

- The regressors CP_t and F_t may be highly persistent. This can give rise to spurious regressions, since the asymptotics of unit root processes are very different from those of stationary processes. In particular, under a local-to-unity specification, it is shown that the standard Wald statistic used to test for significance is not asymptotic chi-squared.

- The lack of strict exogeneity, as well as persistent regressors, leads to the underestimation of standard errors and thus spurious rejections of the null hypothesis.

For this reason, a new bootstrapping method for testing $H_0 : b = 0$ is developed in Bauer and Hamilton (2018). It is found that traditional tests reject the spanning hypothesis way too often. In other words, the bootstrapping method reveals the evidence against the spanning hypothesis is much weaker than previously thought.

5.1.3 A Model of Unspanned Macro Risks

The evidence contained in Ludvigson and Ng (2009), as well as other works, suggests that the macro-finance ATSM in Ang and Piazzesi (2003), which implies that macro variables are spanned, must be modified to reflect the finding that macro variables contain unspanned information. To this end, Joslin, Priebsch, and Singleton (2014) propose imposing a knife-edge restriction. This restriction, which we will study in more depth below, allows for macro variables contained in the ATSM to affect excess bond returns even when yield factors have been controlled for, while no longer being an affine function of yields.

Formally, recall that the short rate dynamics and risk-neutral factor dynamics are given as

$$r_t = \delta + \beta_f' f_t + \beta_m' M_t$$

$$\begin{pmatrix} f_{t+1} \\ M_{t+1} \end{pmatrix} = \begin{pmatrix} K_f^i \\ K_m^i \end{pmatrix} + \begin{pmatrix} G_{ff}^i & G_{fm}^i \\ G_{mf}^i & G_{mm}^i \end{pmatrix} \begin{pmatrix} f_t \\ M_t \end{pmatrix} + \begin{pmatrix} \Sigma_{ff} & O_{n\times k} \\ \Sigma_{mf} & \Sigma_{mm} \end{pmatrix} \cdot \begin{pmatrix} v_{f,t+1}^i \\ v_{m,t+1}^i \end{pmatrix} \quad \text{for } i = \mathbb{P}, \mathbb{Q}$$

in the baseline macro-finance model of Ang and Piazzesi (2003). Under affine risk-neutral and physical factor dynamics, the market prices of risk are also determined as an affine function of the latent and macro factors f_t and M_t:

$$\underbrace{\begin{pmatrix} \lambda_{f,t} \\ \lambda_{m,t} \end{pmatrix}}_{\lambda_t} = \begin{pmatrix} \lambda_f \\ \lambda_m \end{pmatrix} + \underbrace{\begin{pmatrix} \Lambda_{ff} & \Lambda_{fm} \\ \Lambda_{mf} & \Lambda_{mm} \end{pmatrix}}_{\Lambda} \begin{pmatrix} f_t \\ M_t \end{pmatrix},$$

where

$$\lambda = \Sigma^{-1}\left(K^{\mathbb{P}} - K^{\mathbb{Q}}\right)$$

$$\Lambda = \Sigma^{-1}\left(G^{\mathbb{P}} - G^{\mathbb{Q}}\right).$$

Our analysis of the standard Gaussian ATSM shows us that yields are affine in the latent and macro factors:

$$Y_t(\tau) = \frac{a(\tau)}{\tau} + \frac{b_f(\tau)'}{\tau} \cdot f_t + \frac{b_m(\tau)'}{\tau} M_t \tag{5.11}$$

and that the one-period ahead expected excess bond return is given as

$$\mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] = -b(\tau-1)'\Sigma\lambda - b(\tau-1)'\Sigma\Lambda_f \cdot f_t - b(\tau-1)'\Sigma\Lambda_m \cdot M_t \tag{5.12}$$

Joslin, Priebsch, and Singleton (2014) seek to impose restrictions on the model parameters so as to replicate the following stylized facts:

SF1: **The Number of Risk Factors is Small**
As documented in Joslin, Le, and Singleton (2013) and others, the number of risk factors that affect the yield curve is quite small, usually around three or four. Formally, this means that, in the SDF

$$\mathcal{M}_{t+1} = \exp\Big(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}^{\mathbb{P}}\Big),$$

the $(n+k)$-dimensional random vector λ_t contains a number of zero elements, so that only a few risk factors contained in $v_{t+1}^{\mathbb{P}}$ actually affect agents' assessment of risk.

SF2: **Macro Risks are Unspanned by the Yield Curve**
Equation (5.11) cannot be inverted so that M_t can be expressed as an affine function of the sample yields. This can be ensured by making it so that time t yields are not determined by time t macro factors, that is, by letting $b_m(\tau) = O_{k\times 1}$ for any maturity τ.

SF3: **Macro Factors Help Predict Excess Bond Returns**
The loading of the macro factors on expected excess bond returns in equation (5.12) must not be zero. This is equivalent to replicating the finding in Ludvigson and Ng (2009) and others that macro variables have predictive power for excess bond returns even when the information in the latent yield factors have been

controlled for.

Joslin, Priebsch, and Singleton (2014) replicate the three stylized facts above in the following manner. First, to replicate SF2, it must be the case that yields are affine functions of only the latent factors, that is, it must be the case that

$$Y_t(\tau) = \frac{a(\tau)}{\tau} + \frac{b_f(\tau)'}{\tau} \cdot f_t.$$

In the previous chapter, we showed that, if the short rate is an affine function of the latent factors f_t and f_t follows a VAR(1) process under the risk-neutral measure, then the yields are also affine in f_t. Therefore, a sufficient condition to obtain the above representation is for the short rate and risk-neutral dynamics to be given as

$$\begin{aligned} r_t &= \delta + \beta_f \cdot f_t \\ f_{t+1} &= K_f^{\mathbb{Q}} + G_{ff}^{\mathbb{Q}} \cdot f_t + \Sigma_{ff} \cdot v_{f,t+1}^{\mathbb{Q}}, \end{aligned}$$

or in other words, for the following restrictions to be imposed:

$$\beta_m = O_{k\times 1} \tag{5.13}$$

$$G_{fm}^{\mathbb{Q}} = O_{n\times k}. \tag{5.14}$$

These $(n+1)k$ zero restrictions are called knife-edge restrictions, and ensure that the information contained in the macro factors are left unspanned by the yield curve.

To replicate SF1, Joslin, Priebsch, and Singleton (2014) choose to put the last k entries of λ_t equal to 0, so that only the first n entries remain non-zero. By implication, only the risk factors $v_{f,t+1}^{\mathbb{P}}$, or the innovations to the latent factors, determine agents' assessment of risk. Formally, the SDF is now given as

$$\mathcal{M}_{t+1} = \exp\Big(-r_t - \frac{1}{2}\lambda_{f,t}'\lambda_{f,t} - \lambda_{f,t}' v_{f,t+1}^{\mathbb{P}}\Big).$$

The choice of $\lambda_{m,t} = O_{k\times 1}$ is ensured by letting

$$K_m^{\mathbb{Q}} = K_m^{\mathbb{P}} \tag{5.15}$$

$$\begin{pmatrix} G_{mf}^{\mathbb{Q}} & G_{mm}^{\mathbb{Q}} \end{pmatrix} = \begin{pmatrix} G_{mf}^{\mathbb{P}} & G_{mm}^{\mathbb{P}} \end{pmatrix}. \tag{5.16}$$

In other words, the macro factors follow the same dynamics under both the risk-neutral and physical measures.

Under the above restrictions, the market prices of risk are given as

$$\lambda_t = \begin{pmatrix} \lambda_{f,t} \\ O_{k\times 1} \end{pmatrix},$$

where

$$\lambda_{f,t} = \Sigma_{ff}^{-1}\left(K_f^{\mathbb{P}} - K_f^{\mathbb{Q}}\right) + \Sigma_{ff}^{-1}\left(G_{ff}^{\mathbb{P}} - G_{ff}^{\mathbb{Q}}\right)\cdot f_t + \Sigma_{ff}^{-1} G_{fm}^{\mathbb{P}} \cdot M_t.$$

Therefore, to replicate SF3, it suffices for

$$G_{fm}^{\mathbb{P}} \neq O_{n\times k}. \tag{5.17}$$

Joslin, Priebsch, and Singleton (2014) choose to identify the model via the JSZ restrictions, and assume that the first three principal components $\mathcal{P}_t$ are observed without error as in Joslin, Singleton, and Zhu (2011), so that the model can be formulated with the PCs $\mathcal{P}_t$ in place of the latent factors f_t. Naturally, the short rate and risk-neutral parameters are specified as in equations (4.28) to (4.32).

Bringing all these restrictions together, the short rate dynamics, risk-neutral dynamics, physical dynamics and market price of risk specification in Joslin, Priebsch, and Singleton (2014) are given as follows:

$$r_t = \delta_{\mathcal{P}} + \beta_{\mathcal{P}}' \mathcal{P}_t \tag{5.18}$$

$$\mathcal{P}_{t+1} = K_{\mathcal{P}}^{\mathbb{Q}} + G_{\mathcal{P}\mathcal{P}}^{\mathbb{Q}} \mathcal{P}_t + \Sigma_{\mathcal{P}\mathcal{P}} \cdot v_{\mathcal{P},t+1}^{\mathbb{P}} \tag{5.19}$$

$$\begin{pmatrix} \mathcal{P}_{t+1} \\ M_{t+1} \end{pmatrix} = \underbrace{\begin{pmatrix} K_{\mathcal{P}}^{\mathbb{P}} \\ K_m^{\mathbb{P}} \end{pmatrix}}_{K^{\mathbb{P}}} + \underbrace{\begin{pmatrix} G_{\mathcal{P}\mathcal{P}}^{\mathbb{P}} & G_{\mathcal{P}m}^{\mathbb{P}} \\ G_{m\mathcal{P}}^{\mathbb{P}} & G_{mm}^{\mathbb{P}} \end{pmatrix}}_{G^{\mathbb{P}}} \begin{pmatrix} \mathcal{P}_t \\ M_t \end{pmatrix} + \underbrace{\begin{pmatrix} \Sigma_{\mathcal{P}\mathcal{P}} & O_{n\times k} \\ \Sigma_{m\mathcal{P}} & \Sigma_{mm} \end{pmatrix}}_{\Sigma} \cdot \begin{pmatrix} v_{\mathcal{P},t+1}^{\mathbb{P}} \\ v_{m,t+1}^{\mathbb{P}} \end{pmatrix} \tag{5.20}$$

$$\lambda_t = \begin{pmatrix} \lambda_{\mathcal{P},t} \\ O_{k\times 1} \end{pmatrix} \tag{5.21}$$

$$\lambda_{\mathcal{P},t} = \Sigma_{\mathcal{P}\mathcal{P}}^{-1}\left(K_{\mathcal{P}}^{\mathbb{P}} - K_{\mathcal{P}}^{\mathbb{Q}}\right) + \Sigma_{\mathcal{P}\mathcal{P}}^{-1}\left(G_{\mathcal{P}\mathcal{P}}^{\mathbb{P}} - G_{\mathcal{P}\mathcal{P}}^{\mathbb{Q}}\right)\cdot \mathcal{P}_t + \Sigma_{\mathcal{P}\mathcal{P}}^{-1} G_{\mathcal{P}m}^{\mathbb{P}} \cdot M_t. \tag{5.22}$$

The log-likelihood function can again be decomposed as

$$l(\mathcal{Y} \mid \theta) = \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid X_t; \theta^{\mathbb{Q}}) + \sum_{t+1}^{T} \log f(X_t \mid X_{t-1}; K^{\mathbb{P}}, G^{\mathbb{P}}, \Sigma),$$

where $\theta^{\mathbb{Q}}$ contains the risk-neutral dynamic parameters that determine parameters $\delta_{\mathcal{P}}, \beta_{\mathcal{P}}, K_{\mathcal{P}}^{\mathbb{Q}}, G_{\mathcal{P}\mathcal{P}}^{\mathbb{Q}}$ and $\Sigma_{\mathcal{P}\mathcal{P}}$.

5.2 Term Structure Models with Regime-Switching

It is often the case that the dynamics of time series undergo structural changes. A simple way to capture these structural changes is to divide the sample into sub-samples based on predetermined, or known, structural break dates and separately estimate the model in these sub-samples. However, this is only applicable when structural break dates are known and fixed.

One way to model structural breaks in time series when structural break dates are unknown is through Markov-switching regimes. In these models, we assume that there are N separate regimes in the economy, and that the model parameters undergo changes whenever the economy shifts from one regime to another. Because we do not know a prior which regime the economy is in, the regime at time t, denoted by s_t, is treated as a discrete random variable that takes values in the set $\{1, \cdots, N\}$. The defining property of a Markov-switching model is that the regime s_t at time t is determined exogenously and only on the basis of the preceding regime s_{t-1}. In this sense, the regime process $\{s_t\}_{t\in\mathbb{Z}}$ is a Markov chain with discrete state space $\{1, \cdots, N\}$, hence the name of the model.

We first study some basic results concerning discrete Markov chains, including their stationary distribution, before investigating how they can be incorporated into the ATSM framework. Our exposition is based on Bansal and Zhou (2002), Dai, Singleton, and Yang (2007), and Chib and Kang (2013), among many other regime-switching ATSMs in the literature.

5.2.1 Discrete Markov Chains

Let $\{s_t\}_{t\in\mathbb{Z}}$ be a regime process that is modeled as a time-homogeneous Markov chain taking values in the discrete state space $\{1, \cdots, N\}$. As with all discrete time-homoegeneous Markov chains, there exists a transition probability P for the regime process $\{s_t\}_{t\in\mathbb{Z}}$. Here, P is an $N \times N$ matrix whose (i, j)th element P_{ij} is defined as

$$P_{ij} := \mathbb{P}_t\left(s_{t+1} = j \mid s_t = i\right) = \mathbb{P}\left(s_{t+1} = j \mid s_t = i\right),$$

that is, P_{ij} is the probability of the economy being in the jth regime at time $t+1$ given that the economy is in the ith regime at time t. The second equality follows from the

definition of a Markov process. By definition,

$$\sum_{j=1}^{N} P_{ij} = 1,$$

so that the rows of P all sum to 1.

Transition probabilities can be used to evaluate future probabilities of the economy being in each regime. For instance, suppose that the (unconditional) probabilities of the economy being at each regime at time t is collected in the N-dimensional vector

$$\mu = \begin{pmatrix} \mathbb{P}(s_t = 1) \\ \vdots \\ \mathbb{P}(s_t = N) \end{pmatrix}.$$

Then, the probability that the economy is at regime $1 \leq j \leq N$ at time $t+1$ is

$$\begin{aligned} \mathbb{P}(s_{t+1} = j) &= \sum_{i=1}^{N} \mathbb{P}(s_{t+1} = j \mid s_t = i) \cdot \mathbb{P}(s_t = i) \\ &= \sum_{i=1}^{N} P_{ij} \cdot \mu_i = \mu' P_j, \end{aligned}$$

where P_j is the jth column of P. Therefore, the probabilities of the economy being at each regime at time $t+1$ is collected in the N-dimensional vector

$$\begin{pmatrix} \mathbb{P}(s_{t+1} = 1) \\ \vdots \\ \mathbb{P}(s_{t+1} = N) \end{pmatrix} = P'\mu.$$

Iterating this process shows us that the probability of being at each regime at time $t+h$ is

$$\begin{pmatrix} \mathbb{P}(s_{t+h} = 1) \\ \vdots \\ \mathbb{P}(s_{t+h} = N) \end{pmatrix} = (P')^h \mu.$$

In the case that

$$\mu = P'\mu,$$

then we say that μ is a stationary distribution of P. Heuristically, this means that, once the probabilities of the economy being in each regime at time t are given as μ, then the probabilities of the economy being in each regime at any time $t+h$ is also given as μ,

since

$$\mu = \left(P'\right)^h \mu$$

for any $h > 0$[4].

Stationary Distribution when $N = 2$

An especially useful result concerns the stationary distribution of the regime process when there are 2 regimes, that is, when $N = 2$. In this case, the transition probability can be written as

$$P = \begin{pmatrix} P_{11} & 1 - P_{11} \\ 1 - P_{22} & P_{22} \end{pmatrix}.$$

Assume that $P_{11} < 1$, $P_{22} < 1$ (irreducibility); in this case, this implies that P has exactly one unit root (ergodicity). Let $\mu = (\mu_1, \mu_2)'$ be the stationary distribution of P. Since

$$P'\mu = \mu,$$

this suggests that μ is an eigenvector of P' corresponding to the eigenvalue 1, whose elements sum to 1. This eigenvector μ is found as the solution to the equation

$$O_{2\times 1} = \left(P' - I_2\right)\mu = \begin{pmatrix} P_{11} - 1 & 1 - P_{22} \\ 1 - P_{11} & P_{22} - 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 P_{11} - \mu_1 + \mu_2 - \mu_2 P_{22} \\ \mu_1 - \mu_1 P_{11} + \mu_2 P_{22} - \mu_2. \end{pmatrix},$$

subject to the constraint $\mu_1 + \mu_2 = 1$. Substituting $1 - \mu_1$ for μ_2, μ_1 solves the equation

$$\begin{aligned} 0 &= \mu_1 P_{11} - \mu_1 + (1 - \mu_1) - (1 - \mu_1) P_{22} \\ &= \mu_1 \left(P_{11} + P_{22} - 2\right) + 1 - P_{22}, \end{aligned}$$

so that

$$\mu_1 = \frac{1 - P_{22}}{2 - P_{11} - P_{22}} \quad \text{and} \quad \mu_2 = \frac{1 - P_{11}}{2 - P_{11} - P_{22}}.$$

[4]Sufficient conditions, such as irreducibility and ergodicity, for a stationary distribution to exist will be covered in Professor Kang's class.

5.2.2 ATSMs with Markov-Switching Regimes

Let there be a Markov-switching regime process $\{s_t\}_{t\in\mathbb{Z}}$ with state space $\{1,\cdots,N\}$ and transition probability P. Consider a standard yields-only ATSM framework where yields are determined by n latent factors f_t, and let $\mathcal{F}_t$ be the information available at time t.

We start by specifying the physical dynamics of the factors f_t. Since the economy depends on N regimes, the parameters of the physical dynamics must also depend on these regimes; formally, we assume that the factors follow the regime-switching VAR(1) process

$$f_{t+1} = K^{\mathbb{P}}_{s_{t+1}} + G^{\mathbb{P}}_{s_{t+1}} f_t + \Sigma_{s_{t+1}} \cdot v^{\mathbb{P}}_{t+1},$$

where

$$v^{\mathbb{P}}_{t+1} \mid \mathcal{F}_t, s_{t+1} \sim \mathcal{N}\left[O_{n\times 1}, I_n\right]$$

under the physical measure. The only difference here lies with how the parameters are dependent on the regime at time $t+1$. This bleeds over to the specification of the distribution of the factor innovation $v^{\mathbb{P}}_{t+1}$ as well; now, it is standard normal conditional on both the information up to time t and the regime at time $t+1$.

We retain the no-arbitrage assumption. As a result, the first fundamental theorem of asset pricing furnishes us with a risk-neutral measure $\mathbb{Q}$ such that

$$P_t(\tau) = \mathbb{E}^{\mathbb{Q}}_t\left[\exp(-r_t)\cdot P_{t+1}(\tau-1)\right], \tag{5.23}$$

where $P_t(\tau)$ remains the time t price of a zero-coupon bond with τ periods to maturity.

As in a classical Gaussian ATSM, the errors $v^{\mathbb{P}}_{t+1}$ will serve as our risk factors, in the sense that the SDF governing the perception of time $t+1$ risk at time t is specified as

$$\mathcal{M}_{t+1,s_{t+1}} = \exp\left(-r_t - \frac{1}{2}\lambda'_{t,s_{t+1}}\lambda_{t,s_{t+1}} - \lambda'_{t,s_{t+1}} v^{\mathbb{P}}_{t+1}\right).$$

Here, note that the market price of risk $\lambda_{t,s_{t+1}}$ depends on both time t information and the regime s_{t+1} at time $t+1$. By definition of the risk-netural measure and the SDF, the following equality holds:

$$\mathbb{E}_t\left[\mathcal{M}_{t+1,s_{t+1}}\cdot X_{t+1}\right] = \mathbb{E}^{\mathbb{Q}}_t\left[\exp(-r_t)\cdot X_{t+1}\right] \tag{5.24}$$

for any X_{t+1} in the payoff space. An important feature to note is that the conditioning information does not contain the time $t+1$ regime s_{t+1}, since investors do not know what s_{t+1} is at time t.

A version of Girsanov's theorem holds here as well. Defining

$$v_{t+1}^{\mathbb{Q}} = v_{t+1}^{\mathbb{P}} + \lambda_{t,s_{t+1}},$$

we have

$$v_{t+1}^{\mathbb{Q}} \mid \mathcal{F}_t, s_{t+1} \sim \mathcal{N}\left[O_{n\times 1}, I_n\right]$$

under the risk-neutral measure. Here, the proof of Girsanov's theorem is complicated by the fact that $v_{t+1}^{\mathbb{Q}}$ is conditioned on both $\mathcal{F}_t$ and s_{t+1}, whereas the relationship (5.24) involves conditional expectations given $\mathcal{F}_t$ only. For a formal proof, refer to appendix D of my paper.

Given $v_{t+1}^{\mathbb{Q}}$, we now assume that the factors follow a regime-switching VAR(1)

$$f_{t+1} = K_{s_{t+1}}^{\mathbb{Q}} + G_{s_{t+1}}^{\mathbb{Q}} f_t + \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{Q}}$$

under the risk-neutral measure. By implication, the market prices of risk are given as

$$\lambda_{t,s_{t+1}} = \Sigma_{s_{t+1}}^{-1}\left(K_{s_{t+1}}^{\mathbb{P}} - K_{s_{t+1}}^{\mathbb{Q}}\right) + \Sigma_{s_{t+1}}^{-1}\left(G_{s_{t+1}}^{\mathbb{P}} - G_{s_{t+1}}^{\mathbb{Q}}\right) \cdot f_t. \tag{5.25}$$

To complete the model, we specify the following short rate dynamics:

$$r_t = \delta_{s_t} + \beta_{s_t}' f_t,$$

where the parameters are once again regime-dependent.

To summarize, the short rate dynamics and factor dynamics of the model are given as

$$r_t = \delta_{s_t} + \beta_{s_t}' f_t \tag{5.26}$$
$$f_{t+1} = K_{s_{t+1}}^{\mathbb{Q}} + G_{s_{t+1}}^{\mathbb{Q}} f_t + \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{Q}} \tag{5.27}$$
$$f_{t+1} = K_{s_{t+1}}^{\mathbb{P}} + G_{s_{t+1}}^{\mathbb{P}} f_t + \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{P}} \tag{5.28}$$

where

$$v_{t+1}^{\mathbb{P}} \mid \mathcal{F}_t, s_{t+1} \sim \mathcal{N}\left[O_{n\times 1}, I_n\right]$$

$$v_{t+1}^{\mathbb{Q}} = v_{t+1}^{\mathbb{P}} + \lambda_{t,s_{t+1}}$$

$$P_{ij} = \mathbb{P}\left(s_{t+1} = j \mid s_t = i\right) = \mathbb{Q}\left(s_{t+1} = j \mid s_t = i\right) \quad \text{for any } 1 \le i, j \le N.$$

The last condition requires the regime process to be identically distributed under both the physical and risk-neutral measures. A model in which this is not the case will be discussed

in the next section.

Bond prices are found by solving (5.23). They can be derived approximately or exactly depending on how the model parameters depend on the regime; we investigate each case in turn.

M1: Mean Reversion Parameters are Regime-Independent

This is the specification chosen in Dai, Singleton, and Yang (2007). Suppose that β and $G^{\mathbb{Q}}$ are regime-independent, so that the short rate and risk-neutral dynamics are given as

$$\begin{aligned} r_t &= \delta_{s_t} + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}}_{s_{t+1}} + G^{\mathbb{Q}} f_t + \Sigma_{s_{t+1}} \cdot v^{\mathbb{Q}}_{t+1}. \end{aligned}$$

In this case, bond prices are given in the familiar exponential-affine form

$$P_t(\tau) = \exp\Big(-a_{s_t}(\tau) - b(\tau)' f_t\Big),$$

where $a_{s_t}(0) = 0$ and $b(0) = O_{n\times 1}$. Note that the factor loadings $b(\tau)$ are regime-independent, whereas the intercept term $a_{s_t}(\tau)$ is regime-dependent.

To find $a_{s_t}(\cdot)$ and $b(\cdot)$, we use the no-arbitrage equation to first find that

$$\begin{aligned} \exp\Big(-a_{s_t}(\tau) - b(\tau)' f_t\Big) &= P_t(\tau) \\ &= \mathbb{E}^{\mathbb{Q}}_t \left[\exp(-r_t) \cdot P_{t+1}(\tau - 1)\right] \\ &= \mathbb{E}^{\mathbb{Q}}_t \left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' f_{t+1}\Big)\right] \\ &= \mathbb{E}^{\mathbb{Q}}_t \left[\mathbb{E}^{\mathbb{Q}}_t \left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' f_{t+1}\Big) \mid s_{t+1}\right]\right] \end{aligned}$$

where the final equality follows from the law of iterated expectations. Using the risk-neutral dynamics, we can write

$$\begin{aligned} &\mathbb{E}^{\mathbb{Q}}_t \left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' f_{t+1}\Big) \mid s_{t+1}\right] \\ &\qquad = \exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' K^{\mathbb{Q}}_{s_{t+1}} - b(\tau-1)' G^{\mathbb{Q}} f_t\Big) \\ &\qquad\qquad \times \mathbb{E}^{\mathbb{Q}}_t \left[\exp\Big(-b(\tau-1)' \Sigma_{s_{t+1}} \cdot v^{\mathbb{Q}}_{t+1}\Big) \mid s_{t+1}\right] \\ &= \exp\Big(-\delta_{s_t} - \beta' f_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' K^{\mathbb{Q}}_{s_{t+1}} - b(\tau-1)' G^{\mathbb{Q}} f_t\Big) \\ &\qquad\qquad \times \exp\Big(\frac{1}{2} b(\tau-1)' \Sigma_{s_{t+1}} \Sigma'_{s_{t+1}} b(\tau-1)\Big) \end{aligned}$$

since $v^{\mathbb{Q}}_{t+1} \mid \mathcal{F}_t, s_{t+1}$ is standard normally distributed. Therefore,

$$\begin{aligned}
&\mathbb{E}_t^{\mathbb{Q}}\left[\mathbb{E}_t^{\mathbb{Q}}\left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)'f_{t+1}\Big) \mid s_{t+1}\right]\right] \\
&\qquad = \exp\Big(-\delta_{s_t} - \beta' f_t - b(\tau-1)'G^{\mathbb{Q}} f_t\Big) \\
&\quad \times \mathbb{E}_t^{\mathbb{Q}}\left[\exp\Big(-a_{s_{t+1}}(\tau-1) - b(\tau-1)'K^{\mathbb{Q}}_{s_{t+1}} - \frac{1}{2}b(\tau-1)'\Sigma_{s_{t+1}}\Sigma'_{s_{t+1}}b(\tau-1)\Big)\right] \\
&\qquad = \exp\Big(-\delta_{s_t} - \beta' f_t - b(\tau-1)'G^{\mathbb{Q}} f_t\Big) \\
&\times \left[\sum_{j=1}^{N}\exp\Big(-a_j(\tau-1) - b(\tau-1)'K_j^{\mathbb{Q}} + \frac{1}{2}b(\tau-1)'\Sigma_j\Sigma_j'b(\tau-1)\Big)\cdot \mathbb{P}_t\left(s_{t+1}=j\right)\right].
\end{aligned}$$

By the Markov property of the regime process,

$$\mathbb{P}_t\left(s_{t+1}=j\right) = \mathbb{P}\left(s_{t+1}=j \mid s_t\right) = P_{s_t,j},$$

so we can see that

$$\begin{aligned}
&\mathbb{E}_t^{\mathbb{Q}}\left[\mathbb{E}_t^{\mathbb{Q}}\left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)'f_{t+1}\Big) \mid s_{t+1}\right]\right] \\
&\qquad = \exp\Big(-\delta_{s_t} - \beta' f_t - b(\tau-1)'G^{\mathbb{Q}} f_t\Big) \\
&\quad \times \left[\sum_{j=1}^{N}\exp\Big(-a_j(\tau-1) - b(\tau-1)'K_j^{\mathbb{Q}} + \frac{1}{2}b(\tau-1)'\Sigma_j\Sigma_j'b(\tau-1)\Big)\cdot P_{s_t,j}\right].
\end{aligned}$$

Returning to the original equation, taking logs on both sides reveals that

$$\begin{aligned}
&-a_{s_t}(\tau) - b(\tau)'f_t \\
&\qquad = \log\mathbb{E}_t^{\mathbb{Q}}\left[\mathbb{E}_t^{\mathbb{Q}}\left[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)'f_{t+1}\Big) \mid s_{t+1}\right]\right] \\
&\qquad = -\delta_{s_t} - \beta' f_t - b(\tau-1)'G^{\mathbb{Q}} f_t \\
&\quad + \log\left[\sum_{j=1}^{N}\exp\Big(-a_j(\tau-1) - b(\tau-1)'K_j^{\mathbb{Q}} + \frac{1}{2}b(\tau-1)'\Sigma_j\Sigma_j'b(\tau-1)\Big)\cdot P_{s_t,j}\right].
\end{aligned}$$

Matching intercept terms and coefficient terms, we can now see that

$$b(\tau) = G^{\mathbb{Q}\prime}b(\tau-1) + \beta \tag{5.29}$$

$$a_{s_t}(\tau) = \delta_{s_t} - \log\left[\sum_{j=1}^{N}\exp\Big(-a_j(\tau-1) - b(\tau-1)'K_j^{\mathbb{Q}} + \frac{1}{2}b(\tau-1)'\Sigma_j\Sigma_j'b(\tau-1)\Big)\cdot P_{s_t,j}\right]. \tag{5.30}$$

The regime-independence of $G^{\mathbb{Q}}$ and β mean that the solution for $b(\cdot)$ is given identi-

cally to the classical Gaussian ATSM. At first glance, the expression for $a_{s_t}(\cdot)$ seems intractable, but a first-order Taylor expansion shows us that we can interchange logs and expectations to approximate $a_{s_t}(\tau)$ as

$$a_{s_t}(\tau) = \sum_{j=1}^{N} \Big(\delta_{s_t} + a_j(\tau-1) + b(\tau-1)' K_j^{\mathbb{Q}} - \frac{1}{2} b(\tau-1)' \Sigma_j \Sigma_j' b(\tau-1) \Big) \cdot P_{s_t,j}. \quad (5.31)$$

The jth term in the summation on the right hand side is the solution for $a(\cdot)$ if the economy is known to be in the jth regime at time $t+1$. Therefore, $a_{s_t}(\tau)$ is the weighted average of the intercept term for each regime, where the weights are given as the transition probabilities.

M2: Mean Reversion Parameters are Regime-Dependent

This is the specification chosen in Bansal and Zhou (2002) and Chib and Kang (2013). Suppose that β and $G^{\mathbb{Q}}$ are regime-dependent, so that the short rate and risk-neutral dynamics are given as in equations (5.26) and (5.27). In this case, bond prices are given in the exponential-affine form

$$P_t(\tau) = \exp\Big(-a_{s_t}(\tau) - b_{s_t}(\tau)' f_t \Big),$$

where $a_{s_t}(0) = 0$ and $b_{s_t}(0) = O_{n\times 1}$; note that this time, the factor loadings are also regime-dependent. As in the preceding case, we use the no-arbitrage equation to derive the equation

$$\exp\Big(-a_{s_t}(\tau) - b_{s_t}(\tau)' f_t \Big) = P_t(\tau) = \mathbb{E}_t^{\mathbb{Q}} \Big[\mathbb{E}_t^{\mathbb{Q}} \Big[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' f_{t+1} \Big) \mid s_{t+1} \Big] \Big],$$

where

$$\begin{aligned} &\mathbb{E}_t^{\mathbb{Q}} \Big[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' f_{t+1} \Big) \mid s_{t+1} \Big] \\ &\quad = \exp\Big(-\delta_{s_t} - \beta_{s_t}' f_t - a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' K_{s_{t+1}}^{\mathbb{Q}} - b_{s_{t+1}}(\tau-1)' G_{s_{t+1}}^{\mathbb{Q}} f_t \Big) \\ &\qquad\qquad \times \exp\Big(\frac{1}{2} b_{s_{t+1}}(\tau-1)' \Sigma_{s_{t+1}} \Sigma_{s_{t+1}}' b_{s_{t+1}}(\tau-1)' \Big). \end{aligned}$$

Taking time t expectations on both sides, we are left with

$$\begin{aligned} \exp\Big(-a_{s_t}(\tau) - b(\tau)' f_t \Big) &= \mathbb{E}_t^{\mathbb{Q}} \Big[\mathbb{E}_t^{\mathbb{Q}} \Big[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b(\tau-1)' f_{t+1} \Big) \mid s_{t+1} \Big] \Big] \\ &= \left[\sum_{j=1}^{N} \mathbb{E}_t^{\mathbb{Q}} \Big[\exp\Big(-r_t - a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' f_{t+1} \Big) \mid s_{t+1} = j \Big] \cdot P_{s_t,j} \right]. \end{aligned}$$

The difference between the preceding model and the current model is that the right hand side depends on both f_t and s_{t+1} in a non-linear way. This makes it impossible for us to express the expression on the right hand side as an exponential-affine function of f_t.

Therefore, in this case we must employ a first-order approximation; specifically, we use the approximation

$$\exp(x) \approx 1 + x$$

for small values of x to rewrite the above equation as

$$\begin{aligned}
&-a_{s_t}(\tau) - b_{s_t}(\tau)' f_t + 1 \\
\approx &\sum_{j=1}^{N} \left(-\delta_{s_t} - \beta_{s_t}' f_t - a_j(\tau-1) - b_j(\tau-1)' K_j^{\mathbb{Q}} - b_j(\tau-1)' G_j^{\mathbb{Q}} f_t + \frac{1}{2} b_j(\tau-1)' \Sigma_j \Sigma_j' b_j(\tau-1) + 1 \right) \cdot P_{s_t,j} \\
= &\sum_{j=1}^{N} \left(-\delta_{s_t} - \beta_{s_t}' f_t - a_j(\tau-1) - b_j(\tau-1)' K_j^{\mathbb{Q}} - b_j(\tau-1)' G_j^{\mathbb{Q}} f_t + \frac{1}{2} b_j(\tau-1)' \Sigma_j \Sigma_j' b_j(\tau-1) \right) \cdot P_{s_t,j} + 1.
\end{aligned}$$

Matching intercept and coefficient terms now yields

$$a_{s_t}(\tau) = \sum_{j=1}^{N} \left(\delta_{s_t} + a_j(\tau-1) + b_j(\tau-1)' K_j^{\mathbb{Q}} - \frac{1}{2} b_j(\tau-1)' \Sigma_j \Sigma_j' b(\tau-1) \right) \cdot P_{s_t,j} \tag{5.32}$$

$$b_{s_t}(\tau) = \sum_{j=1}^{N} \left(G_j^{\mathbb{Q}\prime} b_j(\tau-1) + \beta_{s_t} \right) \cdot P_{s_t,j}. \tag{5.33}$$

These solutions have the same interpretation as weighted averages of the solutions of classical Gaussian ATSMs, as in the prceding case.

Despite the first model leading to exact bond prices and the second to approximate bond prices, in practice it is convenient to use the approximation (5.31) for the first model when deriving bond excess returns. Therefore, going forward we assume that the solution to the bond pricing formula are given as in (5.32) and (5.33).

Given the solutions above, yields are given as affine functions of the factors, conditioned on regime:

$$Y_t(\tau) = \underbrace{\frac{a_{s_t}(\tau)}{\tau}}_{\alpha_{s_t}(\tau)} + \underbrace{\frac{b_{s_t}(\tau)'}{\tau}}_{\beta_{s_t}'(\tau)} f_t. \tag{5.34}$$

It remains to derive closed-form solutions for bond excess returns. By definition,

$$\begin{aligned}
exr_{t+1}^{(\tau)} &= \log(P_{t+1}(\tau-1)) - \log(P_t(\tau)) - r_t \\
&= -a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' f_{t+1} + a_{s_t}(\tau) + b_{s_t}(\tau)' f_t - \delta_{s_t} - \beta_{s_t}' f_t \\
&= -a_{s_{t+1}}(\tau-1) - b_{s_{t+1}}(\tau-1)' K_{s_{t+1}}^{\mathbb{Q}} - b_{s_{t+1}}(\tau-1)' G_{s_{t+1}}^{\mathbb{Q}} f_t - b_{s_{t+1}}(\tau-1)' \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{Q}} \\
&\quad + a_{s_t}(\tau) + b_{s_t}(\tau)' f_t - \delta_{s_t} - \beta_{s_t}' f_t,
\end{aligned}$$

and taking expectations on both sides yields

$$\begin{aligned}
RP_t^{(\tau)} &= \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] \\
&= -\sum_{j=1}^{N}\left[a_j(\tau-1) - b_j(\tau-1)' K_j^{\mathbb{Q}} - b_j(\tau-1)' G_j^{\mathbb{Q}} f_t\right] \cdot P_{s_t,j} - \mathbb{E}_t\left[b_{s_{t+1}}(\tau-1)' \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{Q}}\right] \\
&\quad + a_{s_t}(\tau) + b_{s_t}(\tau)' f_t - \delta_{s_t} - \beta_{s_t}' f_t.
\end{aligned}$$

Since

$$\begin{aligned}
\mathbb{E}_t\left[b_{s_{t+1}}(\tau-1)' \Sigma_{s_{t+1}} \cdot v_{t+1}^{\mathbb{Q}} \mid s_{t+1}\right] &= b_{s_{t+1}}(\tau-1)' \Sigma_{s_{t+1}} \cdot \lambda_{t,s_{t+1}} \\
&= b_{s_{t+1}}(\tau-1)'\left(K_{s_{t+1}}^{\mathbb{P}} - K_{s_{t+1}}^{\mathbb{Q}}\right) + b_{s_{t+1}}(\tau-1)'\left(G_{s_{t+1}}^{\mathbb{P}} - G_{s_{t+1}}^{\mathbb{Q}}\right) \cdot f_t,
\end{aligned}$$

using equations (5.32) and (5.33) we can see that

$$\begin{aligned}
RP_t^{(\tau)} &= \mathbb{E}_t\left[exr_{t+1}^{(\tau)}\right] \\
&= \mathbb{E}_t\left[b_{s_{t+1}}(\tau-1)'\left(K_{s_{t+1}}^{\mathbb{P}} - K_{s_{t+1}}^{\mathbb{Q}}\right)\right] + \mathbb{E}_t\left[b_{s_{t+1}}(\tau-1)'\left(G_{s_{t+1}}^{\mathbb{P}} - G_{s_{t+1}}^{\mathbb{Q}}\right)\right] f_t \\
&\quad - \frac{1}{2}\sum_{j=1}^{N} b_j(\tau-1)' \Sigma_j \Sigma_j b_j(\tau-1) \cdot P_{s_t,j} \\
&= \underbrace{\sum_{j=1}^{N}\left[b_j(\tau-1)'\left(K_j^{\mathbb{P}} - K_j^{\mathbb{Q}}\right) - \frac{1}{2}\sum_{j=1}^{N} b_j(\tau-1)' \Sigma_j \Sigma_j b_j(\tau-1)\right] \cdot P_{s_t,j}}_{\text{Constant Part}} \\
&\quad + \underbrace{\left[\sum_{j=1}^{N} b_j(\tau-1)'\left(G_j^{\mathbb{P}} - G_j^{\mathbb{Q}}\right) \cdot P_{s_t,j}\right] f_t}_{\text{Factor Part}}.
\end{aligned}$$

Once again, the one-period ahead bond risk premium is given as an affine function of the time t factors.

Recall from equation (4.9) that we can express the term premium of a long term bond

as the average expected one-period ahead excess return across the life of the bond:

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[exr_{t+h+1}^{(\tau-h)}\right].$$

Using the formula for the one-period ahead risk premium and the law of iterated expectations, we can see that

$$\begin{aligned}TP_t(\tau) = &\frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[b_{s_{t+h}}(\tau-h-1)'\left(K_{s_{t+h}}^{\mathbb{P}}-K_{s_{t+h}}^{\mathbb{Q}}\right)-\frac{1}{2}b_{s_{t+h}}(\tau-h-1)'\Sigma_{s_{t+h}}\Sigma_{s_{t+h}}b_{s_{t+h}}(\tau-h-1)\right]\\ &+\frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[b_{s_{t+h}}(\tau-h-1)'\left(G_{s_{t+h}}^{\mathbb{P}}-G_{s_{t+h}}^{\mathbb{Q}}\right)f_{t+h}\right].\end{aligned}$$

The constant part is easy to compute due to the Markovian property of the regime process. The factor part, however, is more difficult, since it is the time t expectation of the product of a regime-dependent variable and f_{t+h}. Refer to appendix J of my paper for a simple algorithm for the computation of the factor part when $b(\cdot)$ is regime-independent.

5.2.3 Time-Varying Transition Probabilities

In the regime-switching ATSM studied in the preceding section, we assumed that the regime process $\{s_t\}_{t\in\mathbb{Z}}$ is a time-homogeneous Markov chain under both the risk-neutral and physical measures, with the same transition probability$P$. However, in many cases we want the transition probability

$$\mathbb{P}_t\left(s_{t+1}=j \mid s_t=i\right)$$

to be dependent on the factors f_t at time t. For example, suppose we are modeling the zero lower bound via regime-switching, so that the economy shifts between a normal regime and a lower bound regime. Realistically, the probability of the economy being in the lower bound regime in the future depends on how close the short term interest rate is to 0 now, so in this case, $\mathbb{P}_t\left(s_{t+1}=j \mid s_t=i\right)$ likely depends on the factor corresponding to the short end of the yield curve. This is the approach taken in works such as Hördahl and Tristani (2019).

Dai, Singleton, and Yang (2007) furnishes a framework in which the transition probability $\mathbb{P}_t\left(s_{t+1}=j \mid s_t=i\right)$ can depend on the time t factors while making minimal changes to the basic Markov-switching ATSM framework. The only change made to the model in the previous section concerns the SDF and the transition probability under the physical measure. Formally, the short-rate, risk-neutral and physical dynamics are given as in equations (5.26) to (5.28). The regime process remains a time-homogeneous Markov chain

with (time-invariant) transition probability $P^{\mathbb{Q}}$ under the risk-neutral measure, that is,

$$P_{ij}^{\mathbb{Q}} = \mathbb{Q}\left(s_{t+1} = j \mid s_t = i\right)$$

for any $1 \leq i, j \leq n$. Since the short rate and risk-neutral dynamics, along with the distribution of the regime process under the risk-neutral measure, determine bond prices, it follows that bond prices are again given as

$$P_t(\tau) = \exp\Big(-a_{s_t}(\tau) - b_{s_t}(\tau)' f_t\Big),$$

with the recursive solutions to $a_{s_t}(\cdot)$ and $b_{s_t}(\cdot)$ given by equations (5.32) and (5.33).

Meanwhile, we denote the transition probability under the physical measure as

$$P_{ij,t}^{\mathbb{P}} = \mathbb{P}_t\left(s_{t+1} = j \mid s_t = i\right) \tag{5.35}$$

for any $1 \leq i, j \leq N$. The SDF is then modified as follows:

$$\mathcal{M}_{t+1,s_{t+1}} = \exp\Big(-r_t - \Gamma_{t,s_t,s_{t+1}} - \frac{1}{2}\lambda_{t,s_{t+1}}'\lambda_{t,s_{t+1}} - \lambda_{t,s_{t+1}}' v_{t+1}^{\mathbb{P}}\Big). \tag{5.36}$$

Note the inclusion of an additional term $\Gamma_{t,s_t,s_{t+1}}$. We define this term as

$$\Gamma_{t,i,j} = \log\left(\frac{P_{ij,t}^{\mathbb{P}}}{P_{ij}^{\mathbb{Q}}}\right) \tag{5.37}$$

for any $1 \leq i, j \leq N$. Then, we can show that, as before, the distribution of

$$v_{t+1}^{\mathbb{Q}} = \lambda_{t,s_{t+1}} + v_{t+1}^{\mathbb{P}},$$

under the risk-neutral measure is given as

$$v_{t+1}^{\mathbb{Q}} \mid \mathcal{F}_t, s_{t+1} \sim \mathcal{N}\left[O_{n\times 1}, I_n\right].$$

The term $\Gamma_{t,s_t,s_{t+1}}$ is called the market price of regime-shifting (MPRS). This is because

$$\Gamma_{t,i,j} \approx \frac{\mathbb{E}_t\left[I_{\{s_{t+1}=j\}} \mid s_t = i\right] - P_{ij}^{\mathbb{Q}}}{P_{ij}^{\mathbb{Q}}};$$

this is essentially the expected excess return from investing in an asset with a payoff of 1 when the economy is in regime j at time $t+1$ and 0 otherwise, given that the economy is in regime i at time t. In other words, the higher $\Gamma_{t,i,j}$, then the greater agents perceive the risk of going from regime i to regime j; $\Gamma_{t,i,j}$ thus represents agents' perception of the risk associated with changes in regime, hence its name.

In the case that there are two regimes, $N = 2$, Dai, Singleton, and Yang (2007) suggest the following parameterization for $P^{\mathbb{P}}_{ij,t}$:

$$P^{\mathbb{P}}_{ii,t} = \frac{\exp\left(\eta^i_0 + \eta^{i\prime}_1 \cdot f_t\right)}{1 + \exp\left(\eta^i_0 + \eta^{i\prime}_1 \cdot f_t\right)}, \quad P^{\mathbb{P}}_{ij,t} = 1 - P^{\mathbb{P}}_{ii,t}$$

for any $1 \leq i, j \leq 2$. Thus, the model parameters to be estimated now include $\eta^1_0, \eta^1_1, \eta^2_0, \eta^2_1$ in addition to the usual ATSM parameters.

5.2.4 Estimating Markov-Switching ATSMs

Returning to the case of time-invariant transition probabilities, suppose there are yields of m maturities $\tau_1, \cdots, \tau_m$ contained in the sample, collected in the vector $\mathcal{Y}_t$. Defining

$$\mathcal{A}_{s_t} = \begin{pmatrix} \alpha_{s_t}(\tau_1) \\ \vdots \\ \alpha_{s_t}(\tau_m) \end{pmatrix} \quad \text{and} \quad \mathcal{B}_{s_t} = \begin{pmatrix} \beta_{s_t}(\tau_1)' \\ \vdots \\ \beta_{s_t}(\tau_m)' \end{pmatrix},$$

the model can be written in state-space form as

$$\mathcal{Y}_t = \mathcal{A}_{s_t} + \mathcal{B}_{s_t} f_t + \Sigma_e \cdot e_t \tag{5.38}$$

$$f_t = K^{\mathbb{P}}_{s_t} + G^{\mathbb{P}}_{s_t} f_{t-1} + \Sigma_{s_t} \cdot v^{\mathbb{P}}_t, \tag{5.39}$$

where e_t represents a vector of standard normally distributed measurement errors. Collect the parameters in the vector

$$\theta = \{\delta_{s_t}, \beta_{s_t}, K^{\mathbb{Q}}_{s_t}, G^{\mathbb{Q}}_{s_t}, K^{\mathbb{P}}_{s_t}, G^{\mathbb{P}}_{s_t}, \Sigma_{s_t}, P, \Sigma_e\}.$$

Assume that the factors are linear combinations of the yields observed without error, as assumed in JSZ and others; formally, if the linear combination

$$\mathcal{P}_t = \mu + W \mathcal{Y}_t$$

of the yields are observed without error, we can formulate the model in terms of $\mathcal{P}_t$ as the latent factors, as shown in JSZ.

The (conditional) log-likelihood can once again be recovered via the prediction error decomposition:

$$l(\mathcal{Y}_T, \cdots, \mathcal{Y}_1 \mid \theta) = \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta)$$

where $\mathcal{G}_{t-1} = \sigma\{\mathcal{Y}_1, \cdots, \mathcal{Y}_{t-1}\}$. For any $1 \leq t \leq T$, we can see that

$$\begin{aligned} f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta) &= \sum_{i=1}^{N} f(\mathcal{Y}_t \mid s_t = i, \mathcal{G}_{t-1}, \theta) \cdot \mathbb{P}(s_t = i \mid \mathcal{G}_{t-1}, \theta) \\ &= \sum_{i=1}^{N} f(\mathcal{Y}_t \mid s_t = i, \mathcal{P}_t, \theta) \cdot f(\mathcal{P}_t \mid \mathcal{P}_{t-1}, s_t = i, \theta) \cdot \mathbb{P}(s_t = i \mid \mathcal{G}_{t-1}, \theta), \end{aligned}$$

where we used Bayes' rule and the fact that information on $\mathcal{P}_{t-1}$ is contained in $\mathcal{G}_{t-1}$ to justify the second equality.

Here, the first two densities are normal, so that we can write

$$f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta) = \sum_{i=1}^{N} \mathcal{N}\left(\mathcal{Y}_t \mid \mathcal{A}_i + \mathcal{B}_i \mathcal{P}_t,\ \Sigma_e \Sigma_e'\right) \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_i^{\mathbb{P}} + G_i^{\mathbb{P}} \mathcal{P}_{t-1},\ \Sigma_i \Sigma_i'\right) \cdot \mathbb{P}(s_t = i \mid \mathcal{G}_{t-1}, \theta).$$

To finish computing the likelihood, we must find the predictive probability

$$\mathbb{P}(s_t = i \mid \mathcal{G}_{t-1}, \theta).$$

To this end, we rely on the Hamilton filter. First, assume that the regime process is at its stationary distribution μ_0 at time 0, and define the following:

$$\alpha_{t|t-1} = \begin{pmatrix} \mathbb{P}(s_t = 1 \mid \mathcal{G}_{t-1}, \theta) \\ \vdots \\ \mathbb{P}(s_t = N \mid \mathcal{G}_{t-1}, \theta) \end{pmatrix}$$

$$\alpha_{t|t} = \begin{pmatrix} \mathbb{P}(s_t = 1 \mid \mathcal{G}_{t}, \theta) \\ \vdots \\ \mathbb{P}(s_t = N \mid \mathcal{G}_{t}, \theta) \end{pmatrix}$$

for any $1 \leq t \leq T$. Since $\mathcal{G}_0$ is just the trivial σ-algebra, we have

$$\alpha_{0|0} = \begin{pmatrix} \mathbb{P}(s_t = 1 \mid \theta) \\ \vdots \\ \mathbb{P}(s_t = N \mid \theta) \end{pmatrix} = \mu_0.$$

Suppose now that we have found $\alpha_{t-1|t-1}$ for some $1 \leq t \leq T$. Then, for any $1 \leq j \leq N$,

$$\begin{aligned} \alpha_{t|t-1,j} = \mathbb{P}(s_t = j \mid \mathcal{G}_{t-1}, \theta) &= \sum_{i=1}^{N} \mathbb{P}(s_t = j \mid s_{t-1} = i, \mathcal{G}_{t-1}, \theta) \cdot \mathbb{P}(s_{t-1} = i \mid \mathcal{G}_{t-1}, \theta) \\ &= \sum_{i=1}^{N} \alpha_{t-1|t-1,i} \cdot P_{ij} = P_j' \alpha_{t-1|t-1}, \end{aligned}$$

where the second equality follows from the Markovian property of the regime process and P_j is the jth column of P. Therefore,

$$\alpha_{t|t-1} = P' \cdot \alpha_{t-1|t-1}.$$

Now we must find the time t filtered probabilities $\alpha_{t|t}$. By Bayes' rule,

$$\begin{aligned}\mathbb{P}(s_t = i \mid \mathcal{G}_t, \theta) &\propto f(\mathcal{Y}_t \mid s_t = i, \mathcal{G}_{t-1}, \theta) \cdot \mathbb{P}(s_t = i \mid \mathcal{G}_{t-1}, \theta) \\ &= f(\mathcal{Y}_t \mid \mathcal{P}_t, s_t = i, \theta^{\mathbb{Q}}) \cdot f(\mathcal{P}_t \mid \mathcal{P}_{t-1}, s_t = i, \theta^{\mathbb{P}}) \cdot \alpha_{t|t-1,i}\end{aligned}$$

for any $1 \leq i \leq N$. Furthermore, note that

$$\sum_{i=1}^{N} \mathcal{N}\left(\mathcal{Y}_t \mid \mathcal{A}_i + \mathcal{B}_i \mathcal{P}_t,\, \Sigma_e \Sigma_e'\right) \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_i^{\mathbb{P}} + G_i^{\mathbb{P}} \mathcal{P}_{t-1},\, \Sigma_i \Sigma_i'\right) \cdot \alpha_{t|t-1,i} = f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta),$$

as per our derivation earlier. As such,

$$\alpha_{t|t} = \frac{1}{f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta)} \begin{pmatrix} \mathcal{N}(\mathcal{Y}_t \mid \mathcal{A}_1 + \mathcal{B}_1 \mathcal{P}_t,\, \Sigma_e \Sigma_e') \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_1^{\mathbb{P}} + G_1^{\mathbb{P}} \mathcal{P}_{t-1},\, \Sigma_1 \Sigma_1'\right) \\ \vdots \\ \mathcal{N}(\mathcal{Y}_t \mid \mathcal{A}_N + \mathcal{B}_N \mathcal{P}_t,\, \Sigma_e \Sigma_e') \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_N^{\mathbb{P}} + G_N^{\mathbb{P}} \mathcal{P}_{t-1},\, \Sigma_N \Sigma_N'\right) \end{pmatrix} \bigodot \alpha_{t|t-1}.$$

We have thus shown that the log-likelihood can be computed recursively as follows:

Step 0: Initialization

We compute the initial stationary distribution of the regime process, μ_0, using the transition probability P, and put $\alpha_{0|0} = \mu_0$. We also specify the initial factor value $\mathcal{P}_0$.

Step 1: Computing Predictive Probabilities

Given $\alpha_{t-1|t-1}$, we construct the predictive probabilities as

$$\alpha_{t|t-1} = P' \cdot \alpha_{t-1|t-1},$$

and obtain the conditional density as

$$f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta) = \sum_{i=1}^{N} \mathcal{N}\left(\mathcal{Y}_t \mid \mathcal{A}_i + \mathcal{B}_i \mathcal{P}_t,\, \Sigma_e \Sigma_e'\right) \cdot \mathcal{N}\left(f_t \mid K_i^{\mathbb{P}} + G_i^{\mathbb{P}} f_{t-1},\, \Sigma_i \Sigma_i'\right) \cdot \alpha_{t|t-1,i}.$$

Step 2: Computing Filtered Probabilities

Given the predictive probability and the log-likelihood, obtain the filtered prob-

abilities as

$$\alpha_{t|t} = \frac{1}{f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta)} \begin{pmatrix} \mathcal{N}\left(\mathcal{Y}_t \mid \mathcal{A}_1 + \mathcal{B}_1 \mathcal{P}_t,\, \Sigma_e \Sigma_e'\right) \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_1^{\mathbb{P}} + G_1^{\mathbb{P}} \mathcal{P}_{t-1},\, \Sigma_1 \Sigma_1'\right) \\ \vdots \\ \mathcal{N}\left(\mathcal{Y}_t \mid \mathcal{A}_N + \mathcal{B}_N \mathcal{P}_t,\, \Sigma_e \Sigma_e'\right) \cdot \mathcal{N}\left(\mathcal{P}_t \mid K_N^{\mathbb{P}} + G_N^{\mathbb{P}} \mathcal{P}_{t-1},\, \Sigma_N \Sigma_N'\right) \end{pmatrix} \bigodot \alpha_{t|t-1}.$$

Step 3: Computing Log-likelihood

If $t < T$, then return to step 1. Otherwise, we compute the log-likelihood as

$$l(\mathcal{Y}_T, \cdots, \mathcal{Y}_1 \mid \theta) = \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}, \theta).$$

The log-likelihood is a complex function of the model parameters, which makes it difficult to numerically find its global maximum. Therefore, in practice, we rely on the EM algorithm to obtain estimates of the model parameters.

5.3 The Zero Lower Bound and Term Structure Models

In recent years, short term interest rates have been repeatedly bound by the ZLB, first during the GFC and also over the COVID pandemic years. Many works have pointed out that Gaussian ATSMs are ill-equipped to accomodate the lower bound restriction.Most notably, since the factor innovation variances are constant in Gaussian ATSMs, the probability of the short rate becoming negtaive is almost the same as the probability that it remains positive when the short rate is near 0. In other words, Gaussian ATSMs fail to capture the inherent asymmetry in the yield curve when the short rate is near 0, where the short rate is more likely to move upward rather than downward. Moreover, structural changes that may take place during the ZLB, such as the compression of yields documented in Swanson and John C Williams (2014) and others, are not reflected in Gaussian ATSMs.

For this reason, practitioners felt a need for a new type of ATSM that takes the ZLB into account. Early attempts to impose the ZLB restriction include the square-root processes of the CIR model and the Dai and Singleton (2000) model, where the factor innovations are square roots of the current factors, and the quadratic term structure model of D.-H. Ahn, Dittmar, and Gallant (2002). These models prove inadequate for our

purposes, however, because they model the lower bound as a reflecting barrier. That is, in the Dai and Singleton (2000) model and the quadratic model, the short rate bounces off of the lower bound, so that the economy remains at the lower bound for only a limited period of time. This is counterfactual to what actually happens when the economy is at the lower bound; during the GFC and the COVID pandemic, the short rate remained near 0 for years on end.

As such, two main approaches to lower bound modeling have taken root in the literature. The first is a regime-switching approach, adopted in works such as Christensen (2013), Hördahl and Tristani (2019), and the paper I co-authored with Professor Kang. In these models, the economy is assumed to shift between two regimes: the normal regime, where the economy is not bound by the lower bound, and the lower bound regime, where the economy is subject to the lower bound restriction. The advantage of this approach is that it allows us to embed in the model structural changes that may occur during the ZLB, such as the aforementioned compression of yields. Indeed, in my paper we show that incorporating such structural changes is integral to studying the way investors' attitude to risk changes when the short rate is stuck at the lower bound, and therefore to the estimation of term premia.

The main weakness of the regime-switching approach is that it does not explicitly impose the lower bound restriction. As noted in Hördahl and Tristani (2019), in regime-switching models of the ZLB, it is possible for the short rate to become negative; it is only that the probability of doing so is negligible. Therefore, in practice, by far the most preferred method of accounting for the ZLB is the shadow-rate approach, pioneered by Black (1995) and developed in works such as Krippner (2013) (for continuous time models) and Wu and Xia (2016) (for discrete time models). Shadow-rate models are based on an overwhelmingly simple intuition. As in Black (1995), consider an investor that holds a hypothetical short term bond whose rate of return, s_t, can be negative. If s_t actually falls below 0, then the investor would have to pay the bank to hold onto this bond, and as such, she would choose to liquidate the bond and hold cash instead. Since cash has a rate of return of 0, the rate of return the investor sees in this case would be equal to 0. Therefore, the short rate r_t, which is the actual rate of return an investor faces from holding a short term bond, must be given as

$$r_t = \max(s_t, 0). \tag{5.40}$$

Here, the rate of return s_t on the hypothetical short term bond is called the shadow rate, and the above equation shows us that the short rate r_t is equal to the shadow rate when it is positive, and equal to 0 otherwise. The shadow rate s_t thus represents what the short rate would have been if the lower bound restriction were not present, hence its name.

In this section, we mainly focus on the models of Ichiue and Ueno (2013) and Wu and

Xia (2016), where the above shadow-rate representation of the short rate is incorporated into the classical Gaussian ATSM framework in discrete time. Due to the non-linearity introduced by equation (5.40), deriving the yield formula requires us to make use of the forward rate representation of yields, that is,

$$Y_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1} f_t^{(h)},$$

where $f_t^{(h)}$ is the h-period ahead forward rate evaluated at time t, and is defined as

$$f_t^{(h)} = (h+1)Y_t(h+1) - h \cdot Y_t(h).$$

Ichiue and Ueno (2013) and Wu and Xia (2016) each use different approximations to recover a tractable closed form expression for $f_t^{(h)}$, using which $Y_t(\tau)$ can be computed. The formal shadow rate term structure model is introduced below.

5.3.1 Shadow Rate Term Structure Models

As in classical Gaussian ATSMs, shadow rate term structure models (SRTSMs) are specified via their short rate dynamics and factor dynamics. As usual, we assume that the no-arbitrage condition holds, so that the no-arbitrage equation can be expressed in terms of the risk-netural measure. In addition, the SDF $\mathcal{M}_{t+1}$ is given in the usual exponential-affine form with n market prices of risk λ_t and n risk factors $v_{t+1}^{\mathbb{P}}$.

Starting with the factor dynamics, letting there be n latent or macro factors f_t, it is assumed that f_t follow a VAR(1) process under both the risk-neutral and physical dynamics:

$$f_{t+1} = K^i + G^i f_t + \Sigma \cdot v_{t+1}^i \quad \text{for } i = \mathbb{P}, \mathbb{Q},$$

where v_{t+1}^i is standard normally distributed under measure i and

$$v_{t+1}^{\mathbb{Q}} = \lambda_t + v_{t+1}^{\mathbb{P}}.$$

The main difference between Gaussian ATSMs and SRTSMs is in the specification of the short rate dynamics. While Gaussian ATSMs assume that the short rate itself is an affine function of the factors, in SRTSMs we assume that the shadow rate s_t is affine in the factors:

$$s_t = \delta + \beta' f_t.$$

Afterward, the short rate follows equation (5.40) and is given as the maximum of the shadow rate s_t and an effective lower bound $\underline{r}$[5]. Finally, the market prices of risk λ_t are given as affine functions of the factors, following the extended affine specification of Cheridito, Filipovic, and Kimmel (2007).

In summary, the basic SRTSM is specified by the following equations:

$$r_t = \max(s_t, 0) \qquad \text{(Short-Rate Dynamics)}$$

$$s_t = \delta + \beta' f_t \qquad \text{(Shadow Rate Dynamics)}$$

$$f_{t+1} = K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}} \qquad \text{(Risk-Neutral Dynamics)}$$

$$f_{t+1} = K^{\mathbb{P}} + G^{\mathbb{P}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{P}}. \qquad \text{(Physical Dynamics)}$$

To derive the formula for the h-period ahead forward rate $f_t^{(h)}$, we first make use of the no-arbitrage equation. Since, for any maturity τ,

$$P_t(\tau) = \mathbb{E}_t^{\mathbb{Q}} \left[\exp(-r_t) P_{t+1}(\tau - 1) \right]$$

holds and the initial condition is $P_t(0) = 1$, iterating ahead we can find that

$$\begin{aligned}
P_t(1) &= \mathbb{E}_t^{\mathbb{Q}} \left[\exp(-r_t) \right] \\
P_t(2) &= \mathbb{E}_t^{\mathbb{Q}} \left[\exp(-r_t) \cdot \mathbb{E}_{t+1}^{\mathbb{Q}} \left[\exp(-r_{t+1}) \right] \right] = \mathbb{E}_t^{\mathbb{Q}} \left[\exp(-r_t - r_{t+1}) \right] \\
&\vdots \\
P_t(\tau) &= \mathbb{E}_t^{\mathbb{Q}} \left[\exp\left(- \sum_{h=0}^{\tau-1} r_{t+h} \right) \right].
\end{aligned}$$

The forward rate $f_t^{(h)}$ is now given as

$$\begin{aligned}
f_t^{(h)} &= \log(P_t(h)) - \log(P_t(h+1)) \\
&= \log \left[\mathbb{E}_t^{\mathbb{Q}} \left[\exp\left(- \sum_{j=0}^{h-1} r_{t+j} \right) \right] \right] - \log \left[\mathbb{E}_t^{\mathbb{Q}} \left[\exp\left(- \sum_{j=0}^{h} r_{t+j} \right) \right] \right].
\end{aligned}$$

Due to the non-linearity of the log and exponential functions, the SRTSM does not admit a closed-form solution of the forward rate. Obtaining a closed-form solution thus requires approximations, and it is in the choice of approximation that the models of Ichiue and

[5]While we may put $\underline{r}$ equal to 0, as in Black (1995), the presence of negative interest rates in Japan and the Euro area, as well as evidence that yields are bound at a level slightly higher than 0 in the U.S., suggests that it would improve model fit to let $\underline{r}$ be a free parameter. This is the approach chosen in both Ichiue and Ueno (2013) and Wu and Xia (2016), while Christensen and Rudebusch (2016) compares a model where $\underline{r}$ is left as a free parameter against a model where it is fixed at 0.

Ueno (2013) and Wu and Xia (2016) diverge. For later use, we define the terms

$$\bar{a}(h) = \delta + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \right] K^{\mathbb{Q}} \tag{5.41}$$

$$a(h) = \bar{a}(h) - \frac{1}{2} \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \right] \Sigma\Sigma' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}\prime} \right)^j \right] \beta \tag{5.42}$$

$$b(h) = \left(G^{\mathbb{Q}\prime} \right)^h \beta \tag{5.43}$$

$$\sigma^{\mathbb{Q}}(h) = \sqrt{\sum_{j=0}^{h-1} \beta' \left(G^{\mathbb{Q}} \right)^j \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^j \beta}. \tag{5.44}$$

The Ichiue-Ueno Model

The SRTSM introduced in Ichiue and Ueno (2013) was designed to study the movement of Japanese yields, which had been bound by the lower bound ever since the early 2000s. To derive a closed-form solution for the forward rate $f_t^{(h)}$, Ichiue and Ueno (2013) assume that the Jensen's inequality term that follows from interchanging logs and expectations is small [6]. This allows the forward rate to be approximated as follows:

$$f_t^{(h)} \approx \mathbb{E}_t^{\mathbb{Q}} \left[\sum_{j=0}^{h} r_{t+j} \right] - \mathbb{E}_t^{\mathbb{Q}} \left[\sum_{j=0}^{h-1} r_{t+j} \right] = \mathbb{E}_t^{\mathbb{Q}} \left[r_{t+h} \right]. \tag{5.45}$$

Recall that, under the expectations hypothesis, the forward rate $f_t^{(h)}$ equals the expected h-period ahead short rate $\mathbb{E}_t[r_{t+h}]$. The above approximation simply states that, when agents' risk aversion is taken into account, $f_t^{(h)}$ must equal the $\mathbb{Q}$-expected value of r_{t+h} instead of its $\mathbb{P}$-expected value.

It remains to compute the expected value $\mathbb{E}_t^{\mathbb{Q}}[r_{t+h}]$. Since the h-period ahead short rate r_{t+h} is a function of the h-period ahead shadow rate s_{t+h}, we first calculate the moments of the shadow rate. For any $h > 0$, the shadow rate dynamics and risk-neutral dynamics imply that

$$\begin{aligned} s_{t+h} &= \delta + \beta' f_{t+h} \\ &= \delta + \beta' K^{\mathbb{Q}} + \beta' \Sigma \cdot v_{t+h}^{\mathbb{Q}} + \beta' G^{\mathbb{Q}} f_{t+h-1} \end{aligned}$$

[6] They acknowledge that, since the Jensen's inequality term increases exponentially as the yield maturity increases, their approximation may be inappropriate when studying yields of longer maturities, such as 30-year bonds. They nevertheless claim that the approximation error from ignoring Jensen's inequality is small for yields of maturities 10 years or less.

$$= \cdots = \delta + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^j\right] K^{\mathbb{Q}} + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^j \Sigma \cdot v_{t+h-j}^{\mathbb{Q}}\right] + \beta' \left(G^{\mathbb{Q}}\right)^h f_t.$$

Since $\{v_t^{\mathbb{Q}}\}_{t\in\mathbb{Z}}$ is pairwise uncorrelated under the risk-neutral measure, it follows that

$$\mathbb{E}_t^{\mathbb{Q}}[s_{t+h}] = \underbrace{\delta + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^j\right] K^{\mathbb{Q}}}_{\bar{a}(h)} + \underbrace{\beta' \left(G^{\mathbb{Q}}\right)^h}_{b(h)'} f_t$$

$$\mathrm{Var}_t^{\mathbb{Q}}(s_{t+h}) = \underbrace{\sum_{j=0}^{h-1} \beta' \left(G^{\mathbb{Q}}\right)^j \Sigma\Sigma' \left(G^{\mathbb{Q}\prime}\right)^j \beta}_{\left(\sigma^{\mathbb{Q}}(h)\right)^2}.$$

We must now express $\mathbb{E}_t^{\mathbb{Q}}[r_{t+h}]$ in terms of the $\mathbb{Q}$-moments of the shadow rate. Since $r_{t+h} = \max(s_{t+h}, \underline{r})$, we have

$$\mathbb{E}_t^{\mathbb{Q}}[r_{t+h}] = \mathbb{E}_t^{\mathbb{Q}}\left[s_{t+h} \cdot I_{\{s_{t+h}\geq \underline{r}\}}\right] + \underline{r} \cdot \mathbb{Q}_t(s_{t+h} < \underline{r}).$$

Since $v_{t+1}^{\mathbb{Q}}, \cdots, v_{t+h}^{\mathbb{Q}}$ are jointly Gaussian and mutually independent under the risk-neutral measure, we can see that

$$s_{t+h} \mid \mathcal{F}_t \sim \mathcal{N}\left[\bar{a}(h) + b(h)' f_t,\ \left(\sigma^{\mathbb{Q}}(h)\right)^2\right],$$

under the risk-neutral measure, and therefore

$$\mathbb{Q}_t(s_{t+h} < \underline{r}) = \Phi\left(\frac{\underline{r} - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}\right)$$

$$\mathbb{E}_t^{\mathbb{Q}}[r_{t+h}] = \frac{1}{\sigma^{\mathbb{Q}}(h)} \int_{\underline{r}}^{\infty} z \cdot \phi\left(\frac{z - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}\right) dz,$$

where $\phi : \mathbb{R} \to (0, +\infty)$ is the standard normal density and $\Phi : \mathbb{R} \to [0,1]$ is the cdf of the standard normal distribution. A simple change of variables tells us that

$$\begin{aligned}
\mathbb{E}_t^{\mathbb{Q}}\left[s_{t+h} \cdot I_{\{s_{t+h}\geq \underline{r}\}}\right] &= \frac{1}{\sigma^{\mathbb{Q}}(h)} \int_{\underline{r}}^{\infty} z \cdot \phi\left(\frac{z - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}\right) dz \\
&= \int_{\frac{\underline{r} - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}}^{\infty} \left(\sigma^{\mathbb{Q}}(h) x + \bar{a}(h) + b(h)' f_t\right) \cdot \phi(x)\, dx \\
&= \sigma^{\mathbb{Q}}(h) \cdot \int_{\frac{\underline{r} - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}}^{\infty} x \cdot \phi(x) dx + \left[\bar{a}(h) + b(h)' f_t\right] \left[1 - \Phi\left(\frac{\underline{r} - \bar{a}(h) - b(h)' f_t}{\sigma^{\mathbb{Q}}(h)}\right)\right].
\end{aligned}$$

Through the fundamental theorem of calculus, we have

$$\int_{\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}}^{\infty} x\cdot\phi(x)dx = -\left[\phi(x)\right]_{\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}}^{\infty} = \phi\left(\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}\right),$$

so it follows that

$$\begin{aligned}
\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] &= \mathbb{E}_t^{\mathbb{Q}}\left[s_{t+h}\cdot I_{\{s_{t+h}\geq\underline{r}\}}\right] + \underline{r}\cdot\mathbb{Q}_t\left(s_{t+h}<\underline{r}\right)\\
&= \sigma^{\mathbb{Q}}(h)\cdot\phi\left(\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}\right) + \left[\bar{a}(h)+b(h)'f_t\right]\left[1-\Phi\left(\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}\right)\right]\\
&\quad + \underline{r}\cdot\Phi\left(\frac{\underline{r}-\bar{a}(h)-b(h)'f_t}{\sigma^{\mathbb{Q}}(h)}\right)\\
&= \sigma^{\mathbb{Q}}(h)\cdot\phi\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + \left[\bar{a}(h)+b(h)'f_t-\underline{r}\right]\Phi\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + \underline{r}\\
&= \underline{r} + \sigma^{\mathbb{Q}}(h)\left[\phi\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + \frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\cdot\Phi\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right)\right],
\end{aligned}$$

where the third equality follows from the symmetry of ϕ and the fact that $1-\Phi(x) = \Phi(-x)$ for any $x\in\mathbb{R}$. Defining the function $g:\mathbb{R}\to\mathbb{R}$ as

$$g(x) = x\cdot\Phi(x)+\phi(x), \tag{5.46}$$

the conditional expectation $\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right]$ and thus the forward rate $f_t^{(h)}$ can be written as

$$f_t^{(h)} \approx \mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] = \underline{r} + \sigma^{\mathbb{Q}}(h)\cdot g\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right). \tag{5.47}$$

It follows that the τ-period yield is given as

$$Y_t(\tau) \approx \underline{r} + \frac{1}{\tau}\left(\sum_{h=1}^{\tau-1}\sigma^{\mathbb{Q}}(h)\cdot g\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + r_t - \underline{r}\right). \tag{5.48}$$

The Wu-Xia Model

Wu and Xia (2016) use a more precise second-order approximation instead of ignoring the Jensen's inequality term outright. Specifically, they use the approximation

$$\log\left(\mathbb{E}\left[\exp(Z)\right]\right) \approx \mathbb{E}\left[Z\right] + \frac{1}{2}\mathrm{Var}\left(Z\right)$$

for any square integrable random variable Z[7]. Now, the forward rate is approximated as follows:

$$\begin{aligned} f_t^{(h)} &= \log\left[\mathbb{E}_t^{\mathbb{Q}}\left[\exp\left(-\sum_{j=0}^{h-1} r_{t+j}\right)\right]\right] - \log\left[\mathbb{E}_t^{\mathbb{Q}}\left[\exp\left(-\sum_{j=0}^{h} r_{t+j}\right)\right]\right] \\ &\approx \mathbb{E}_t^{\mathbb{Q}}\left[\sum_{j=0}^{h} r_{t+j}\right] - \mathbb{E}_t^{\mathbb{Q}}\left[\sum_{j=0}^{h-1} r_{t+j}\right] + \frac{1}{2}\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=0}^{h-1} r_{t+j}\right) - \frac{1}{2}\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=0}^{h} r_{t+j}\right) \\ &= \mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] + \frac{1}{2}\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} r_{t+j}\right) - \frac{1}{2}\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} r_{t+j}\right), \end{aligned}$$

This approximation is more precise than that employed in Ichiue and Ueno (2013), since it is based on a second-order approximation instead of a first-order one. Using further approximations to the variances and covariances of truncated normal random vectors, Wu and Xia show in their paper (and us in the appendix) that the h-period ahead forward rate can be approximated as

$$f_t^{(h)} \approx \underline{r} + \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{a(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right). \tag{5.49}$$

Therefore, the τ-period yield at time t is approximated as

$$Y_t(\tau) \approx \underline{r} + \frac{1}{\tau}\left(\sum_{h=1}^{\tau-1} \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{a(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + r_t - \underline{r}\right). \tag{5.50}$$

Note that the only difference between the formulas in equations (5.48) and (5.50) lies in the use of $a(h)$ instead of $\bar{a}(h)$ in the latter. Furthermore, the only difference between the

[7]To derive this approximation, we employ two Taylor approximations. First, a second degree Taylor approxmiation of the exponential function around $\mathbb{E}[Z]$ yields the formula

$$\exp(Z) \approx \exp(\mathbb{E}[Z]) + \exp(\mathbb{E}[Z]) \cdot (Z - \mathbb{E}[Z]) + \frac{1}{2}\exp(\mathbb{E}[Z]) \cdot (Z - \mathbb{E}[Z])^2.$$

Taking expectations on both sides now yields

$$\mathbb{E}[\exp(Z)] \approx \exp(\mathbb{E}[Z])\left[1 + \frac{1}{2}\mathrm{Var}(Z)\right].$$

Finally, taking logs on both sides and using the first degree Taylor approxmiation $\log(1+x) \approx x$ for small values of x shows us that

$$\begin{aligned} \log(\mathbb{E}[\exp(Z)]) &\approx \mathbb{E}[Z] + \log\left(1 + \frac{1}{2}\mathrm{Var}(Z)\right) \\ &\approx \mathbb{E}[Z] + \frac{1}{2}\mathrm{Var}(Z). \end{aligned}$$

The accuracy of the two approximations rely on small second moments of the random variable Z.

two terms, as seen in equation (5.42), is in the Jensen's inequality term

$$\frac{1}{2}\beta' \left[\sum_{j=0}^{h-1}\left(G^{\mathbb{Q}}\right)^j\right]\Sigma\Sigma'\left[\sum_{j=0}^{h-1}\left(G^{\mathbb{Q}\prime}\right)^j\right]\beta.$$

Therefore, we can interpret the formula in equation (5.50) as a version of equation (5.48) that does not completely disregard Jensen's inequality. Nevertheless, because the Wu-Xia SRTSM also involves approximations to the variance and covariance terms of the current and future short rates, it is unclear which model yields more precise approximations of the forward rate.

The Term Premium in SRTSMs

The term premium of a τ-maturity bond is given as the difference between a τ-maturity yield and its expectation hypothesis component:

$$TP_t(\tau) = Y_t(\tau) - \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[r_{t+h}\right].$$

In classical ATSMs, we use the formula in Cochrane and Piazzesi (2008) to formulate the term premium as an affine function of the factors directly from formula for one-period ahead bond excess returns. In this case, because there is no simple analytical formula for bond prices, we instead compute the expectations component first.

Mirroring the process through which we derived $\mathbb{E}_t^{\mathbb{Q}}[r_{t+h}]$ for $h > 1$, we can see that

$$\mathbb{E}_t\left[r_{t+h}\right] = \underline{r} + \sigma^{\mathbb{P}}(h)\cdot g\left(\frac{\bar{\alpha}(h)+\beta(h)'f_t-\underline{r}}{\sigma^{\mathbb{P}}(h)}\right), \tag{5.51}$$

where

$$\bar{\alpha}(h) = \delta + \beta'\left[\sum_{j=0}^{h-1}\left(G^{\mathbb{P}}\right)^j\right]K^{\mathbb{P}}$$

$$\beta(h) = \left(G^{\mathbb{P}\prime}\right)^h\beta$$

$$\sigma^{\mathbb{P}}(h) = \sqrt{\sum_{j=0}^{h-1}\beta'\left(G^{\mathbb{P}}\right)^j\Sigma\Sigma'\left(G^{\mathbb{P}\prime}\right)^j\beta}.$$

Therefore,

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=1}^{\tau-1}\left[\sigma^{\mathbb{Q}}(h)\cdot g\left(\frac{\bar{a}(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) - \sigma^{\mathbb{P}}(h)\cdot g\left(\frac{\bar{\alpha}(h)+\beta(h)'f_t-\underline{r}}{\sigma^{\mathbb{P}}(h)}\right)\right] \tag{5.52}$$

in the Ichiue-Ueno model, and

$$TP_t(\tau) = \frac{1}{\tau}\sum_{h=1}^{\tau-1}\left[\sigma^{\mathbb{Q}}(h)\cdot g\left(\frac{a(h)+b(h)'f_t-\underline{r}}{\sigma^{\mathbb{Q}}(h)}\right)-\sigma^{\mathbb{P}}(h)\cdot g\left(\frac{\bar{\alpha}(h)+\beta(h)'f_t-\underline{r}}{\sigma^{\mathbb{P}}(h)}\right)\right] \quad (5.53)$$

in the Wu-Xia model.

5.3.2 Estimating Shadow Rate Models

For estimation, both Ichiue and Ueno (2013) and Wu and Xia (2016) choose the identify the model via the JSZ canonical restrictions, so that $\delta = 0$, $\beta = \iota_n$, $K^{\mathbb{Q}}$ is the zero vector except for its first element $k_\infty^{\mathbb{Q}}$, and $G^{\mathbb{Q}}$ is a matrix in Jordan form, with eigenvalues collected in the vector $\lambda^{\mathbb{Q}}$. Both models estimate a 3-factor model, and Wu and Xia (2016) assume the existence of a repeated eigenvalue in $G^{\mathbb{Q}}$. However, these two models do not assume the existence of a set of observable portfolios of yields, choosing to retain the latent factor specification.

The workhorse estimation method for SRTSMs is maximum likelihood/Bayesian estimation via the extended Kalman filter. To motivate the use of the extended Kalman filter, suppose we have a sample of monthly yields of maturities 1 to $m+1$ months. Then, we can construct m forward rates

$$\mathcal{Y}_t = \begin{pmatrix} f_t^{(1)} \\ \vdots \\ f_t^{(m)} \end{pmatrix} = \begin{pmatrix} 2Y_t(2) - Y_t(1) \\ \vdots \\ (m+1)Y_t(m+1) - mY_t(m) \end{pmatrix}.$$

As with classical ATSMs, we can now express SRTSMs in the following state-space form:

$$\mathcal{Y}_t = L(f_t;\theta^{\mathbb{Q}}) + \Sigma_\eta \cdot \eta_t \quad (5.54)$$

$$f_t = K^{\mathbb{P}} + G^{\mathbb{P}} f_{t-1} + \Sigma \cdot v_t^{\mathbb{P}}, \quad (5.55)$$

where $\theta^{\mathbb{Q}}$ collects the $\mathbb{Q}$-parameters

$$\{k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma\}$$

and

$$L(f_t;\theta^{\mathbb{Q}}) = \begin{pmatrix} \underline{r} + \sigma^{\mathbb{Q}}(1)\cdot g\left(\frac{\bar{a}(1)+b(1)'f_t}{\sigma^{\mathbb{Q}}(1)}\right) \\ \vdots \\ \underline{r} + \sigma^{\mathbb{Q}}(m)\cdot g\left(\frac{\bar{a}(m)+b(m)'f_t}{\sigma^{\mathbb{Q}}(m)}\right) \end{pmatrix}$$

under the model of Ichiue and Ueno (2013); replacing $\bar{a}(h)$ with $a(h)$ yields the model of Wu and Xia (2016). $\Sigma_\eta \cdot \eta_t$ is a vector of forward rate measurement errors, with η_t being standard normally distributed.

Clearly, $L(\cdot;\theta^{\mathbb{Q}})$ is a non-linear function, so we cannot apply the Kalman filter for linear models directly. Therefore, we use the extended Kalman filter, which is the workhorse for estimating non-linear state space models. The non-linearity of the measurement equation means that the resulting filtered values are not optimal, but in practice they turn out to be good enough approximations to the true optimal values.

To start the extended Kalman filter iterations, we define the quantities

$$\begin{aligned} f_{t|t-1} &= \mathbb{E}\left[f_t \mid \mathcal{G}_{t-1}\right] \\ f_{t|t} &= \mathbb{E}\left[f_t \mid \mathcal{G}_t\right] \\ P_{t|t-1} &= \mathrm{Var}\left(f_t \mid \mathcal{G}_{t-1}\right) \\ P_{t|t} &= \mathrm{Var}\left(f_t \mid \mathcal{G}_t\right) \\ \mathcal{Y}_{t|t-1} &= \mathbb{E}\left[\mathcal{Y}_t \mid \mathcal{G}_{t-1}\right] \\ V_{t|t-1} &= \mathbb{E}\left[\mathcal{Y}_t \mid \mathcal{G}_{t-1}\right], \end{aligned}$$

where $\mathcal{G}_t$ is the information contained in the sample $\{\mathcal{Y}_1, \cdots, \mathcal{Y}_t\}$. The algorithm is initialized with

$$f_0 \sim \mathcal{N}\left[f_{0|0},\, P_{0|0}\right],$$

where $f_{0|0}$ and $P_{0|0}$ are chosen as in the usual Kalman filter for the stationary and non-stationary cases.

Suppose that we have obtained $f_{t-1|t-1}$ and $P_{t-1|t-1}$, and that f_{t-1}, η_t and $v_t^{\mathbb{P}}$ are approximately jointly normal given $\mathcal{G}_{t-1}$:

$$\begin{pmatrix} f_{t-1} \\ \eta_t \\ v_t^{\mathbb{P}} \end{pmatrix} \mid \mathcal{G}_{t-1} \overset{\text{approx}}{\sim} \mathcal{N}\left[\begin{pmatrix} f_{t-1|t-1} \\ O_{(n+m)\times 1} \end{pmatrix},\, \mathrm{diag}\left(P_{t-1|t-1}, I_{n+m}\right)\right].$$

Then,

$$f_{t|t-1} = K^{\mathbb{P}} + G^{\mathbb{P}} \cdot f_{t-1|t-1}$$

$$P_{t|t-1} = G^{\mathbb{P}} P_{t-1|t-1} G^{\mathbb{P}\prime} + \Sigma\Sigma'$$

and

$$\begin{pmatrix} f_t \\ \eta_t \end{pmatrix} \mid \mathcal{G}_{t-1} \overset{\text{approx}}{\sim} \mathcal{N}\left[\begin{pmatrix} f_{t|t-1} \\ O_{m\times 1} \end{pmatrix}, \begin{pmatrix} P_{t|t-1} & O_{n\times m} \\ O_{m\times n} & I_m \end{pmatrix} \right],$$

which is the same as in the linear Kalman filter because the transition equation is linear. Now note that

$$\frac{\partial L(x;\theta^{\mathbb{Q}})}{\partial x'} = \begin{pmatrix} \frac{\partial L_1(x;\theta^{\mathbb{Q}})}{\partial x_1} & \dots & \frac{\partial L_1(x;\theta^{\mathbb{Q}})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial L_m(x;\theta^{\mathbb{Q}})}{\partial x_1} & \dots & \frac{\partial L_m(x;\theta^{\mathbb{Q}})}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \Phi\left(\frac{\bar{a}(1)+b(1)'x}{\sigma^{\mathbb{Q}}(1)}\right)\cdot b(1)' \\ \vdots \\ \Phi\left(\frac{\bar{a}(m)+b(m)'x}{\sigma^{\mathbb{Q}}(m)}\right)\cdot b(m)' \end{pmatrix}$$

for any $x \in \mathbb{R}^n$, so that a first order Taylor approximation of $L(f_t;\theta^{\mathbb{Q}})$ around $f_{t|t-1}$ yields

$$L(f_t;\theta^{\mathbb{Q}}) \approx L(f_{t|t-1};\theta^{\mathbb{Q}}) + \frac{\partial L(x;\theta^{\mathbb{Q}})}{\partial x'}|_{x=f_{t|t-1}}\left(f_t - f_{t|t-1}\right).$$

This approximation is precise if $f_t \approx f_{t|t-1}$. Defining

$$\mathbb{L}_{t|t-1} := \frac{\partial L(x;\theta^{\mathbb{Q}})}{\partial x'}|_{x=f_{t|t-1}} = \begin{pmatrix} \Phi\left(\frac{\bar{a}(1)+b(1)'f_{t|t-1}}{\sigma^{\mathbb{Q}}(1)}\right)\cdot b(1)' \\ \vdots \\ \Phi\left(\frac{\bar{a}(m)+b(m)'f_{t|t-1}}{\sigma^{\mathbb{Q}}(m)}\right)\cdot b(m)' \end{pmatrix},$$

we have the linearization

$$\mathcal{Y}_t \approx L(f_{t|t-1};\theta^{\mathbb{Q}}) + \mathbb{L}_{t|t-1}\left(f_t - f_{t|t-1}\right) + \Sigma_\eta \cdot \eta_t. \tag{5.56}$$

Therefore, we end up with the approximations

$$\begin{aligned} \mathcal{Y}_{t|t-1} &\approx L(f_{t|t-1};\theta^{\mathbb{Q}}) \\ V_{t|t-1} &\approx \mathbb{L}_{t|t-1}P_{t|t-1}\mathbb{L}'_{t|t-1} + \Sigma_\eta\Sigma'_\eta. \end{aligned}$$

We can further see from equation (5.56) that f_t and $\mathcal{Y}_t$ are approximately jointly normal conditional on $\mathcal{G}_{t-1}$:

$$\begin{pmatrix} f_t \\ \mathcal{Y}_t \end{pmatrix} \mid \mathcal{G}_{t-1} \overset{\text{approx}}{\sim} \mathcal{N}\left[\begin{pmatrix} f_{t|t-1} \\ \mathcal{Y}_{t|t-1} \end{pmatrix}, \begin{pmatrix} P_{t|t-1} & P_{t|t-1}\mathbb{L}'_{t|t-1} \\ \mathbb{L}_{t|t-1}P_{t|t-1} & V_{t|t-1} \end{pmatrix} \right].$$

The usual updating formula for jointly normally distributed variables tells us that f_t is

approximately normal given $\mathcal{G}_t$, with mean and variance given as

$$f_{t|t} \approx f_{t|t-1} + K_{t|t-1}\left(\mathcal{Y}_t - \mathcal{Y}_{t|t-1}\right)$$
$$P_{t|t} \approx \left[I_n - K_{t|t-1}\mathbb{L}_{t|t-1}\right] P_{t|t-1},$$

where

$$K_{t|t-1} = P_{t|t-1}\mathbb{L}'_{t|t-1}V^{-1}_{t|t-1}$$

is the near-optimal Kalman gain; it is optimal when $\mathcal{Y}_t$ and f_t are actually jointly normally distributed.

Finally, we can see that, because $\mathcal{G}_t$ and f_t are independent of η_{t+1} and $v^{\mathbb{P}}_{t+1}$, f_t is conditionally independent of the two error terms. The approximate normality of f_t given $\mathcal{G}_t$ and the normality of the errors now imply

$$\begin{pmatrix} f_t \\ \eta_{t+1} \\ v^{\mathbb{P}}_{t+1} \end{pmatrix} \mid \mathcal{G}_t \overset{\text{approx}}{\sim} \mathcal{N}\left[\begin{pmatrix} f_{t|t} \\ O_{(n+m)\times 1} \end{pmatrix}, \operatorname{diag}\left(P_{t|t}, I_{n+m}\right)\right].$$

Thus, we can recursively recover the approximate filtered and predictive values from the extended Kalman filter as follows:

$$f_{t|t-1} = K^{\mathbb{P}} + G^{\mathbb{P}} \cdot f_{t-1|t-1} \tag{5.57}$$

$$P_{t|t-1} = G^{\mathbb{P}} P_{t-1|t-1} G^{\mathbb{P}\prime} + \Sigma\Sigma' \tag{5.58}$$

$$\mathbb{L}_{t|t-1} = \begin{pmatrix} \Phi\left(\frac{\bar{a}(1)+b(1)' f_{t|t-1}}{\sigma^{\mathbb{Q}}(1)}\right) \cdot b(1)' \\ \vdots \\ \Phi\left(\frac{\bar{a}(m)+b(m)' f_{t|t-1}}{\sigma^{\mathbb{Q}}(m)}\right) \cdot b(m)' \end{pmatrix} \tag{5.59}$$

$$\mathcal{Y}_{t|t-1} \approx L(f_{t|t-1}; \theta^{\mathbb{Q}}) \tag{5.60}$$

$$V_{t|t-1} = \mathbb{L}_{t|t-1} P_{t|t-1} \mathbb{L}'_{t|t-1} + \Sigma_\eta \Sigma'_\eta \tag{5.61}$$

$$K_{t|t-1} = P_{t|t-1}\mathbb{L}'_{t|t-1}V^{-1}_{t|t-1} \tag{5.62}$$

$$f_{t|t} \approx f_{t|t-1} + K_{t|t-1}\left(\mathcal{Y}_t - \mathcal{Y}_{t|t-1}\right) \tag{5.63}$$

$$P_{t|t} \approx \left[I_n - K_{t|t-1}\mathbb{L}_{t|t-1}\right] P_{t|t-1} \tag{5.64}$$

$$\mathcal{Y}_t \mid \mathcal{G}_{t-1} \overset{\text{approx}}{\sim} \mathcal{N}\left[\mathcal{Y}_{t|t-1},\, V_{t|t-1}\right]. \tag{5.65}$$

with $a(h)$ in place of $\bar{a}(h)$ if we use the Wu-Xia model over the Ichiue-Ueno model. The approximate conditional log-likelihood is now given as

$$\begin{aligned} l(\mathcal{Y}_T, \cdots, \mathcal{Y}_1 \mid \theta) &= \sum_{t=1}^{T} \log f(\mathcal{Y}_t \mid \mathcal{G}_{t-1}; \theta) \\ &\approx -\frac{mT}{2}\log(2\pi) - \frac{1}{2}\left[\sum_{t=1}^{T}\left(\log\left|V_{t|t-1}\right| + \left(\mathcal{Y}_t - \mathcal{Y}_{t|t-1}\right)' V_{t|t-1}^{-1}\left(\mathcal{Y}_t - \mathcal{Y}_{t|t-1}\right)\right)\right]. \end{aligned}$$

5.4 Term Structure Models with Falling Stars

Most of the term structure models developed in the 2000s and early 2010s assumed stationary physical factor dynamics; examples include Dai and Singleton (2000) and Joslin, Singleton, and Zhu (2011). A subset of the literature, however, had consistently pointed out that, because yield factors actually exhibit near-unit root behavior, modeling their physical dynamics as stationary likely incurs severe small-sample bias during estimation. Bauer, Rudebusch, and Wu (2012) was one of the first papers to address this issue, demonstrating that standard Gaussian ATSMs suffer from high degress of small sample bias, in particular predicting unrealistically fast rates of mean reversion for the short rate. In the SRTSM of Christensen and Rudebusch (2016), the physical factor dynamics contain a unit root, not because the authors believe that the factor dynamics are truly non-stationary, but because they claim it helps alleviate estimation bias when the yield factors are close to non-stationary. Bauer and Hamilton (2018) developed a bootstrapping algorithm to obtain more robust standard errors in the presence of precisely this small sample bias.

Bauer and Rudebusch (2020) is the natural culmination of this trend, in which is developed a Gaussian ATSM that explicitly accounts for the non-stationarity and cointegration properties of bond yields. They first establish further stylized facts concerning the role of the real interest rate trend r_t^*, the inflation trend π_t^*, and the nominal interest rate trend i_t^* pertaining to the term structure of interest rates. Subsequently, a Gaussian ATSM that is able to replicate these facts is formulated.

5.4.1 Definition of and Proxies for Macroeconomic Trends

In Bauer and Rudebusch (2020), the long run trend of a macroeconomic variable X_t is defined as its Beveridge-Nelson trend. Formally, suppose that X_t is an I(1) process whose first difference ΔX_t is a weakly stationary causal linear process with MA(∞) representa-

tion

$$\Delta X_t = \delta + \Psi(L)\varepsilon_t.$$

Here, ε_t is an i.i.d. white noise process, and $\Psi(L)$ is an MA(∞) lag polynomial with one-summable coefficients. It follows from the Beveridge-Nelson decomposition that there exists a sequence of absolutely summable coefficients $\{\alpha_j\}_{j\in\mathbb{N}}$ and a mean zero weakly stationary causal linear process $\eta_t = \alpha(L)\varepsilon_t$ such that

$$X_t = \delta t + \Psi(1)\left(\sum_{s=1}^{t}\varepsilon_s\right) + \eta_t + (X_0 - \eta_0)$$

for any $t > 0$. Suppose that the initial values are chosen so that $X_0 - \eta_0$ equals a nonrandom constant μ. Then,

$$X_t^* = \mu + \delta t + \Psi(1)\left(\sum_{s=1}^{t}\varepsilon_s\right)$$

is defined as the Beveridge-Nelson trend of the variable X_t. Similarly, η_t represents the cyclical component of X_t.

When $\delta = 0$, and the initial values are chosen so that there exists a very convenient expression of the Beveridge-Nelson trend as the current forecast of future values of X_t as the forecast horizon goes to infinity. Indeed, this is how Bauer and Rudebusch (2020) define the trend of an I(1) macroeconomic variable. From the above decomposition, we can see that, for any $t > 0$ and $h > 0$,

$$X_{t+h} = \mu + \Psi(1)\left(\sum_{s=1}^{t+h}\varepsilon_s\right) + \eta_{t+h},$$

so that

$$\begin{aligned}\mathbb{E}_t\left[X_{t+h}\right] &= \mu + \Psi(1)\left(\sum_{s=1}^{t+h}\mathbb{E}_t\left[\varepsilon_s\right]\right) + \mathbb{E}_t\left[\eta_{t+h}\right] \\ &= \mu + \Psi(1)\left(\sum_{s=1}^{t}\varepsilon_s\right) + \mathbb{E}_t\left[\eta_{t+h}\right],\end{aligned}$$

where the second equality follows because ε_t is an i.i.d. white noise process. Note also that

$$\begin{aligned}\mathbb{E}_t\left[\eta_{t+h}\right] &= \sum_{j=0}^{\infty}\alpha_j \cdot \mathbb{E}_t\left[\varepsilon_{t+h-j}\right] \\ &= \sum_{j=h}^{\infty}\alpha_j \cdot \varepsilon_{t+h-j}.\end{aligned}$$

Since

$$\left\|\sum_{j=h}^{\infty}\alpha_j\cdot\varepsilon_{t+h-j}\right\|_2=\left(\mathbb{E}\left|\sum_{j=h}^{\infty}\alpha_j\cdot\varepsilon_{t+h-j}\right|^2\right)^{\frac{1}{2}}$$
$$\leq\sum_{j=h}^{\infty}\alpha_j^2\cdot\left\|\varepsilon_{t+h-j}\right\|_2=\left\|\varepsilon_0\right\|_2\cdot\left(\sum_{j=h}^{\infty}\alpha_j^2\right)$$

for any $h>0$, the absolute summability of $\{\alpha_j\}_{j\in\mathbb{N}}$ implies its square summability and

$$\mathbb{E}_t\left[\eta_{t+h}\right]=\sum_{j=h}^{\infty}\alpha_j\cdot\varepsilon_{t+h-j}\xrightarrow{L^2}0$$

as $h\to\infty$. In other words,

$$\mathbb{E}_t\left[X_{t+h}\right]=X_t^*+\mathbb{E}_t\left[\eta_{t+h}\right]\xrightarrow{L^2}X_t^*$$

as $h\to\infty$, so that the Beveridge-Nelson trend X_t^* is the L^2-limit of the forecast $\mathbb{E}_t\left[X_{t+h}\right]$ as the forecast horizion goes to infinity. We write this relationship as

$$X_t^*=\lim_{h\to\infty}\mathbb{E}_t\left[X_{t+h}\right].$$

For this reason, the long run trend X_t^* is sometimes referred to as the "endpoint" of the process X_t. If X_t does not contain any stochastic trend, that is, if $\Psi(1)=0$, then X_t^* is equal to μ, its (constant) long run mean. On the other hand, a non-zero $\Psi(1)$ implies that X_t^* is stochastic. Therefore, I(1) processes without deterministic time trends are sometimes referred to as processes with shifting endpoints, in contrast to I(0) processes, which have constant endpoints equal to their constant long run means.

Using the above definition and formulation of the long run trend of I(1) macroeconomic variables, the following relationship holds if we assume that the real interest rate, nominal interest rate and inflation are I(1) variables with no deterministic time trend:

$$i_t^*=\lim_{h\to\infty}\mathbb{E}_t\left[i_{t+h}\right]=\lim_{h\to\infty}\left(\mathbb{E}_t\left[r_{t+h}\right]+\mathbb{E}_t\left[\mathbb{E}_{t+h}\left[\pi_{t+h+1}\right]\right]\right)$$
$$=\lim_{h\to\infty}\left(\mathbb{E}_t\left[r_{t+h}\right]+\mathbb{E}_t\left[\pi_{t+h+1}\right]\right)=r_t^*+\pi_t^*,$$

where the second equality follows from the Fisher equation and the third from the law of iterated expectations. This shows us that the nominal interest rate trend is just the sum of the real interest rate trend and trend inflation.

We can also study the role of the nominal interest rate trend when it comes to the EH and TP components of long term yields. Recall that τ-maturity yield can be decomposed

as

$$Y_t(\tau) = \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[i_{t+h}\right] + TP_t(\tau),$$

where the first term on the right hand side represents the EH component. The EH component can be further decomposed as

$$\frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[i_{t+h}\right] = i_t^* + \frac{1}{\tau}\sum_{h=0}^{\tau-1}\mathbb{E}_t\left[i_{t+h}^c\right],$$

where i_{t+h}^c is the cyclical component of i_{t+h}, and we used the martingale property of the nominal interest rate trend. In other words, i_t^* loads on the EH component with unity, indicating that the EH component must possess a stochastic trend. This is a feature neglected in many stationary Gaussian ATSMs, leading to the trend in long term yields to be imputed to the term premium instead.

To establish stylized facts about the role of r_t^*, π_t^* and i_t^* in the term structure of interest rates, Bauer and Rudebusch (2020) first procure some empirical proxies of the originally unobservable quantities r_t^* and π_t^*. Exploiting the fact that trend inflation π_t^* can be interpreted macroeconomically as the central bank's inflation target, Bauer and Rudebusch (2020) use the perceived target rate (PTR) from the FRB/US model as their proxy for π_t^*.

On the other hand, the estimation of the real interest rate trend r_t^* is a hotly debated topic in the empirical macroeconomic literature, as r_t^* proves to be very model-sensitive. To avoid relying on one specific model, Bauer and Rudebusch (2020) choose to take as their proxy for r_t^* the average estimates obtained from three different types of models. The first are empirical time-series model estimated via Bayesian methods, exemplified by Del Negro et al. (2017) [8]. A second class of models collects those that extract r_t^* from formal DSGE models, or at least semi-structural models, as in Laubach and John C. Williams (2003). Finally, Bauer and Rudebusch (2020) also provide their own estimates

[8] This model, referred to as DGGT, is important in its own right, so we provide here a brief summary of its contents.

In DGGT, the natural rate of interest, which is defined as the real interest rate in the absence of monetary policy, is extracted from a multivariate unobserved components model containing variables such as short/long term nominal interest rates, inflation, survey expectations, and the rate of return on corporate bonds. The results are found to be comparable to those of a formal DSGE model.

The authors find that the marked secular decline in the natural rate since the Great Recession can be attributed to the rise in convenience yields, which are the premia investors pay in exchange for safety and liquidity. Specifically, in exchange for the greater safety and liquidity provided by treasury bonds, investors accept a lower rate of return compared to corporate bonds with the same return structure, which explains why the trend of the real return to treasury bonds have plummeted in response to the savings glut following the onset of the Great Recession. This has proven to be one of the most influential explanations for the secular decline in the real interest rate trend. Indeed, macro trends in Bauer and Rudebusch (2020) are called "falling stars" precisely because of the marked decline in r_t^*.

of r_t^*, obtained from simple trend-cycle decompositions of the real interest rate.

5.4.2 Stylized Facts about Macroeconomic Trends and the Yield Curve

Using the empirical proxies for r_t^* and π_t^*, Bauer and Rudebusch (2020) establish the following stylized facts:

Fact 1: The Nominal Interest Rate Trend is the Common Trend among Yields

First, Bauer and Rudebusch (2020) find that the nominal interest rate trend is precisely the common trend affecting yields. As far back as Campbell and Shiller (1987), it has been known that yields are best modeled as cointegrated with a single common component, while yield spreads are largely stationary. The contribution in Bauer and Rudebusch (2020) is that they provide evidence pointing to i_t^* being this common component.

To show this, Bauer and Rudebusch (2020) run regressions of the form

$$Y_t(\tau) = \beta_0 + \beta_1' X_t + \varepsilon_t,$$

where τ represents the maturity corresponding to 10 years and X_t contains the macroeconomic trends of interest. If the trends in X_t are truly the common trends driving the yields, then the stochastic trend of $Y_t(\tau)$ would be given as a linear combination of X_t, meaning that there exist values of β_0, β_1 such that $Y_t(\tau) - \beta_0 - \beta_1' X_t$ is stationary[9]. Based on this intuition, Bauer and Rudebusch (2020) estimate β_0, β_1 by the Dynamic OLS method of Saikkonen (1992), and test whether the resulting residuals are stationary[10].

The combinations of X_t used in Bauer and Rudebusch (2020) are

$$X_t = \{\pi_t^*\}, \quad \{r_t^*, \pi_t^*\}, \quad \{i_t^*\}.$$

Trend inflation is always included as part of the regressors, since many works in the literature document the predictive power of trend inflation π_t^* for excess bond returns. In particular, Cieslak and Povala (2015) take inspiration from the Fisher equation and

[9] This is all predicated on the assumption that yields do not contain a deterministic time trend.

[10] OLS estimates of β_0, β_1 are also consistent for the true parameter values, albeit with an asymptotic distribution that contains nuisance parameters, and Phillips and Ouliaris (1990) derives the asymptotic distribution of the Dickey-Fuller test statistic of the OLS residuals (this test is referred to as the Phillips-Ouliaris test).

The reason the authors in Bauer and Rudebusch (2020) opt for Dynamic OLS instead of standard OLS is probably to obtain estimates of β_0, β_1 with pivotal asymptotic distributions, which allows them to conduct hypothesis tests concerning β_0 and β_1.

decompose nominal yields into a trend inflation component and a component that affects the real part of the yields. Bauer and Rudebusch (2020) find that, if only π_t^* is included in X_t, then there is insufficient evidence to conclude that the residuals are stationary, based on unit root tests such as the ADF and PP tests, as well as more general stationarity tests like the low-frequency cointegration test of Müller and Watson (2013). On the other hand, if both r_t^* and π_t^* are included as regressors, we can conclude on the basis of the above tests that the residual process is stationary. Finally, including i_t^* as the sole regressor yields similar results, so that i_t^* can serve as the single common component among yields of various maturities.

In addition, the Dynamic OLS estimate of β_1 from putting $X_t = i_t^*$ reveals that the loading of i_t^* on the 10 year yield is significantly greater than 1. This suggests that the common component i_t^* loads on yields with a loading greater than unity, that is, a unit increase in the common trend i_t^* results in a larger than unit increase in the yields. Since i_t^* loads on the EH component with unity, this implies that i_t^* loads positively on the term premium. In other words, even though the EH and TP components now share the common trend in long term yields, the term premium remains non-stationary to some degree.

Fact 2: The Nominal Interest Rate Trend helps predict Excess Bond Returns

To investigate the relationship between the common yield trend i_t^* and excess bond returns, Bauer and Rudebusch (2020) test whether β_2 is equal to 0 in the regression

$$exr_{t+1}^{(\tau)} = \beta_0 + \beta_1' \mathcal{P}_t + \beta_2 \cdot \cdot i_t^* + u_{t+1}, \tag{5.66}$$

where $exr_{t+1}^{(\tau)}$ is the one-period ahead excess bond return for the τ-maturity bond defined as

$$exr_{t+1}^{(\tau)} = (\tau - 1)Y_{t+1}(\tau - 1) - \tau \cdot Y_t(\tau) - r_t,$$

and $\mathcal{P}_t$ contain the first three principal components of the yield curve. Essentially, Bauer and Rudebusch (2020) test whether the information in i_t^* is spanned by the yield curve by running a regression similar to that in equation (5.10), where the information contained in the yield curve is controlled for using the PCs rather than the CP factor. Equation (5.66) is estimated using the bootstrap method of Bauer and Hamilton (2018), which helps control for small sample bias in the presence of trending regressors and errors that are not strictly exogenous.

Results show that β_2 is significantly lower than 0, indicating that an increase in the common trend i_t^* lowers excess bond returns, even when the information in the yield curve

has been controlled for. This suggests that i_t^* contains information about excess bond returns that is unspanned by the yield curve. Another interesting finding is that, unlike the results in Campbell and Shiller (1991) and Cochrane and Piazzesi (2005), which find that the yield spread or the CP factor contain relevant information about excess bond returns, the level factor becomes a strong predictor for excess bond returns once i_t^* is included in the regression (5.66).

5.4.3 No-Arbitrage under Falling Stars

We have seen how Bauer and Rudebusch (2020) established stylized facts indicating that i_t^* is not only the common trend moving all yields, but also that i_t^* contains unspanned information about the yield curve. This motivates them to formulate a Gaussian ATSM with i_t^*. This ATSM, called the falling stars (FS) model, is formally almost identical to a macro-finance ATSM with i_t^* assuming the role of the unspanned macro factor.

As with the usual Gaussian ATSM, the FS model is a model of no-arbitrage. Under the no-arbitrage condition, there exists an SDF process $\{\mathcal{M}_{t+1}\}_{t\in\mathbb{N}}$ and a risk-neutral measure $\mathbb{Q}$ under which the price of each asset is its expected one-period ahead payoff discounted by the (nominal) short rate i_t. We assume that the SDF is given in the exponential-affine form

$$\mathcal{M}_{t+1} = \exp\Big(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t' v_{t+1}^{\mathbb{P}}\Big),$$

where $v_{t+1}^{\mathbb{P}}$ is an n-dimensional random vector of risk factors, which we assume to be standard normally distributed. Defining $v_{t+1}^{\mathbb{Q}}$ as

$$v_{t+1}^{\mathbb{Q}} = v_{t+1}^{\mathbb{P}} + \lambda_t,$$

$v_{t+1}^{\mathbb{Q}}$ is standard normally distributed under the risk-neutral measure.

Bauer and Rudebusch (2020) choose the JSZ identification scheme, so that, in terms of the n latent factors f_t, the short rate and risk-neutral factor dynamics of the model are given as

$$\begin{aligned} i_t &= \iota' f_t \\ f_{t+1} &= \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ O_{(n-1)\times 1} \end{pmatrix} + J^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where $\lambda^{\mathbb{Q}}$ contains the n distinct real eigenvalues that determine the Jordan matrix $J^{\mathbb{Q}}$[11].

[11] Bauer and Rudebusch (2020) assume that the eigenvalues collected in $\lambda^{\mathbb{Q}}$ are all less than unity, in order to prevent the counterfactual implication of bond prices and forward rates diverging to infinity as

The sample of m yields, $\mathcal{Y}_t$, is then given as an affine function of the latent factors, where the factor loadings and intercept are functions of the $\mathbb{Q}$-parameters $k_\infty^\mathbb{Q}, J^\mathbb{Q}$ and Σ:

$$\mathcal{Y}_t = \mathcal{A} + \mathcal{B} \cdot f_t.$$

Bauer and Rudebusch (2020) further assume that there exists an n-dimensional portfolio $\mathcal{P}_t = W\mathcal{Y}_t$ of yields that is observed without error. This indicates that $\mathcal{P}_t$ is an invariant affine transformation of f_t:

$$\mathcal{P}_t = W\mathcal{A} + W\mathcal{B} \cdot f_t.$$

This allows the short rate and risk-neutral dynamics can be written in terms of $\mathcal{P}_t$ as the factors:

$$i_t = \delta_{0,\mathcal{P}} + \delta_{1,\mathcal{P}}' f_t$$

$$\mathcal{P}_{t+1} = K_\mathcal{P}^\mathbb{Q} + G_\mathcal{P}^\mathbb{Q} \mathcal{P}_t + \Sigma_\mathcal{P} \cdot v_{t+1}^\mathbb{Q},$$

where

$$\delta_\mathcal{P} = -\iota' \left[W\mathcal{B}\right] \tag{5.67}$$

$$\beta_\mathcal{P} = \left[W\mathcal{B}\right]^{-1\prime} \iota \tag{5.68}$$

$$K_\mathcal{P}^\mathbb{Q} = \left[W\mathcal{B}\right] \left(I_n - J^\mathbb{Q}\right) \left[W\mathcal{B}\right]^{-1} \cdot W\mathcal{A} + \left[W\mathcal{B}\right]^{-1} \begin{pmatrix} k_\infty^\mathbb{Q} \\ O_{(n-1)\times 1} \end{pmatrix} \tag{5.69}$$

$$G_\mathcal{P}^\mathbb{Q} = \left[W\mathcal{B}\right] J^\mathbb{Q} \left[W\mathcal{B}\right]^{-1} \tag{5.70}$$

$$\Sigma_\mathcal{P} = \left[W\mathcal{B}\right] \Sigma. \tag{5.71}$$

The sample yields are now given as affine functions of the observed factors $\mathcal{P}_t$:

$$\mathcal{Y}_t = \mathcal{A}_\mathcal{P} + \mathcal{B}_\mathcal{P} \cdot \mathcal{P}_t,$$

where

$$\mathcal{A}_\mathcal{P} = \mathcal{A} - \mathcal{B} \left[W\mathcal{B}\right]^{-1} W\mathcal{A}$$

the maturity increases.

This is in contrast to the AFNS model, in which $G^\mathbb{Q}$ does contain a unit root. However, the tradeoff for the unfortunate implications is the ability to easily interpret the factors and a superb fit for the yield curve.

$$\mathcal{B}_{\mathcal{P}} = \mathcal{B}\left[W\mathcal{B}\right]^{-1}.$$

So far, the FS model has closely followed the JSZ model. The main difference between the two models comes with the physical factor dynamics. In the FS model, we assume that the factors $\mathcal{P}_t$ are non-stationary; specifically, the trend-cycle representation of $\mathcal{P}_t$ is given as

$$\mathcal{P}_t = \underbrace{\mu + \gamma \cdot \tau_t}_{\mathcal{P}_t^*} + \tilde{\mathcal{P}}_t, \tag{5.72}$$

where $\mu, \gamma \in \mathbb{R}^{n\times 1}$. τ_t represents the single trend that drives the n factors and therefore the yields, and it follows a random walk:

$$\tau_t = \tau_{t-1} + \sigma_\eta \cdot \eta_t.$$

Meanwhile, the cyclical component $\tilde{\mathcal{P}}_t$ of $\mathcal{P}_t$ is assumed to follow a mean zero weakly stationary VAR(1) process:

$$\tilde{\mathcal{P}}_t = \Phi \cdot \tilde{\mathcal{P}}_{t-1} + \Sigma_u \cdot u_t,$$

where the eigenvalues of Φ are all contained in the unit circle. The error processes η_t and u_t are i.i.d. standard normal, and assumed to be mutually independent. In this context, equation (5.72) specifies the factors $\mathcal{P}_t$ as following a trend-cycle VAR(1).

Note that $\mathcal{P}_t^*$ is the Beveridge-Nelson trend of the model, since τ_t, being a random walk, is a martingale and $\tilde{\mathcal{P}}_t$ has an absolutely summable causal linear process representation:

$$\mathbb{E}_t\left[\mathcal{P}_{t+h}\right] \xrightarrow{L^2} \mu + \gamma \cdot \tau_t = \mathcal{P}_t^*$$

as $h \to \infty$.

It remains to match the two stylized facts established above. First, we must ensure that the common trend τ_t equals the nominal interest rate trend i_t^*. From the short rate dynamics, we can see that

$$i_t^* = \delta_{\mathcal{P}} + \beta_{\mathcal{P}}' \mathcal{P}_t^*.$$

Substituting equation (5.72) into the above equation, we end up with

$$i_t^* = \delta_{\mathcal{P}} + \beta_{\mathcal{P}}' \mu + \beta_{\mathcal{P}}' \gamma \cdot \tau_t.$$

To match the stylized facts, we must thus impose the identification restrictions

$$\delta_{\mathcal{P}} + \beta_{\mathcal{P}}'\mu = 0 \tag{5.73}$$

$$\beta_{\mathcal{P}}'\gamma = 1. \tag{5.74}$$

To match the second stylized fact, we direct our focus to the market prices of risk in the FS model:

$$\begin{aligned}\lambda_t &= \Sigma_{\mathcal{P}}^{-1}\left[\mathbb{E}_t\left[\mathcal{P}_{t+1}\right] - \mathbb{E}_t^{\mathbb{Q}}\left[\mathcal{P}_{t+1}\right]\right] \\ &= \Sigma_{\mathcal{P}}^{-1}\left[\mathcal{P}_t^* + \Phi\cdot\tilde{\mathcal{P}}_t - K_{\mathcal{P}}^{\mathbb{Q}} - G_{\mathcal{P}}^{\mathbb{Q}}\mathcal{P}_t\right] \\ &= -\Sigma_{\mathcal{P}}^{-1}K_{\mathcal{P}}^{\mathbb{Q}} + \Sigma_{\mathcal{P}}^{-1}\left(I_n - \Phi\right)\mathcal{P}_t^* + \Sigma_{\mathcal{P}}^{-1}\left[\Phi - G_{\mathcal{P}}^{\mathbb{Q}}\right]\mathcal{P}_t.\end{aligned}$$

Since $I_n - \Phi$ is non-zero, even when the factors $\mathcal{P}_t$ are controlled for, the trend component $\mathcal{P}_t^*$, and by extension i_t^*, still loads on the market prices of risk λ_t. Since excess bond returns are linear in λ_t, this indicates that, in the FS model, i_t^* remains unspanned, which is precisely the content of the second stylized fact.

In contrast, the model-implied yields $\mathcal{Y}_t$ are affine functions of $\mathcal{P}_t$ only. In other words, we cannot recover the trend $\mathcal{P}_t^*$ and by extension i_t^* separately from the cyclical component $\tilde{\mathcal{P}}_t$ by inverting the current yields. This demonstrates that the FS model satisfies the knife-edge condition of unspanned macro-finance ATSMs, and that i_t^* serves the same role here as an unspanned macro factor.

5.4.4 Estimating the Falling Stars Model

For the purposes of estimation, the FS model can be written as a state-space model with $(i_t^*, \mathcal{P}_t')'$ serving as the factors:

$$\mathcal{Y}_t = \mathcal{A}_{\mathcal{P}} + \mathcal{B}_{\mathcal{P}}\mathcal{P}_t + \Sigma_e \cdot e_t$$

$$\begin{pmatrix} i_t^* \\ \mathcal{P}_t \end{pmatrix} = \begin{pmatrix} 0 \\ (I_n - \Phi)\mu \end{pmatrix} + \begin{pmatrix} 1 & O_{1\times n} \\ (I_n - \Phi)\gamma & \Phi \end{pmatrix}\begin{pmatrix} i_{t-1}^* \\ \mathcal{P}_{t-1} \end{pmatrix} + \begin{pmatrix} \sigma_\eta & O_{1\times n} \\ \gamma\cdot\sigma_\eta & \Sigma_u \end{pmatrix}\begin{pmatrix} \eta_t \\ u_t \end{pmatrix},$$

where e_t is an i.i.d. yield pricing error process that is standard normally distributed. Estimation of the model should now proceed via ML or Bayesian methods that make use of the Kalman filter, but the famous "pile-up" problem makes the estimation of the trend component difficult. Del Negro et al. (2017), for instance, circumvents this problem by imposing a very tight prior on σ_η around 0, but in general, estimation using the Kalman filter proves challenging. For this reason, Bauer and Rudebusch (2020) suggest using the proxy for i_t^* discussed earlier in place of i_t^* in the above state space model. This means that

the factors $(i_t^*, \mathcal{P}_t')'$ are now all observed, which greatly facilitates estimation. The model estimated in this manner is referred to as the observed shifting endpoint (OSE) model. Bauer and Rudebusch (2020) also estimate the model through the traditional Kalman filter method, which they call the estimated shifting endpoint (ESE) model.

Bibliography

Adrian, T., R. K. Crump, and E. Moench (2013). "Pricing the Term Structure with Linear Regressions". In: *Journal of Financial Economics* 110, pp. 110–138.

Ahn, Dong-Hyun, Robert F Dittmar, and A Ronald Gallant (2002). "Quadratic term structure models: Theory and evidence". In: *The Review of financial studies* 15.1, pp. 243–288.

Ahn, Seung C and Alex R Horenstein (2013). "Eigenvalue ratio test for the number of factors". In: *Econometrica* 81.3, pp. 1203–1227.

Ang, Andrew and Monika Piazzesi (2003). "A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables". In: *Journal of Monetary economics* 50.4, pp. 745–787.

Bai, Jushan (2003). "Inferential Theory for Factor Models of Large Dimensions". In: *Econometrica* 71.1, pp. 135–171. (Visited on 08/23/2023).

Bai, Jushan and Serena Ng (2002). "Determining the number of factors in approximate factor models". In: *Econometrica* 70.1, pp. 191–221.

Bansal, Ravi and Hao Zhou (2002). "Term Structure of Interest Rates with Regime Shifts". In: *The Journal of Finance* 57.5, pp. 1997–2043.

Barigozzi, Matteo and Matteo Luciani (2019). "Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the EM algorithm". In: *arXiv preprint arXiv:1910.03821*.

Bauer, Michael D. and James D Hamilton (2018). "Robust bond risk premia". In: *The Review of Financial Studies* 31.2, pp. 399–448.

Bauer, Michael D. and Glenn D. Rudebusch (2016). "Monetary Policy Expectations at the Zero Lower Bound". In: *Journal of Money, Credit and Banking* 48.7, pp. 1439–1465.

— (2020). "Interest Rates under Falling Stars". In: *American Economic Review* 110.5, pp. 1316–54.

Bauer, Michael D., Glenn D. Rudebusch, and Jing Cynthia Wu (2012). "Correcting Estimation Bias in Dynamic Term Structure Models". In: *Journal of Business & Economic Statistics* 30.3, pp. 454–467. ISSN: 07350015. (Visited on 08/23/2023).

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz (2005). "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach". In: *The Quarterly journal of economics* 120.1, pp. 387–422.

Black, Fischer (1995). "Interest Rates as Options". In: *The Journal of Finance* 50.5, pp. 1371–1376.

Campbell, John Y and John H. Cochrane (1999). "By force of habit: A consumption-based explanation of aggregate stock market behavior". In: *Journal of political Economy* 107.2, pp. 205–251.

Campbell, John Y and Robert J Shiller (1987). "Cointegration and Tests of Present Value Models". In: *Journal of Political Economy* 95.5, pp. 1062–1088.

— (1988). "The dividend-price ratio and expectations of future dividends and discount factors". In: *The Review of Financial Studies* 1.3, pp. 195–228.

— (1991). "Yield Spreads and Interest Rate Movements: A Bird's Eye View". In: *The Review of Economic Studies* 58.3, pp. 495–514.

Cheridito, Patrick, Damir Filipovic, and Robert L. Kimmel (2007). "Market Price of Risk Specifications for Affine Models: Theory and Evidence". In: *Journal of Financial Economics* 83.1, pp. 123–170.

Chib, Siddhartha and Kyu Ho Kang (2013). "Change-points in affine arbitrage-free term structure models". In: *Journal of Financial Econometrics* 11.2, pp. 302–334.

Christensen, Jens H. E. (2013). *A Regime-Switching Model of the Yield Curve at the Zero Bound*. Tech. rep. FRB of San Francisco.

Christensen, Jens H. E., Francis X. Diebold, and Glenn D. Rudebusch (2011). "The Affine Arbitrage-free Class of Nelson–Siegel Term Structure Models". In: *Journal of Econometrics* 164.1, pp. 4–20.

Christensen, Jens H. E. and Glenn D. Rudebusch (2016). "Modeling Yields at the Zero Lower Bound: Are Shadow Rates the Solution?" In: *Advances in Econometrics* 35, pp. 75–125.

Cieslak, Anna and Pavol Povala (2015). "Expected returns in Treasury bonds". In: *The Review of Financial Studies* 28.10, pp. 2859–2901.

Cochrane, John H. and Monika Piazzesi (2005). "Bond Risk Premia". In: *American Economic Review* 95.1, pp. 138–160.

— (2008). "Decomposing the Yield Curve". In: *Unpublished manuscript*.

Cox, John C, Jonathan E Ingersoll, and Stephen A Ross (1985). "A Theory of the Term Structure of Interest Rates". In: *Econometrica* 53.2, p. 385.

Dai, Qiang and Kenneth J. Singleton (2000). "Specification Analysis of Affine Term Structure Models". In: *The Journal of Finance* 55.5, pp. 1943–1978.

— (2002). "Expectation Puzzles, Time-varying Risk Premia, and Affine Models of the Term Structure". In: *Journal of Financial Economics* 63.3, pp. 415–441.

Dai, Qiang, Kenneth J. Singleton, and Wei Yang (2007). "Regime Shifts in a Dynamic Term Structure Model of U.S. Treasury Bond Yields". In: *Review of Financial Studies* 20.5, pp. 1669–1706.

Del Negro, Marco et al. (2017). "Safety, Liquidity, and the Natural Rate of Interest". In: *Brookings Papers on Economic Activity* 2017.1, pp. 235–316.

Diebold, Francis X. and Canlin Li (2006). "Forecasting the Term Structure of Government Bond Yields". In: *Journal of Econometrics* 130.2, pp. 337–364.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin (2011). "A Two-step Estimator for Large Approximate Dynamic Factor Models based on Kalman Filtering". In: *Journal of Econometrics* 164.1, pp. 188–205.

— (2012). "A quasi–maximum likelihood approach for large, approximate dynamic factor models". In: *Review of economics and statistics* 94.4, pp. 1014–1024.

Duffee, Gregory R (2002). "Term premia and interest rate forecasts in affine models". In: *The Journal of Finance* 57.1, pp. 405–443.

Duffie, Darrell and Rui Kan (1996). "A yield-factor model of interest rates". In: *Mathematical finance* 6.4, pp. 379–406.

Durbin, James and Siem Jan Koopman (2012). *Time series analysis by state space methods*. Vol. 38. OUP Oxford.

Fama, Eugene F and Robert R Bliss (1987). "The information in long-maturity forward rates". In: *The American Economic Review*, pp. 680–692.

Goliński, Adam and Peter Spencer (2021). "Estimating the term structure with linear regressions: Getting to the roots of the problem". In: *Journal of Financial Econometrics* 19.5, pp. 960–984.

Gurkaynak, Refet S., Brian Sack, and Jonathan H. Wright (2007). "The US Treasury Yield Curve: 1961 to the Present". In: *Journal of Monetary Economics* 54.8, pp. 2291–2304.

Hamilton, James D (2020). *Time series analysis*. Princeton university press.

Hamilton, James D and Jing Cynthia Wu (2012). "Identification and Estimation of Gaussian Affine Term Structure Models". In: *Journal of Econometrics* 168.2, pp. 315–331.

Hansen, Lars Peter and Scott F Richard (1987). "The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models". In: *Econometrica: Journal of the Econometric Society*, pp. 587–613.

Hördahl, Peter and Oreste Tristani (2019). "Modelling Yields at the Lower Bound through Regime Shifts". In: *Available at SSRN 3464245*.

Ichiue, Hibiki and Yoichi Ueno (2013). *Estimating Term Premia at the Zero Bound: An Analysis of Japanese, US, and UK Yields*. Tech. rep. Bank of Japan.

Joslin, Scott, Anh Le, and Kenneth J. Singleton (2013). "Why Gaussian macro-finance term structure models are (nearly) unconstrained factor-VARs". In: *Journal of Financial Economics* 109.3, pp. 604–622.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton (2014). "Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks". In: *The Journal of Finance* 69.3, pp. 1197–1233.

Joslin, Scott, Kenneth J. Singleton, and Haoxiang Zhu (2011). "A New Perspective on Gaussian Dynamic Term Structure Models". In: *The Review of Financial Studies* 24.3, pp. 926–970.

Krippner, Leo (2013). "Measuring the Stance of Monetary Policy in Zero Lower Bound Environments". In: *Economics Letters* 118.1, pp. 135–138.

Laubach, Thomas and John C. Williams (2003). "Measuring the Natural Rate of Interest". In: *Review of Economics and Statistics* 85.4, pp. 1063–1070.

Litterman, Robert B and Jose Scheinkman (1991). "Common factors affecting bond returns". In: *The journal of fixed income* 1.1, pp. 54–61.

Liu, Yan and Jing Cynthia Wu (2021). "Reconstructing the yield curve". In: *Journal of Financial Economics* 142.3, pp. 1395–1425.

Ludvigson, Sydney C and Serena Ng (2009). "Macro factors in bond risk premia". In: *The Review of Financial Studies* 22.12, pp. 5027–5067.

Markowitz, Harry (1952). "Portfolio Selection". In: *The Journal of Finance* 7.1, pp. 77–91. ISSN: 00221082, 15406261.

Müller, Ulrich K and Mark W Watson (2013). "Low-frequency robust cointegration testing". In: *Journal of Econometrics* 174.2, pp. 66–81.

Nelson, Charles R and Andrew F Siegel (1987). "Parsimonious modeling of yield curves". In: *Journal of business*, pp. 473–489.

Newey, W and D MacFadden (1994). "Large Sample Estimation and Hypothesis Testing, Chapter 36". In: *Handbook of Econometrics Vol* 4.

Niu, Linlin and Gengming Zeng (2012). "The Discrete-Time Framework of the Arbitrage-Free Nelson-Siegel Class of Term Structure Models". In: *Available at SSRN 2015858*.

Phillips, Peter CB and Sam Ouliaris (1990). "Asymptotic properties of residual based tests for cointegration". In: *Econometrica: journal of the Econometric Society*, pp. 165–193.

Ross, Stephen A (1976). "The arbitrage theory of capital asset pricing". In: *Journal of Economic Theory* 13.3, pp. 341–360. ISSN: 0022-0531.

Saikkonen, Pentti (1992). "Estimation and testing of cointegrated systems by an autoregressive approximation". In: *Econometric theory* 8.1, pp. 1–27.

Singleton, Kenneth J. (2006). *Empirical dynamic asset pricing: model specification and econometric assessment*. Princeton University Press.

Swanson, Eric T and John C Williams (2014). "Measuring the Effect of the Zero Lower Bound on Medium-and Longer-term Interest Rates". In: *American Economic Review* 104.10, pp. 3154–3185.

Vasicek, Oldrich (1977). "An equilibrium characterization of the term structure". In: *Journal of financial economics* 5.2, pp. 177–188.

Wu, Jing Cynthia and Fan Dora Xia (2016). "Measuring the macroeconomic impact of monetary policy at the zero lower bound". In: *Journal of Money, Credit and Banking* 48.2-3, pp. 253–291.

Appendices

A Consistency of Non-Parametric Estimator of Nelson-Siegel Model

Here we show that the non-parametric estimators $\overline{\kappa}$ and $\overline{F}$ of the Nelson-Siegel decay parameter and factors, obtained as

$$\begin{aligned}\overline{\kappa} &= \underset{\kappa\in[\epsilon,1-\epsilon]}{\operatorname{argmin}} \quad \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y}M_{\Lambda(\kappa)}\mathcal{Y}'\right)\\ \overline{F} &= \mathcal{Y}\Lambda(\overline{\kappa})\left(\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right)^{-1}\end{aligned}$$

are consistent for the true values κ_0 and F^0 of the decay parameter and factors under certain assumptions, where

$$M_{\Lambda(\kappa)} = I_m - \Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'.$$

First, we need to specify how we determine the maturities to be included in the sample as the cross-sectional dimension m goes to infinity. Given the context of the yield curve, it is unrealistic to assume that the sample contains yields of maturities 1 to m. A solution is to let τ_{max} be, in months, the longest possible sample maturity (often 120 or 360 months) and assume that we have a sample on yields of maturities $\tau_1, \cdots, \tau_m$, where

$$\tau_i = \frac{\tau_{max}}{m} i$$

for any $1 \leq i \leq m$. Thus, a cross sectional size of τ_{max} indicates that the sample contains yields of maturities 1 to τ_{max} months, and a higher cross-sectional dimension means that the sample contains yields that mature in the middle of the month. This normalization

ensures that the limit

$$\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m} = \begin{pmatrix} 1 & \frac{1}{m}\sum_{i=1}^{m}\beta_2(\tau_i;\kappa) & \frac{1}{m}\sum_{i=1}^{m}\beta_3(\tau_i;\kappa) \\ \frac{1}{m}\sum_{i=1}^{m}\beta_2(\tau_i;\kappa) & \frac{1}{m}\sum_{i=1}^{m}\beta_2(\tau_i;\kappa)^2 & \frac{1}{m}\sum_{i=1}^{m}\beta_2(\tau_i;\kappa)\beta_3(\tau_i;\kappa) \\ \frac{1}{m}\sum_{i=1}^{m}\beta_3(\tau_i;\kappa) & \frac{1}{m}\sum_{i=1}^{m}\beta_2(\tau_i;\kappa)\beta_3(\tau_i;\kappa) & \frac{1}{m}\sum_{i=1}^{m}\beta_3(\tau_i;\kappa)^2 \end{pmatrix}$$

converges to a positive definite matrix $\Omega(\kappa)$ for any $\kappa \in [\epsilon, 1-\epsilon]$. The positive definiteness of $\Omega(\kappa)$, as well as the uniformity of the convergence, is confirmed numerically.

We can now make the following assumptions:

A1. Stationarity of Factors

The true factor process $\{f_t^0\}_{t\in\mathbb{Z}}$ is assumed to be be weakly stationary as well as mean and variance ergodic, so that

$$\frac{F^{0\prime}\iota_T}{T} = \frac{1}{T}\sum_{t=1}^{T} f_t^0 \xrightarrow{p} \mu_F$$
$$\frac{F^{0\prime}F}{T} = \frac{1}{T}\sum_{t=1}^{T} f_t^0 f_t^{0\prime} \xrightarrow{p} \Omega_F$$

as $T \to \infty$ for some vector $\mu_F \in \mathbb{R}^{3\times 1}$ and positive definite 3×3 matrix Ω_F, where ι_T is the T-dimensional vector comprised of 1s.

A2. Stationarity and Cross-Sectional Independence of Errors

The measurement error process $\{e_{it}\}_{t\in\mathbb{Z}}$ associated with yields of maturity i is weakly stationary. Furthermore, $\{e_{it}\}_{t\in\mathbb{Z}}$ and $\{e_{jt}\}_{t\in\mathbb{Z}}$ are independent processes for any $i \neq j$, and

$$\sup_{i\in N_+,\, t\in\mathbb{Z}} \mathbb{E}|e_{it}|^4 < +\infty.$$

We also assume that the error processes are homoskedastic, that is, $\mathbb{E}|e_{it}|^2 = \sigma^2 > 0$ for any i, t.

A3. Uniform Convergence of Factor Loadings

For any $\kappa \in [\epsilon, 1-\epsilon]$,

$$\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m} \to \Omega(\kappa)$$
$$\frac{\Lambda(\kappa)'\Lambda^0}{m} \to \Omega_0(\kappa).$$

as $m \to \infty$. Furthermore, this convergence is uniform on $[\epsilon, 1-\epsilon]$, which implies that the suprema

$$\sup_{\kappa \in [\epsilon, 1-\epsilon]} \left\| \frac{\Lambda(\kappa)'\Lambda(\kappa)}{m} \right\|$$
$$\sup_{\kappa \in [\epsilon, 1-\epsilon]} \left\| \frac{\Lambda(\kappa)'\Lambda^0}{m} \right\|$$

are bounded.

We first show that $\overline{\kappa}$ is consistent for κ_0. To this end, we rely on the general consistency result in Newey and MacFadden (1994). Denote the parameter space by

$$\Theta = [\epsilon, 1-\epsilon],$$

which is a compact set, and define

$$V_{m,T}(\kappa) = \frac{1}{mT} \operatorname{tr}\left(\mathcal{Y} M_{\Lambda(\kappa)} \mathcal{Y}'\right)$$

for any m, T and $\kappa \in [\epsilon, 1-\epsilon]$. In Newey and MacFadden (1994), it is shown that, if:

i) Θ is compact,

ii) $V_{m,T}(\kappa)$ converges in probability to some continuous function $V_0 : \Theta \to \mathbb{R}$ uniformly on Θ as $m, T \to \infty$, that is,

$$\sup_{\kappa \in \Theta} \left| V_{m,T}(\kappa) - V_0(\kappa) \right| \xrightarrow{p} 0$$

as $m, T \to \infty$, and

iii) V_0 is uniquely minimized at κ_0,

then the minimizer $\overline{\kappa}$ of $V_{m,T}(\kappa)$ converges in probability to the true value κ_0 of the decay parameter. Here we construct the function V_0 that satisfies the two conditions above, which will conclude the proof.

First note that

$$\mathcal{Y} = F^0 \Lambda^{0\prime} + e,$$

and as such that, for any fixed $\kappa \in \Theta$,

$$\begin{aligned} V_{m,T}(\kappa) = \frac{1}{mT}\operatorname{tr}\left(\mathcal{Y} M_{\Lambda(\kappa)} \mathcal{Y}'\right) &= \frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)} \mathcal{Y}'\mathcal{Y}\right) \\ &= \frac{1}{mT}\operatorname{tr}\left[M_{\Lambda(\kappa)}\left(\Lambda^0 F^{0\prime} + e'\right)\left(F^0 \Lambda^{0\prime} + e\right)\right] \\ &= \operatorname{tr}\left[\left(\frac{\Lambda^{0\prime}\Lambda^0}{m}\right)\left(\frac{F^{0\prime}F^0}{T}\right)\right] \\ &- \operatorname{tr}\left[\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\left(\frac{\Lambda(\kappa)'\Lambda^0}{m}\right)\left(\frac{F^{0\prime}F^0}{T}\right)\left(\frac{\Lambda(\kappa)'\Lambda^0}{m}\right)'\right] \\ &+ 2\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)}\Lambda^0 F^{0\prime} e\right) + \frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)} e'e\right). \end{aligned}$$

We now derive the probability limits of each of the three terms.

Term 2 We start with the sole term that converges in probability to 0. We can decompose the second term as

$$\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)}\Lambda^0 F^{0\prime} e\right) = \underbrace{\frac{1}{mT}\operatorname{tr}\left(F^{0\prime} e \Lambda^0\right)}_{I} - \underbrace{\frac{1}{mT}\operatorname{tr}\left(\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\Lambda^0 F^{0\prime} e \Lambda(\kappa)\right)}_{II}$$

We study term II first.

$$\begin{aligned} II &= \frac{1}{mT}\operatorname{tr}\left(\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\Lambda^0 F^{0\prime} e \Lambda(\kappa)\right) \\ &= \operatorname{tr}\left[\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\left(\frac{\Lambda(\kappa)'\Lambda^0}{m}\right)\left(\frac{1}{mT}F^{0\prime} e \Lambda(\kappa)\right)\right]. \end{aligned}$$

Here, $\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}$ and $\frac{\Lambda(\kappa)'\Lambda^0}{m}$ are all $O(1)$ $r \times r$ matrices, so we need only show that

$$\frac{1}{mT}F^{0\prime} e \Lambda(\kappa) = o_p(1).$$

To this end, note that

$$e\Lambda(\kappa) = \sum_{i=1}^{m} \tilde{e}_i \beta(\tau_i;\kappa)' = \begin{pmatrix} \sum_{i=1}^m e_{i1}\beta(\tau_i;\kappa)' \\ \vdots \\ \sum_{i=1}^m e_{iT}\beta(\tau_i;\kappa)' \end{pmatrix},$$

so that

$$\|e\Lambda(\kappa)\|^2 = \sum_{t=1}^{T} \left| \sum_{i=1}^{m} e_{it}\beta(\tau_i;\kappa)' \right|^2 = \sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{j=1}^{m} e_{it}e_{jt}\beta(\tau_i;\kappa)'\beta(\tau_j;\kappa).$$

Taking expectations on both sides implies, by the independence of the measurement errors across maturities,

$$\begin{aligned}\mathbb{E}\|e\Lambda(\kappa)\|^2 &= \sum_{t=1}^{T}\sum_{i=1}^{m} \mathbb{E}\left[e_{it}^2\right] |\beta(\tau_i;\kappa)|^2 \\ &= T\sigma^2 \left(\sum_{i=1}^{m} |\beta(\tau_i;\kappa)|^2 \right) = T\sigma^2 \cdot \|\Lambda(\kappa)\|^2.\end{aligned}$$

Thus,

$$\mathbb{E}\left\| \frac{1}{m\sqrt{T}} e\Lambda(\kappa) \right\|^2 \leq \frac{\sigma^2}{m} \operatorname{tr}\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m} \right),$$

where the right hand side is $O(1/m)$, indicating that

$$\left\| \frac{1}{m\sqrt{T}} e\Lambda(\kappa) \right\| = O_p\left(\frac{1}{\sqrt{m}} \right).$$

Since

$$\left\| \frac{1}{mT} F^{0\prime} e\Lambda(\kappa) \right\| \leq \left\| \frac{\sqrt{T}}{F}^0 \right\| \cdot \left\| \frac{1}{mT} e\Lambda(\kappa) \right\|,$$

where $\left\| \frac{\sqrt{T}}{F}^0 \right\| = O_p(1)$, it follows that

$$\frac{1}{mT} F^{0\prime} e\Lambda(\kappa) = O_p\left(\frac{1}{\sqrt{m}} \right),$$

which implies that it is $o_p(1)$.

Moving onto the first term, we have

$$\begin{aligned} I &= \frac{1}{mT} \operatorname{tr}\left(F^{0\prime} e\Lambda^0 \right) \\ &= \operatorname{tr}\left(\frac{1}{mT} F^{0\prime} e\Lambda(\kappa_0) \right) \end{aligned}$$

We just showed that $\frac{1}{mT} F^{0\prime} e\Lambda(\kappa_0) = o_p(1)$, so it follows that both I and II are

$o_p(1)$. This allows us to conclude that

$$\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)}\Lambda^0 F^{0\prime} e\right) = o_p(1).$$

Term 3 This term pertains to the variances of the error terms e_{it}. We will consider the difference

$$\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)} e'e\right) - \sigma^2.$$

Since

$$\sigma^2 = \frac{1}{m-3}\sigma^2 \operatorname{tr}\left(M_{\Lambda(\kappa)}\right),$$

the difference above can be rewritten as

$$\begin{aligned}
\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)} e'e\right) - \sigma^2 &= \frac{1}{m}\operatorname{tr}\left[M_{\Lambda(\kappa)}\left(\frac{1}{T}e'e - \frac{m}{m-3}\sigma^2 I_m\right)\right] \\
&= \underbrace{\frac{1}{m}\operatorname{tr}\left(\frac{1}{T}e'e - \sigma^2 I_m\right)}_{I} \\
&\quad \underbrace{- \frac{1}{m}\operatorname{tr}\left[\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\left(\frac{1}{T}e'e - \sigma^2 I_m\right)\right]}_{II} \\
&\quad + \underbrace{\frac{1}{m}\operatorname{tr}\left(M_{\Lambda(\kappa)}\right)\frac{3}{m-3}\sigma^2}_{III}.
\end{aligned}$$

Clearly,

$$III = \frac{1}{m}\operatorname{tr}\left(M_{\Lambda(\kappa)}\right)\frac{3}{m-3}\sigma^2 = \frac{3\sigma^2}{m\sqrt{m-3}}$$

goes to 0 as $m, T \to \infty$. On the other hand, we have

$$\begin{aligned}
II &= \frac{1}{m}\operatorname{tr}\left[\Lambda(\kappa)\left(\Lambda(\kappa)'\Lambda(\kappa)\right)^{-1}\Lambda(\kappa)'\left(\frac{1}{T}e'e - \sigma^2 I_m\right)\right] \\
&= \frac{1}{m}\operatorname{tr}\left[\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\left(\frac{1}{\sqrt{m}}\Lambda(\kappa)'\right)\left(\frac{1}{T}e'e - \sigma^2 I_m\right)\left(\frac{1}{\sqrt{m}}\Lambda(\kappa)\right)\right]
\end{aligned}$$

$$\leq \frac{1}{m}\left\|\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\right\| \cdot \left\|\left(\frac{1}{\sqrt{m}}\Lambda(\kappa)'\right)\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\left(\frac{1}{\sqrt{m}}\Lambda(\kappa)\right)\right\|$$

$$\leq \left\|\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\right\| \cdot \left\|\frac{1}{\sqrt{m}}\Lambda(\kappa)\right\| \cdot \left\|\frac{1}{m^{3/2}}\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|,$$

where the first inequality used the Cauchy-Schwarz inequality for the trace inner product. $\left\|\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\right\|$ and $\left\|\frac{1}{\sqrt{m}}\Lambda(\kappa)\right\|$ are $O(1)$, so we need only show that $\left\|\frac{1}{m^{3/2}}\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|$ converges in probability to 0.

Define $\sigma_{ij} = \begin{cases}\sigma^2 & \text{if } i=j \\ 0 & \text{if } i\neq j\end{cases}$. Letting $\{v_1^{(m)},\cdots,v_m^{(m)}\}$ be the standard basis of $\mathbb{R}^m$, we can see that

$$\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right) = \left(\Lambda(\kappa)'\left(\tfrac{1}{T}e'\tilde{e}_1-\sigma^2 v_1^{(m)}\right) \quad \cdots \quad \Lambda(\kappa)'\left(\tfrac{1}{T}e'\tilde{e}_m-\sigma^2 v_m^{(m)}\right)\right),$$

so that

$$\left\|\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|^2 = \sum_{i=1}^{m}\left|\Lambda(\kappa)'\left(\frac{1}{T}e'\tilde{e}_i-\sigma^2 v_i^{(m)}\right)\right|^2.$$

For any $1\leq i\leq m$,

$$\begin{aligned}\Lambda(\kappa)'\left(\frac{1}{T}e'\tilde{e}_i-\sigma^2 v_i^{(m)}\right) &= \frac{1}{T}\left(\beta(\tau_1;\kappa) \quad \cdots \quad \beta(\tau_m;\kappa)\right)\begin{pmatrix}\tilde{e}_1' \\ \vdots \\ \tilde{e}_m'\end{pmatrix}\tilde{e}_i \\ &\quad -\left(\beta(\tau_1;\kappa) \quad \cdots \quad \beta(\tau_m;\kappa)\right)\sigma^2 v_i^{(m)} \\ &= \sum_{j=1}^{m}\beta(\tau_j;\kappa)\left(\frac{1}{T}\tilde{e}_j'\tilde{e}_i\right)-\beta(\tau_i;\kappa)\sigma^2 \\ &= \sum_{j=1}^{m}\beta(\tau_j;\kappa)\left(\frac{1}{T}\tilde{e}_j'\tilde{e}_i-\sigma_{ij}\right).\end{aligned}$$

Defining

$$\zeta_{ij,t} = e_{jt}e_{it}-\sigma_{ij}$$

for any $i,j,t\in N_+$, we have

$$\frac{1}{T}\tilde{e}_j'\tilde{e}_i-\sigma_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left(e_{jt}e_{it}-\sigma_{ij}\right) = \frac{1}{T}\sum_{t=1}^{T}\zeta_{ij,t},$$

so that

$$\mathbb{E}\left\|\frac{1}{m^{3/2}}\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|^2 = \frac{1}{m^3}\sum_{i=1}^{m}\mathbb{E}\left|\sum_{j=1}^{m}\beta(\tau_j;\kappa)\left(\frac{1}{T}\sum_{t=1}^{T}\zeta_{ij,t}\right)\right|^2$$

$$= \frac{1}{m^3}\sum_{j=1}^{m}\sum_{k=1}^{m}\left(\beta(\tau_j;\kappa)'\beta(\tau_k;\kappa)\right)\left(\frac{1}{T^2}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\zeta_{ij,t}\zeta_{ik,s}\right]\right).$$

The independence of measurement errors for different maturities indicates that

$$\mathbb{E}\left[\zeta_{ij,t}\zeta_{ik,s}\right] = 0$$

whenever $j \neq k$. Letting

$$\sup_{i,j\in N_+,\, t,s\in\mathbb{Z}}\mathbb{E}\left[\zeta_{ij,t}\zeta_{ij,s}\right] = \mu_4 < +\infty,$$

where the supremum is bounded due to the stationarity of the errors and the assumption of finite fourth moments, we can see that

$$\mathbb{E}\left\|\frac{1}{m^{3/2}}\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|^2 = \frac{1}{m^3}\sum_{j=1}^{m}\left|\beta(\tau_j;\kappa)\right|^2\left(\frac{1}{T^2}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\zeta_{ij,t}\zeta_{ij,s}\right]\right)$$

$$\leq \frac{\mu_4}{m}\cdot\operatorname{tr}\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right).$$

Since $\operatorname{tr}\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right) = O(1)$, we can see that

$$\left\|\frac{1}{m^{3/2}}\Lambda(\kappa)'\left(\frac{1}{T}e'e-\sigma^2 I_m\right)\right\|^2 = O_p\left(\frac{1}{m}\right).$$

By implication,

$$II = O_p\left(\frac{1}{\sqrt{m}}\right)$$

and thus converges in probability to 0.

Finally, looking at the first term I shows us that

$$I = \frac{1}{m}\operatorname{tr}\left(\frac{1}{T}e'e-\sigma^2 I_m\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{T}\tilde{e}_i'\tilde{e}_i-\sigma^2\right) = \frac{1}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\left(e_{it}^2-\sigma^2\right).$$

We now have

$$\mathbb{E}\Big|\frac{1}{m}\operatorname{tr}\Big(\frac{1}{T}e'e-\sigma^2 I_m\Big)\Big|^2=\frac{1}{m^2T^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\left(e_{it}^2-\sigma^2\right)\left(e_{js}^2-\sigma^2\right)\right]$$

$$=\frac{1}{m^2T^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[\left(e_{it}^2-\sigma^2\right)\left(e_{is}^2-\sigma^2\right)\right]\leq\frac{\mu_4}{m},$$

where the second inequality follows from the independence of measurement errors for different maturities. By implication, $I=O_p\left(\frac{1}{\sqrt{m}}\right)$ as well, so that

$$\frac{1}{mT}\operatorname{tr}\left(M_{\Lambda(\kappa)}e'e\right)-\sigma^2=O_p\left(\frac{1}{\sqrt{m}}\right).$$

Term 1 This is the easiest term to deal with. By assumption,

$$\operatorname{tr}\left[\left(\frac{\Lambda^{0\prime}\Lambda^0}{m}\right)\left(\frac{F^{0\prime}F^0}{T}\right)\right]-\operatorname{tr}\left[\left(\frac{\Lambda(\kappa)'\Lambda(\kappa)}{m}\right)^{-1}\left(\frac{\Lambda(\kappa)'\Lambda^0}{m}\right)\left(\frac{F^{0\prime}F^0}{T}\right)\left(\frac{\Lambda(\kappa)'\Lambda^0}{m}\right)'\right]$$

$$\xrightarrow{p}\operatorname{tr}\left(\Omega_0\Omega_F\right)-\operatorname{tr}\left(\Omega(\kappa)^{-1}\Omega_0(\kappa)\Omega_F\Omega_0(\kappa)'\right)=\operatorname{tr}\left[\left(\Omega_0-\Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\cdot\Omega_F\right]$$

as $m,T\to\infty$.

Now define the function $V_0:\Theta\to\mathbb{R}$ as

$$V_0(\kappa)=\operatorname{tr}\left[\left(\Omega_0-\Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\cdot\Omega_F\right]+\sigma^2$$

for any $\kappa\in\Theta$. We saw above that

$$\left|V_{m,T}(\kappa)-V_0(\kappa)\right|=o_p(1),$$

and since the convergence of the terms involving κ are uniform with respect to κ on Θ,

$$\sup_{\kappa\in\Theta}\left|V_{m,T}(\kappa)-V_0(\kappa)\right|=o_p(1).$$

Finally, we need to show that V_0 is uniquely minimized at κ_0. To this end, note that

$$\Omega_0-\Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)$$

is positive semidefinite, so that

$$\operatorname{tr}\left[\left(\Omega_0 - \Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\cdot\Omega_F\right] = \operatorname{tr}\left[\Omega_F^{\frac{1}{2}\prime}\left(\Omega_0 - \Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\Omega_F^{\frac{1}{2}}\right] \geq 0$$

for any κ. The trace of a positive semidefinite matrix is equal to the sum of its eigenvalues, which are non-negative, so equality holds in this case if and only if

$$\Omega_F^{\frac{1}{2}\prime}\left(\Omega_0 - \Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\Omega_F^{\frac{1}{2}} = O_{3\times 3},$$

or equivalently,

$$\Omega_0 - \Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa) = O_{3\times 3}.$$

This is only the case when $\kappa = \kappa_0$, so it follows that

$$V_0(\kappa_0) = \sigma^2 < \sigma^2 + \operatorname{tr}\left[\left(\Omega_0 - \Omega_0(\kappa)'\Omega(\kappa)^{-1}\Omega_0(\kappa)\right)\cdot\Omega_F\right] = V_0(\kappa)$$

for any $\kappa \neq \kappa_0$. As such, V_0 is uniquely minimized at κ_0, and $\overline{\kappa}$ is consistent for κ_0.

Using the consistency of $\overline{\kappa}$, we can prove the consistency of $\overline{F}$ for the true factors F^0. The time t factor estimator is given by

$$\begin{aligned}\overline{f}_t &= \left(\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right)^{-1}\Lambda(\overline{\kappa})'\mathcal{Y}_t \\ &= \left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1}\left(\frac{\Lambda(\overline{\kappa})'\Lambda^0}{m}\right)f_t^0 + \left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1}\left(\frac{1}{m}\Lambda(\overline{\kappa})'e_t\right).\end{aligned}$$

Starting with the error term,

$$\frac{1}{m}\Lambda(\overline{\kappa})'e_t = \frac{1}{m}\sum_{i=1}^{m}\beta(\tau_i;\overline{\kappa})e_{it},$$

so that

$$\begin{aligned}\mathbb{E}\left|\frac{1}{m}\Lambda(\overline{\kappa})'e_t\right|^2 &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\beta(\tau_i;\overline{\kappa})'\beta(\tau_j;\overline{\kappa})\mathbb{E}\left[e_{it}e_{jt}\right] \\ &= \sigma^2\frac{1}{m^2}\sum_{i=1}^{m}\left|\beta(\tau_i;\overline{\kappa})\right|^2 \\ &= \sigma^2\frac{1}{m}\operatorname{tr}\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right).\end{aligned}$$

Since the last term on the right is $O_p(1)$, this shows us that

$$\frac{1}{m}\Lambda(\overline{\kappa})' e_t = o_p(1).$$

Now we study the asymptotic behavior of $\Lambda(\overline{\kappa})$. We work with the following norm:

$$\left\| \frac{1}{\sqrt{m}} \left(\Lambda(\overline{\kappa}) - \Lambda^0 \right) \right\|^2 = \frac{1}{m} \sum_{i=1}^{m} |\beta_2(\tau_i; \overline{\kappa}) - \beta_2(\tau_i; \kappa_0)|^2 + \frac{1}{m} \sum_{i=1}^{m} |\beta_3(\tau_i; \overline{\kappa}) - \beta_3(\tau_i; \kappa_0)|^2.$$

By the mean value theorem,

$$\beta_2(\tau_i; \overline{\kappa}) - \beta_2(\tau_i; \kappa_0) = \frac{1}{k^2 \tau_i} \left((\tau_i k + 1) \exp(-\tau_i k) - 1 \right) (\overline{\kappa} - \kappa_0)$$

for some k between $\overline{\kappa}$ and κ_0. Since

$$\begin{aligned} \left| \frac{1}{k^2 \tau_i} \left((\tau_i k + 1) \exp(-\tau_i k) - 1 \right) \right| &= \frac{1}{k^2 \tau_i} \left(1 - (\tau_i k + 1) \exp(-\tau_i k) \right) \\ &\leq \frac{1}{\epsilon} \left(\frac{1 - \exp(-\tau_i \epsilon)}{\tau_i \epsilon} \right) = \frac{1}{\epsilon} \beta_2(\tau_i; \epsilon), \end{aligned}$$

we have

$$|\beta_2(\tau_i; \overline{\kappa}) - \beta_2(\tau_i; \kappa_0)|^2 \leq \frac{1}{\epsilon^2} \beta_2(\tau_i; \epsilon)^2 |\overline{\kappa} - \kappa_0|^2.$$

By implication,

$$\frac{1}{m} \sum_{i=1}^{m} |\beta_2(\tau_i; \overline{\kappa}) - \beta_2(\tau_i; \kappa_0)|^2 = \frac{1}{\epsilon^2} |\overline{\kappa} - \kappa_0|^2 \left(\frac{1}{m} \sum_{i=1}^{m} \beta_2(\tau_i; \epsilon)^2 \right),$$

where the average in the rightmost term converges to the (2,2) element of $\Omega(\epsilon)$ be assumption. The consistency of $\overline{\kappa}$ for κ_0 now reveals that

$$\frac{1}{m} \sum_{i=1}^{m} |\beta_2(\tau_i; \overline{\kappa}) - \beta_2(\tau_i; \kappa_0)|^2 = o_p(1).$$

In a similar manner, the mean value theorem reveals that

$$|\exp(-\tau_i \overline{\kappa}) - \exp(-\tau_i \kappa_0)|^2 \leq \tau_{max}^2 \exp(-\tau_i \epsilon)^2 |\overline{\kappa} - \kappa_0|^2,$$

where we can numerically check that the sequence

$$\left\{ \frac{1}{m} \sum_{i=1}^{m} \exp(-\tau_i \epsilon)^2 \right\}_{m \in N_+}$$

converges. Therefore,

$$\frac{1}{m}\sum_{i=1}^{m}|\beta_3(\tau_i;\overline{\kappa})-\beta_3(\tau_i;\kappa_0)|^2 \leq \frac{2}{m}\sum_{i=1}^{m}|\beta_2(\tau_i;\overline{\kappa})-\beta_2(\tau_i;\kappa_0)|^2+\frac{2}{m}\sum_{i=1}^{m}|\exp(-\tau_i\overline{\kappa})-\exp(-\tau_i\kappa_0)|^2$$
$$\leq \frac{2}{m}\sum_{i=1}^{m}|\beta_2(\tau_i;\overline{\kappa})-\beta_2(\tau_i;\kappa_0)|^2+2\tau_{max}^2\left(\frac{1}{m}\sum_{i=1}^{m}\exp(-\tau_i\epsilon)^2\right)|\overline{\kappa}-\kappa_0|^2,$$

where both terms on the right are $o_p(1)$. We have shown that

$$\left\|\frac{1}{\sqrt{m}}\left(\Lambda(\overline{\kappa})-\Lambda^0\right)\right\|^2=o_p(1).$$

By implication,

$$\left\|\frac{\Lambda(\overline{\kappa})'\Lambda^0}{m}-\frac{\Lambda^{0\prime}\Lambda^0}{m}\right\| \leq \left\|\frac{1}{\sqrt{m}}\left(\Lambda(\overline{\kappa})-\Lambda^0\right)\right\|\cdot\left\|\frac{1}{\sqrt{m}}\Lambda^0\right\|=o_p(1),$$

so that

$$\frac{\Lambda(\overline{\kappa})'\Lambda^0}{m}\xrightarrow{p}\Omega_0.$$

Similarly,

$$\left\|\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}-\frac{\Lambda(\overline{\kappa})'\Lambda^0}{m}\right\| \leq \left\|\frac{1}{\sqrt{m}}\Lambda(\overline{\kappa})\right\|\cdot\left\|\frac{1}{\sqrt{m}}\left(\Lambda(\overline{\kappa})-\Lambda^0\right)\right\|=o_p(1),$$

implying that

$$\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\xrightarrow{p}\Omega_0.$$

Putting these results together, it now follows easily that

$$\overline{f}_t\xrightarrow{p}f_t^0.$$

To see consistency for the mean of the entire factors, note that

$$\frac{1}{\sqrt{T}}\overline{F}=\mathcal{Y}\Lambda(\overline{\kappa})\left(\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})\right)^{-1}$$
$$=\frac{1}{\sqrt{T}}F^0\left(\frac{\Lambda^{0\prime}\Lambda(\overline{\kappa})}{m}\right)\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1}+\left(\frac{1}{m\sqrt{T}}e\Lambda(\overline{\kappa})\right)\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1}.$$

Since

$$\Lambda(\overline{\kappa})'e' = \begin{pmatrix} \Lambda(\overline{\kappa})'e_1 & \cdots & \Lambda(\overline{\kappa})'e_T \end{pmatrix},$$

we have

$$\|e\Lambda(\overline{\kappa})\|^2 = \sum_{t=1}^{T} \left|\Lambda(\overline{\kappa})'e_t\right|^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{m\sqrt{T}}e\Lambda(\overline{\kappa})\right\|^2 &= \frac{1}{mT}\sum_{t=1}^{T}\mathbb{E}\left|\Lambda(\overline{\kappa})'e_t\right|^2 \\
&= \frac{1}{m^2T}\sum_{t=1}^{T}\sum_{i=1}^{m}\sum_{j=1}^{m}\beta(\tau_i;\overline{\kappa})'\beta(\tau_j;\overline{\kappa})\mathbb{E}\left[e_{it}e_{jt}\right] \\
&= \sigma^2\frac{1}{m^2}\sum_{i=1}^{m}\left|\beta(\tau_i;\overline{\kappa})\right|^2 \\
&= \sigma^2\frac{1}{mT}\operatorname{tr}\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right).
\end{aligned}$$

It follows that

$$\frac{1}{m\sqrt{T}}e\Lambda(\overline{\kappa}) = o_p(1).$$

Therefore,

$$\begin{aligned}
\frac{1}{\sqrt{T}}\left(\overline{F} - F^0\right) = \frac{1}{\sqrt{T}}F^0\left[\left(\frac{\Lambda^{0\prime}\Lambda(\overline{\kappa})}{m}\right)\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1} - I_r\right] \\
+ \left(\frac{1}{m\sqrt{T}}e\Lambda(\overline{\kappa})\right)\left(\frac{\Lambda(\overline{\kappa})'\Lambda(\overline{\kappa})}{m}\right)^{-1},
\end{aligned}$$

where the terms on the right hand side are all $o_p(1)$. It follows that

$$\frac{1}{\sqrt{T}}\left(\overline{F} - F^0\right) = o_p(1),$$

and by implication,

$$\frac{1}{T}\left\|\overline{F} - F^0\right\| = o_p(1).$$

This tells us that

$$\frac{1}{T}\overline{F}'\iota_T \xrightarrow{p} \mu_F,$$

$$\frac{1}{T}\overline{F}'\overline{F}, \quad \frac{1}{T}\overline{F}'F^0 \xrightarrow{p} \Omega_F,$$

among other things.

B Kalman Smoother for the Singular Case

The exposition in this appendix follows Durbin and Koopman (2012) almost verbatim, only filling in the necessary details. The smoothed factors and factor variances derived below can also be found in Barigozzi and Luciani (2019).

In this section, we use innovations to derive an equivalent formulation of the Kalman smoother that holds even when $q < r$ and $P_{t+1|t}(\theta)$ is singular. For the sake of notational brevity, we omit the dependence of the quantities below on the model parameters θ. As in the main text, we assume Gaussianity and independent idiosyncratic errors/factor innovations.

We must first define what is meant by "innovations". The time t innovation v_t is defined as

$$v_t = x_t - x_{t|t-1} = \Lambda(f_t - f_{t|t-1}) + \Sigma e_t.$$

In other words, v_t is the part of x_t that is not predicted at time $t-1$. In light of the interpretation of the conditional expectation $x_{t|t-1} = \mathbb{E}\left[x_t \mid \mathcal{F}_{t-1}\right]$ as the orthogonal projection of x_t onto the closed linear subspace $\mathcal{F}_{t-1}$, v_t represents the orthogonal projection of x_t onto the orthogonal complement of $\mathcal{F}_{t-1}$.

A related concept is the state estimation error

$$\varepsilon_t = f_t - f_{t|t-1} = G(f_{t-1} - f_{t-1|t-1}) + Hu_t,$$

which is the part of the factors f_t that is not predicted at time $t-1$. The innovations and state estimation errors have the following relationship:

$$v_t = \Lambda\varepsilon_t + \Sigma e_t.$$

Note that $P_{t|t-1}$ is the variance of state estimation error, or the mean squared state estimation error:

$$P_{t|t-1} = \mathbb{E}\left[(f_t - f_{t|t-1})(f_t - f_{t|t-1})' \mid \mathcal{F}_{t-1}\right] = \operatorname{Var}\left(\varepsilon_t \mid \mathcal{F}_{t-1}\right).$$

We can also obtain a recursive relationship for the state estimation error:

$$\begin{aligned}
\varepsilon_{t+1} &= f_{t+1} - f_{t+1|t} \\
&= G\left(f_t - f_{t|t}\right) + Hu_{t+1} \\
&= G\left(f_t - f_{t|t-1}\right) - GK_{t|t-1}v_t + Hu_{t+1} \\
&= G\left(I_r - K_{t|t-1}\Lambda\right)\varepsilon_t + \left(Hu_{t+1} - GK_{t|t-1}\Sigma e_t\right).
\end{aligned}$$

Therefore, the DFM can be written as a state space model of the innovations and state estimation errors as follows:

$$\begin{aligned} v_t &= \Lambda\varepsilon_t + \Sigma e_t \\ \varepsilon_{t+1} &= L_t\varepsilon_t + \left(Hu_{t+1} - GK_{t|t-1}\Sigma e_t\right), \end{aligned}$$

where we define

$$L_t = G\left(I_r - K_{t|t-1}\Lambda\right).$$

These relationships will prove invaluable later on.

We first show that the innovations $\{v_1, \cdots, v_T\}$ are mutually independent and follow a jointly Gaussian distribution. Note initially that the random vector $(v_1', \cdots, v_T')'$ is a non-random linear transformation of the data $(x_1', \cdots, x_T')'$, so that the joint density of $v_1, \cdots, v_T$ is exactly that of $x_1, \cdots, x_T$. Furthermore, for any $1 \leq t \leq T$, the likelihood of x_t given $\mathcal{F}_{t-1}$ can be written as

$$f(x_t \mid \mathcal{F}_{t-1}) = \left(\frac{1}{2\pi}\right)^N \left|V_{t|t-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} v_t' V_{t|t-1}^{-1} v_t\right).$$

The recursive formulas for $V_{t|t-1}$ indicates that it depends only on the (non-random) initial value $P_{0|0}$, so that the distribution on the right hand side of

$$v_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}\left[O_{N\times 1}, V_{t|t-1}\right]$$

does not depend on $\mathcal{F}_{t-1}$. It follows that v_t is independent of $\mathcal{F}_{t-1}$ with unconditional distribution $\mathcal{N}\left[O_{N\times 1}, V_{t|t-1}\right]$, so that $f(x_t \mid \mathcal{F}_{t-1})$ is simply the unconditional density of v_t, denoted $f(v_t)$. Putting these results together, we have

$$f(v_1, \cdots, v_T) = f(x_1, \cdots, x_T) = \prod_{t=1}^{T} f(x_t \mid \mathcal{F}_{t-1}) = \prod_{t=1}^{T} f(v_t).$$

The joint density of $v_1, \cdots, v_T$ can be expressed as the product of their marginal densities, so they are mutually independent. Furthermore, for any $1 \leq t < T$, $v_t, \cdots, v_T$ are independent of $\mathcal{F}_{t-1}$, so that

$$\begin{pmatrix} v_t \\ \vdots \\ v_T \end{pmatrix} \mid \mathcal{F}_{t-1} \sim \mathcal{N}\left[O_{N(T-t)\times 1}, \quad \operatorname{diag}\left(V_{t|t-1}, \cdots, V_{T|T-1}\right)\right].$$

Now we are ready to derive the Kalman smoother. Note that, given $\mathcal{F}_{t-1}$ and v_t, we can construct x_t as $x_t = v_t + x_{t|t-1}$, where $x_{t|t-1} \in \mathcal{F}_{t-1}$. Conversely, v_t and $\mathcal{F}_{t-1}$ are known quantities when $\mathcal{F}_t$ is known. In other words,

$$\mathcal{F}_t = \sigma\{\mathcal{F}_{t-1}, v_t\}.$$

By implication,

$$\mathcal{F}_T = \sigma\{\mathcal{F}_{t-1}, v_t, \cdots, v_T\},$$

so that the information contained in the data can be partitioned into $\mathcal{F}_{t-1}$, the information up to time $t-1$, and $\{v_t, \cdots, v_T\}$, the information on the innovations from time t onward. Note that

$$\begin{pmatrix} f_t \\ v_t \\ \vdots \\ v_T \end{pmatrix}$$

can be represented as a linear transformation of f_t, v_t and the errors $e_{t+1}, \cdots, e_T, u_{t+1}, \cdots, u_T$, which are jointly normal given $\mathcal{F}_{t-1}$. For instance,

$$\begin{aligned} v_{t+1} = x_{t+1} - x_{t+1|t} &= \Lambda f_{t+1} + \Sigma e_{t+1} - \Lambda(c + G f_{t|t}) \\ &= \Lambda G f_t + \Lambda H u_{t+1} + \Sigma e_{t+1} - \Lambda G f_{t|t-1} + G K_{t|t-1} v_t. \end{aligned}$$

Therefore, the random vector above is a mean zero jointly normally distributed random vector conditional on $\mathcal{F}_{t-1}$, and we can use the updating formula for jointly normally distributed random variables to see that

$$\begin{aligned} f_{t|T} &= \mathbb{E}[f_t \mid \mathcal{F}_T] \\ &= \mathbb{E}[f_t \mid \mathcal{F}_{t-1}, v_t, \cdots, v_T] \\ &= f_{t|t-1} + \sum_{h=0}^{T-t} \mathbb{E}\left[f_t v_{t+h}' \mid \mathcal{F}_{t-1}\right] V_{t+h|t+h-1}^{-1} v_{t+h}, \end{aligned}$$

where we used the fact that $v_t, \cdots, v_T$ are conditionally independent given $\mathcal{F}_{t-1}$ to justify the last equality.

For any $0 \leq h \leq T-t$,

$$\begin{aligned} \mathbb{E}\left[f_t v_{t+h}' \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\varepsilon_t v_{t+h}' \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\varepsilon_t (\Lambda \varepsilon_{t+h} + \Sigma e_{t+h})' \mid \mathcal{F}_{t-1}\right] \end{aligned}$$

$$= \mathbb{E}\left[\varepsilon_t \varepsilon_{t+h}' \mid \mathcal{F}_{t-1}\right] \Lambda' = P_{t|t-1} \left(\prod_{i=0}^{h-1} L_{t+h-i}' \right) \Lambda',$$

where we used the fact that ε_t is independent of $e_t, e_{t+1}, \cdots, e_T, u_{t+1}, \cdots, u_T$ and the recursive formula for the state estimation error. It follows that

$$f_{t|T} = f_{t|t-1} + P_{t|t-1} \underbrace{\left[\sum_{h=0}^{T-t} \left(\prod_{i=0}^{h-1} L_{t+h-i}' \right) \Lambda' V_{t+h|t+h-1}^{-1} v_{t+h} \right]}_{r_{t-1}}.$$

This must hold for any $1 \leq t \leq T$, so

$$\begin{aligned} r_{t-1} &= \sum_{h=0}^{T-t} \left(\prod_{i=0}^{h-1} L_{t+h-i}' \right) \Lambda' V_{t+h|t+h-1}^{-1} v_{t+h} \\ &= \Lambda' V_{t|t-1}^{-1} v_t + L_t' \sum_{h=0}^{T-t-1} \left(\prod_{i=0}^{h-1} L_{t+1+h-i}' \right) \Lambda' V_{t+1+h|t+1+h-1}^{-1} v_{t+1+h} \\ &= \Lambda' V_{t|t-1}^{-1} v_t + L_t' r_t. \end{aligned}$$

As for the smoothed factor variance, by the same updating formula as above we have

$$\begin{aligned} P_{t|T} &= \mathrm{Var}\left(f_t \mid \mathcal{F}_T\right) \\ &= \mathrm{Var}\left(f_t \mid \mathcal{F}_{t-1}, v_t, \cdots, v_T\right) \\ &= P_{t|t-1} - \sum_{h=0}^{T-t} \mathbb{E}\left[f_t v_{t+h}' \mid \mathcal{F}_{t-1}\right] V_{t+h|t+h-1}^{-1} \mathbb{E}\left[v_{t+h} f_t' \mid \mathcal{F}_{t-1}\right] \\ &= P_{t|t-1} - P_{t|t-1} \underbrace{\left[\sum_{h=0}^{T-t} \left(\prod_{i=0}^{h-1} L_{t+h-i}' \right) \Lambda' V_{t+h|t+h-1}^{-1} \Lambda \left(\prod_{i=0}^{h-1} L_{t+h-i} \right) \right]}_{N_{t-1}} P_{t|t-1}. \end{aligned}$$

Again, we can see that

$$\begin{aligned} N_{t-1} &= \sum_{h=0}^{T-t} \left(\prod_{i=0}^{h-1} L_{t+h-i}' \right) \Lambda' V_{t+h|t+h-1}^{-1} \Lambda \left(\prod_{i=0}^{h-1} L_{t+h-i} \right) \\ &= \Lambda' V_{t|t-1}^{-1} \Lambda + L_t' N_t L_t. \end{aligned}$$

In summary, the smoothed factors and factor variances are given recursively by the

formulas

$$
\begin{aligned}
f_{t|T} &= f_{t|t-1} + P_{t|t-1} r_{t-1} \\
P_{t|T} &= \left(I_r - P_{t|t-1} N_{t-1}\right) P_{t|t-1} \\
r_{t-1} &= \Lambda' V_{t|t-1}^{-1} v_t + L_t' r_t \\
N_{t-1} &= \Lambda' V_{t|t-1}^{-1} \Lambda + L_t' N_t L_t \\
r_T &= O_{r \times 1} \\
N_T &= O_{r \times r} \\
L_t &= G\left(I_r - K_{t|t-1}\Lambda\right)
\end{aligned}
$$

for any $1 \leq t \leq T$.

C Why the EM Algorithm Works

We prove here that, under certain regularity conditions, the algorithm above yields a sequence of estimates $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ that converges to some stationary point $\hat{\theta}$ of the log-likelihood function $L(\cdot) := L(\cdot \mid \mathcal{Y}_T)$. First we introduce some notations.

Throughout, we assume the outcome $\omega \in \Omega$ is fixed, so that the likelihoods below can be treated as non-random. Let Θ be the parameter space, and define the correspondence $\mathcal{M} : \Theta \to \Theta$ as

$$\mathcal{M}(\theta) = \underset{\phi\in\Theta}{\operatorname{argmax}} \quad E(\phi \mid \theta)$$

for any $\theta \in \Theta$. Likewise, for any $\theta \in \Theta$ and $1 \leq t \leq T$, define the function $H(\cdot \mid \theta)$ on Θ is defined as

$$H(\phi \mid \theta) = \int \log(l(F \mid Y, \phi)) \cdot l(F \mid Y, \theta) dF$$

for any $\phi \in \Theta$.

Finally, assume that $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ is a sequence of EM iterates given the outcome ω. Our objective is to show that this sequence converges to a stationary point of the log likelihood under certain regularity assumptions.

Assumptions

We make the following assumptions:

1) The parameter space Θ is a compact subset of $\mathbb{R}^d$.

2) The log-likelihood $L(\cdot)$ is continuous on Θ and differentiable on the interior of Θ.

3) For any $\theta \in \Theta$, the mapping $E(\cdot \mid \theta)$ is differentiable on the interior of Θ.

4) For any $1 \leq t \leq T$ and $\theta \in \Theta$, $H(\cdot \mid \theta)$ is differentiable on the interior of Θ.

5) Integration and differentiation can be interchanged.

6) $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ is contained in the interior Θ^o of Θ.

7) Defining $Q : \Theta^o \times \Theta \to \mathbb{R}$ as

$$Q(\theta_0, \theta) = \frac{\partial E(\phi \mid \theta)}{\partial \phi}|_{\phi=\theta_0}$$

for any $\theta_0 \in \Theta^o$ and $\theta \in \Theta$, Q is continuous on $\Theta^o \times \Theta$.

We briefly mention some implications of the above assumptions. For one, note that, for any $\theta \in \Theta^o$, the interior of Θ,

$$\begin{aligned}\frac{\partial H(\phi \mid \theta)}{\partial \phi}|_{\phi=\theta} &= \int \frac{\partial \log(l(F \mid Y,\phi))}{\partial \phi}|_{\phi=\theta} \cdot l(F \mid Y,\theta) dF \\ &= \int \frac{\partial l(F \mid Y,\phi)}{\partial \phi}|_{\phi=\theta}\, dF \\ &= \left[\frac{\partial}{\partial \phi}\int l(F \mid Y,\phi) dF\right]|_{\phi=\theta} = 0,\end{aligned}$$

where the last equality follows because $l(F \mid Y,\phi)$ integrates to 1 for any $\phi \in \Theta^o$.

In addition, the Kullback-Leibler inequality holds: for any $\phi,\theta \in \Theta$,

$$\begin{aligned}H(\phi \mid \theta) - H(\theta \mid \theta) &= \mathbb{E}\left[\log(l(F \mid Y,\phi)) \mid Y,\theta\right] - \mathbb{E}\left[\log(l(F \mid Y,\theta)) \mid Y,\theta\right] \\ &= \mathbb{E}\left[\log\left(\frac{l(F \mid Y,\phi)}{l(F \mid Y,\theta)}\right)|Y,\theta\right] \\ &\leq \log\left(\mathbb{E}\left[\frac{l(F \mid Y,\phi)}{l(F \mid Y,\theta)}|Y,\theta\right]\right) \\ &= \log\left(\int l(F \mid Y,\phi)F\right) = \log(1) = 0,\end{aligned}$$

where the inequality is justified by Jensen's inequality.

Furthermore, for any $i \in N_+$, since $\theta^{(i+1)} \in \mathcal{M}(\theta^{(i)})$, $\theta^{(i+1)} \in \Theta^o$, and $E(\cdot \mid \theta^{(i)})$ is differentiable on Θ^o, it follows from the first order necessary condition of maximization that

$$Q(\theta^{(i+1)},\theta^{(i)}) := \frac{\partial E(\theta \mid \theta^{(i)})}{\partial \theta}|_{\theta=\theta^{(i+1)}} = \mathbf{0}.$$

Monotonicity of the Log-Likelihood

First, we show that every step of the algorithm updates the parameter estimates in a manner that increases the value of the log likelihood. Since

$$l(F \mid Y,\theta) = \frac{l(Y \mid F,\theta) \cdot l(F \mid \theta)}{l(Y \mid \theta)} = \frac{l(Y,F \mid \theta)}{l(Y \mid \theta)}$$

by Bayes' rule, we have

$$\begin{aligned}L(\theta) &= \log(l(Y \mid \theta)) \\ &= \log(l(Y,F \mid \theta)) - \log(l(F \mid Y,\theta)).\end{aligned}$$

Since $l(F \mid Y, \theta^{(i-1)})$ is a density, it integrates to 1 and thus

$$\begin{aligned} L(\theta) &= \int L(\theta) l(F \mid Y, \theta^{(i-1)}) dF \\ &= \int \log(l(Y, F \mid \theta)) l(F \mid Y, \theta^{(i-1)}) dF - \int \log(l(F \mid Y, \theta)) l(F \mid Y, \theta^{(i-1)}) dF \\ &= E(\theta \mid \theta^{(i-1)}) - H(\theta \mid \theta^{(i-1)}). \end{aligned}$$

By the Kullback-Leibler inequality,

$$H(\theta \mid \theta^{(i-1)}) \leq H_t(\theta^{(i-1)} \mid \theta^{(i-1)})$$

for any $\theta \in \Theta$.

It follows that

$$\begin{aligned} L(\theta) - L(\theta^{(i-1)}) &= E(\theta \mid \theta^{(i-1)}) - E(\theta^{(i-1)} \mid \theta^{(i-1)}) \\ &\quad + H(\theta^{(i-1)} \mid \theta^{(i-1)}) - H(\theta \mid \theta^{(i-1)}) \\ &\geq E(\theta \mid \theta^{(i-1)}) - E(\theta^{(i-1)} \mid \theta^{(i-1)}) \end{aligned}$$

for any θ in the parameter space. Since $\theta^{(i)}$ is chosen as the maximizer of $E(\theta \mid \theta^{(i-1)})$, we have

$$E(\theta^{(i)} \mid \theta^{(i-1)}) - E(\theta^{(i-1)} \mid \theta^{(i-1)}) \geq 0,$$

and therefore

$$L(\theta^{(i)}) - L(\theta^{(i-1)}) \geq 0.$$

Convergence of the Algorithm

The preceding result tells us that $\{L(\theta^{(i)})\}_{i \in \mathbb{N}}$ is a monotonically increasing sequence, and therefore it has a limit $L^* \in [-\infty, +\infty]$. We can further see that L^* must be real-valued, since asumption (2), along with the compactness of Θ, implies that the image $L(\Theta)$ is compact. Under an additional assumption, we can easily show that $\{\theta^{(i)}\}_{i \in \mathbb{N}}$ converges to some $\hat{\theta}$:

The inverse image $L^{-1}(L^*)$ is a singleton.

As this assumption states, suppose that $\theta^* \in \Theta$ is the only point in Θ such that $L(\theta^*) = L^*$, and choose any convergent subsequence $\{\theta^{(i_k)}\}_{k \in N_+}$ of $\{\theta^{(i)}\}_{i \in \mathbb{N}}$ with limit θ_0. Then, by

the continuity of L on Θ,

$$L(\theta_0) = \lim_{k\to\infty} L(\theta^{(i_k)}) = L^*,$$

which implies that $\theta_0 = \theta^*$. Therefore, any convergent subsequence of $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ converges to θ^*. Finally, since $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ is a sequence in the compact set Θ, every subsequence of $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ has a convergent subsequence (by sequential compactness). These two results together reveal that $\{\theta^{(i)}\}_{i\in\mathbb{N}}$ converges to θ^*. We denote $\hat{\theta} = \theta^*$.

Now we show that $\hat{\theta}$ is a stationary point of L. To this end, note that the equality

$$L(\theta) = E(\theta \mid \hat{\theta}) - H(\theta \mid \hat{\theta})$$

implies that

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} &= \frac{\partial E(\theta \mid \hat{\theta})}{\partial \theta}\Big|_{\theta=\hat{\theta}} - \frac{\partial H(\theta \mid \hat{\theta})}{\partial \theta}\Big|_{\theta=\hat{\theta}} \\ &= \frac{\partial E(\theta \mid \hat{\theta})}{\partial \theta}\Big|_{\theta=\hat{\theta}} = Q(\hat{\theta}, \hat{\theta}).\end{aligned}$$

By the assumed continuity of Q in both arguments, we can see that

$$\frac{\partial L(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = \lim_{i\to\infty} Q(\theta^{(i+1)}, \theta^{(i)}) = \mathbf{0}.$$

Therefore, $\hat{\theta}$ is a stationary point of L.

D Consistency of Two-Step Estimation Method

We work with the basic model of appendix A, with an added VAR(1) specification for the factors. To prove the consistency result of interest, we make the following assumptions in addition to assumptions A1 to A3 of appendix A:

A4. Factor Innovations as an MDS

We assume that the factor innovation process $\{u_t\}_{t\in\mathbb{Z}}$ is an r-dimensional martingale difference sequence (MDS) with respect to the filtration that it generates. We also assume that $\mathbb{E}[u_t u_t'] = I_r$ for any $t \in \mathbb{Z}$, so that $\{u_t\}_{t\in\mathbb{Z}}$ is L^2-bounded. By implication, the martingale WLLN tells us that

$$\frac{1}{T}\sum_{t=1}^{T} u_t \xrightarrow{p} O_{r\times 1}.$$

Finally, we assume variance ergodicity for $\{u_t\}_{t\in\mathbb{Z}}$, that is,

$$\frac{1}{T}\sum_{t=1}^{T} u_t u_t' \xrightarrow{p} I_r.$$

A5. Stationarity of VAR(1) Representation

We assume that the eigenvalues of G are all within the unit circle. This, in addition to the L^2-boundedness of $\{f_t\}_{t\in\mathbb{Z}}$ implied by assumption A1, implies that $\{f_t\}_{t\in\mathbb{Z}}$ is an absolutely summable causal linear process with innovation process $\{u_t\}_{t\in\mathbb{Z}}$.

By implication, letting $\mathcal{G} = \{\mathcal{G}_t \mid t \in \mathbb{Z}\}$ be the filtration generated by $\{u_t\}_{t\in\mathbb{Z}}$, f_{t-1} is $\mathcal{G}_{t-1}$-measurable and each entry of the matrix process $\{f_{t-1}u_t'\}_{t\in\mathbb{Z}}$ is an L^2-bounded MDS with respect to $\mathcal{G}$. It follows that the following martingale difference WLLN holds:

$$\frac{1}{T}\sum_{t=1}^{T} f_{t-1}u_t' \xrightarrow{p} O_{r\times r}.$$

First, denote

$$\overline{F}_{+1} = \begin{pmatrix} \overline{f}_2' \\ \vdots \\ \overline{f}_T' \end{pmatrix}, \quad \overline{F}_{-1} = \begin{pmatrix} \overline{f}_1' \\ \vdots \\ \overline{f}_{T-1}' \end{pmatrix},$$

and analogously define F^0_{+1} and F^0_{-1}. The results on the least square estimators of the

N-S model tell us that

$$\frac{1}{\sqrt{T}}\left\|\overline{F}_{+1}-F_{+1}^{0}\right\|,\quad \frac{1}{\sqrt{T}}\left\|\overline{F}_{-1}-F_{-1}^{0}\right\|=o_p(1),$$

so that

$$\frac{1}{T}\overline{F}_{-1}'\iota_{T-1}\xrightarrow{p}\mu_F$$

$$\frac{1}{T}\overline{F}_{-1}'\overline{F}_{-1},\quad \frac{1}{T}\overline{F}_{-1}'F_{-1}^{0}\xrightarrow{p}\Omega_F$$

$$\left\|\frac{1}{T}\overline{F}_{-1}'\overline{F}_{+1}-\frac{1}{T}F_{-1}^{0\prime}F_{+1}^{0}\right\|=o_p(1).$$

Define

$$\Pi=\begin{pmatrix}c'\\G'\end{pmatrix},\quad \Gamma^f=HH',$$

so that the true values of c,G and HH' are contained in the parameters Π^0 and Γ^{f0}. Recall that the two-step estimator of Π and Γ^f are denoted

$$\overline{\Pi}=\begin{pmatrix}\overline{c}'\\\overline{G}'\end{pmatrix}=\begin{pmatrix}1 & \frac{1}{T-1}\iota_{T-1}'\overline{F}_{-1}\\ \frac{1}{T-1}\overline{F}_{-1}'\iota_{T-1} & \frac{1}{T-1}\overline{F}_{-1}'\overline{F}_{-1}\end{pmatrix}^{-1}\begin{pmatrix}\frac{1}{T-1}\iota_{T-1}'\overline{F}_{+1}\\ \frac{1}{T-1}\overline{F}_{-1}'\overline{F}_{+1}\end{pmatrix}$$

$$\overline{\Gamma^f}=\frac{1}{T}\left[\overline{F}_{+1}-\begin{pmatrix}\iota_{T-1} & \overline{F}_{-1}\end{pmatrix}\overline{\Pi}\right]'\left[\overline{F}_{+1}-\begin{pmatrix}\iota_{T-1} & \overline{F}_{-1}\end{pmatrix}\overline{\Pi}\right].$$

Defining

$$\overline{Q}_T=\begin{pmatrix}1 & \frac{1}{T-1}\iota_{T-1}'\overline{F}_{-1}\\ \frac{1}{T-1}\overline{F}_{-1}'\iota_{T-1} & \frac{1}{T-1}\overline{F}_{-1}'\overline{F}_{-1}\end{pmatrix}$$

$$Q_T=\begin{pmatrix}1 & \frac{1}{T-1}\iota_{T-1}'F_{-1}^{0}\\ \frac{1}{T-1}F_{-1}^{0\prime}\iota_{T-1} & \frac{1}{T-1}F_{-1}^{0\prime}F_{-1}^{0}\end{pmatrix}$$

$$Q=\begin{pmatrix}1 & \mu_F'\\ \mu_F & \Omega_F\end{pmatrix},$$

Q is a positive definite $(r+1)\times(r+1)$ matrix such that

$$\overline{Q}_T,\quad Q_T\xrightarrow{p}Q.$$

Thsi implies that $(\overline{Q}_T)^{-1} = O_p(1)$ with probability limit equal to Q^{-1}. In addition,

$$\begin{pmatrix} \frac{1}{T-1}\iota'_{T-1}\overline{F}_{+1} \\ \\ \frac{1}{T-1}\overline{F}'_{-1}\overline{F}_{+1} \end{pmatrix} - \begin{pmatrix} \frac{1}{T-1}\iota'_{T-1}F^0_{+1} \\ \\ \frac{1}{T-1}F^{0\prime}_{-1}F^0_{+1} \end{pmatrix} = o_p(1),$$

so that

$$\overline{\Pi} = (\overline{Q}_T)^{-1} \begin{pmatrix} \frac{1}{T-1}\iota'_{T-1}F^0_{+1} \\ \\ \frac{1}{T-1}F^{0\prime}_{-1}F^0_{+1} \end{pmatrix} + o_p(1).$$

Since

$$f^{0\prime}_t = c^{0\prime} + f^{0\prime}_{t-1}G^{0\prime} + u'_t H^{0\prime},$$

for any $t \in \mathbb{Z}$, we have

$$F^0_{+1} = \begin{pmatrix} \iota_{T-1} & F^0_{-1} \end{pmatrix} \Pi^0 + U \cdot H^{0\prime},$$

where we define $U = (u_2, \cdots, u_T)'$. Therefore,

$$\begin{aligned} \begin{pmatrix} \frac{1}{T-1}\iota'_{T-1}F^0_{+1} \\ \\ \frac{1}{T-1}F^{0\prime}_{-1}F^0_{+1} \end{pmatrix} &= \frac{1}{T-1}\begin{pmatrix} \iota'_{T-1} \\ F^{0\prime}_{-1} \end{pmatrix} F^0_{+1} \\ &= \frac{1}{T-1}\begin{pmatrix} \iota'_{T-1} \\ F^{0\prime}_{-1} \end{pmatrix}\begin{pmatrix} \iota_{T-1} & F^0_{-1} \end{pmatrix}\Pi^0 + \frac{1}{T-1}\begin{pmatrix} \iota'_{T-1} \\ F^{0\prime}_{-1} \end{pmatrix} U \cdot H^{0\prime} \\ &= Q_T \cdot \Pi^0 + \begin{pmatrix} \frac{1}{T-1}\sum_{t=2}^T u'_t \\ \\ \frac{1}{T-1}\sum_{t=2}^T f^0_{t-1}u'_t \end{pmatrix} H^{0\prime}. \end{aligned}$$

We showed above that, due to our assumptions,

$$\begin{aligned} \frac{1}{T-1}\sum_{t=2}^T u_t &\xrightarrow{p} O_{r\times 1} \\ \frac{1}{T-1}\sum_{t=2}^T f^0_{t-1}u'_t &\xrightarrow{p} O_{r\times r}. \end{aligned}$$

As such,

$$\overline{\Pi} = (\overline{Q}_T)^{-1}Q_T\Pi^0 + o_p(1),$$

and as $N,T \to \infty$,

$$\overline{\Pi} \xrightarrow{p} \Pi^0,$$

which proves the consistency of $\bar{c}$ and $\overline{G}$.

The consistency of $\overline{\Gamma^f}$ now follows easily. Note that

$$\begin{aligned}
\overline{\Gamma^f} &= \frac{1}{T}\left[\overline{F}_{+1} - \begin{pmatrix}\iota_{T-1} & \overline{F}_{-1}\end{pmatrix}\overline{\Pi}\right]'\left[\overline{F}_{+1} - \begin{pmatrix}\iota_{T-1} & \overline{F}_{-1}\end{pmatrix}\overline{\Pi}\right] \\
&= \frac{1}{T}\overline{F}'_{+1}\overline{F}_{+1} - \overline{\Pi}'\frac{1}{T}\begin{pmatrix}\iota'_{T-1} \\ \overline{F}'_{-1}\end{pmatrix}\overline{F}_{+1} \\
&\quad - \frac{1}{T}\overline{F}'_{+1}\begin{pmatrix}\iota_{T-1} & \overline{F}_{-1}\end{pmatrix}\overline{\Pi} + \overline{\Pi}'\overline{Q}_T\overline{\Pi}.
\end{aligned}$$

Using the fact that $\overline{\Pi} = O_p(1)$ and the results proven above, we can see that

$$\begin{aligned}
\overline{\Gamma^f} &= \frac{1}{T}F^{0\prime}_{+1}F^0_{+1} - \overline{\Pi}'\frac{1}{T}\begin{pmatrix}\iota'_{T-1} \\ F^{0\prime}_{-1}\end{pmatrix}F^0_{+1} \\
&\quad - \frac{1}{T}F^{0\prime}_{+1}\begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\overline{\Pi} + \overline{\Pi}'Q_T\overline{\Pi} + o_p(1) \\
&= \frac{1}{T}\left[F^0_{+1} - \begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\overline{\Pi}\right]'\left[F^0_{+1} - \begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\overline{\Pi}\right] + o_p(1).
\end{aligned}$$

Now we proceed as in the usual OLS case: since

$$F^0_{+1} = \begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\Pi^0 + U\cdot H^{0\prime},$$

we can see that

$$\begin{aligned}
\overline{\Gamma^f} &= \frac{1}{T}\left[\begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\left(\Pi^0 - \overline{\Pi}\right) + U\cdot H^{0\prime}\right]'\left[\begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\left(\Pi^0 - \overline{\Pi}\right) + U\cdot H^{0\prime}\right] + o_p(1) \\
&= \left(\Pi^0 - \overline{\Pi}\right)'Q_T\left(\Pi^0 - \overline{\Pi}\right) + H^0\cdot\left[\frac{1}{T}U'\begin{pmatrix}\iota_{T-1} & F^0_{-1}\end{pmatrix}\right]\left(\Pi^0 - \overline{\Pi}\right) \\
&\quad + \left(\Pi^0 - \overline{\Pi}\right)'\left[\frac{1}{T}\begin{pmatrix}\iota'_{T-1} \\ F^{0\prime}_{-1}\end{pmatrix}U\right]\cdot H^{0\prime} + H^0\left(\frac{1}{T}U'U\right)H^{0\prime} + o_p(1).
\end{aligned}$$

We have already shown that

$$\begin{aligned}
\Pi^0 - \overline{\Pi} &= o_p(1), \\
Q_T &= O_p(1)
\end{aligned}$$

$$\frac{1}{T}\begin{pmatrix}\iota'_{T-1}\\ F^{0\prime}_{-1}\end{pmatrix}U=\begin{pmatrix}\frac{1}{T}\sum_{t=2}^{T}u'_t\\ \frac{1}{T}\sum_{t=2}^{T}f^0_{t-1}u'_t\end{pmatrix}=o_p(1),$$

and by assumption,

$$\frac{1}{T}U'U=\frac{1}{T}\sum_{t=2}^{T}u_tu'_t\xrightarrow{p}I_r.$$

Therefore,

$$\overline{\Gamma^f}\xrightarrow{p}H^0H^{0\prime}=\Gamma^{f0},$$

establishing the consistency of $\overline{\Gamma^f}$.

E Proof of Dai-Singleton Canonical Model Identification

Consider a Gaussian ATSM with short rate and risk-neutral dynamics specified as

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where we assume $G^{\mathbb{Q}}$ has real and distinct eigenvalues within the unit circle. No restrictions are imposed on the physical factor dynamics of the model.

Below we show that this ATSM is observationally equivalent to a Gaussian ATSM satisfying the restrictions of the canonical Dai-Singleton model, and that this canonical model is identified against invariant affine transformations.

1) **Equivalence of True Model to Canonical Form**
We proceed in steps. First, define

$$f_t^{(1)} = \Sigma^{-1} \cdot f_t.$$

Under this rotation, the short rate and risk-neutral dynamics are given as

$$\begin{aligned} r_t &= \delta + \beta^{(1)\prime} f_t \\ f_{t+1}^{(1)} &= K^{\mathbb{Q}(1)} + G^{\mathbb{Q}(1)} f_t^{(1)} + v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where

$$\begin{aligned} \beta^{(1)} &= \Sigma' \beta \\ K^{\mathbb{Q}(1)} &= \Sigma^{-1} K^{\mathbb{Q}} \\ G^{\mathbb{Q}(1)} &= \Sigma^{-1} G^{\mathbb{Q}} \Sigma. \end{aligned}$$

This has the effect of normalizing the scale and cross-correlation of the factor innovations to 1 and 0, respectively. $G^{\mathbb{Q}}$ and $G^{\mathbb{Q}(1)}$ share the same eigenvalues, so the eigenvalues of the latter are real and within the unit circle as well.

Now consider the Schur decomposition

$$G^{\mathbb{Q}(1)\prime} = U L^{\mathbb{Q}\prime} U^*$$

of $G^{\mathbb{Q}(1)\prime}$, where U is a complex Hermitian matrix and $L^{\mathbb{Q}\prime}$ is an upper triangular

matrix with diagonals equal to the eigenvalues of $G^{\mathbb{Q}\prime}$. Since the eigenvalues of $G^{\mathbb{Q}(1)}$ are real-valued, $L^{\mathbb{Q}\prime}$ and U are both real-valued matrices; by implication, U is an orthogonal matrix. Otherwise, if there exists a complex eigenvalue of $G^{\mathbb{Q}}$ and thus $G^{\mathbb{Q}(1)}$, then the Schur decomposition involves complex matrices, so that we must amend our approach. This is the reason for requiring $G^{\mathbb{Q}}$ to have real eigenvalues.

We can further choose $L^{\mathbb{Q}}$ so that its diagonal entries are decreasing. This results in the equality

$$L^{\mathbb{Q}} = U' G^{\mathbb{Q}(1)} U,$$

where $L^{\mathbb{Q}}$ is a lower triangular matrix.

Define the rotation

$$f_t^{(2)} = U' \cdot f_t^{(1)}$$

of the factors. Under $f_t^{(2)}$, the short rate and risk-neutral dynamics become

$$\begin{aligned} r_t &= \delta + \beta^{(2)\prime} \cdot f_t^{(2)} \\ f_{t+1}^{(2)} &= K^{\mathbb{Q}(2)} + L^{\mathbb{Q}} \cdot f_t^{(2)} + \Sigma^{(2)} \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where

$$\begin{aligned} \beta^{(2)\prime} &= U' \cdot \beta^{(1)} \\ K^{\mathbb{Q}(2)} &= U' \cdot K^{\mathbb{Q}(1)} \\ \Sigma^{(2)} &= U'. \end{aligned}$$

Since $\Sigma^{(2)}\Sigma^{(2)\prime} = U'U = I_n$ due to the orthogonality of U, the factor innovation variance remains unchanged under $f_t^{(2)}$ compared to $f_t^{(1)}$. This means that the short rate and risk-netural dynamics can be written as

$$\begin{aligned} r_t &= \delta + \beta^{(2)\prime} \cdot f_t^{(2)} \\ f_{t+1}^{(2)} &= K^{\mathbb{Q}(2)} + L^{\mathbb{Q}} \cdot f_t^{(2)} + v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

Finally, consider the translation

$$f_t^{(3)} = -\left(I_n - L^{\mathbb{Q}}\right)^{-1} K^{\mathbb{Q}(2)} + f_t^{(2)}.$$

Note that the inverse $\left(I_n - L^{\mathbb{Q}}\right)^{-1}$ is well-defined because none of the eigenvalues of

$L^{\mathbb{Q}}$, which are equal to its diagonal elements, are equal to 1. Under $f_t^{(3)}$, the short rate and risk-neutral dynamics become

$$\begin{aligned} r_t &= \left[\delta^{(2)} + \beta^{(2)\prime}\left(I_n - L^{\mathbb{Q}}\right)^{-1} K^{\mathbb{Q}(2)}\right] + \beta^{(2)\prime} \cdot f_t^{(3)} \\ f_{t+1}^{\mathbb{Q}} &= L^{\mathbb{Q}} \cdot f_t^{(3)} + v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

To set all the signs of $\beta^{(2)}$ to be non-negative, we need only rotate the factors one last time so that, if $\beta_i^{(2)}$ is negative, then $f_{it}^{(2)}$ is multiplied by -1. The resulting model satisfies the restrictions of the canonical Dai-Singleton model.

2) **Uniqueness of Canonical Form**

Now we show that an ATSM in the Dai-Singleton canonical form is identified against invariant affine transformations. Suppose the short rate and risk-neutral dynamics are given in the Dai-Singleton canonical form under the factors f_t:

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= G^{\mathbb{Q}} f_t + v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where $G^{\mathbb{Q}}$ is a real lower triangular matrix. Let X_t be an invariant affine transformation of f_t:

$$X_t = A + B f_t,$$

and supose the short rate and risk-neutral dynamics

$$\begin{aligned} r_t &= \delta_X + \beta_X' f_t \\ X_{t+1} &= K_X^{\mathbb{Q}} + G_X^{\mathbb{Q}} \cdot X_t + \Sigma_X \cdot v_{t+1}^{\mathbb{Q}} \end{aligned}$$

under X_t also satisfy the restrictions placed on the canonical Dai-Singleton model.

First, we show that $B = I_n$. The factor innovation variance, mean reversion coefficient, and factor loadings on the short rate are given as

$$\begin{aligned} \Sigma_X \Sigma_X' &= I_n = BB' \\ G_X^{\mathbb{Q}} &= B G^{\mathbb{Q}} B^{-1} \\ \beta_X &= B^{-1\prime}\beta. \end{aligned}$$

The first equation tells us that B is an orthogonal matrix, and the second equation shows us that $BG^{\mathbb{Q}}B^{-1}$ is a Schur decomposition of $G_X^{\mathbb{Q}}$, which is also a lower triangular matrix with ordered diagonals. Furthermore, the last equation tells us that the signs of $B^{-1\prime}\beta = B\beta$ must all be non-negative.

In Hamilton and Wu (2012), it is shown that there exist orthogonal matrices B that are not equal to the identity matrix but nonetheless make $BG^{\mathbb{Q}}B'$ lower triangular and $B\beta$ a vector with non-negative entries. The resulting lower triangular matrix shares diagonal entries with $G^{\mathbb{Q}}$, albeit with their order switched around. The only value of B that makes $BG^{\mathbb{Q}}B'$ lower triangular and $B\beta$ a vector with non-negative entries, while simultaneously preserving the order of the diagonal entries, is I_n. It is for this reason that we impose the condition that the diagonal entries of $G^{\mathbb{Q}}$ must be ordered, and that the eigenvalues comprising the diagonals are distinct.

We have shown that $B = I_n$. It remains to show that $A = O_{n\times 1}$. This can be shown by studying the equation

$$K_X^{\mathbb{Q}} = O_{n\times 1} = \left(I_n - BG^{\mathbb{Q}}B'\right)A.$$

Since $G^{\mathbb{Q}}$ and $BG^{\mathbb{Q}}B' = G_X^{\mathbb{Q}}$ share the same eigenvalues, which are all contained in the unit circle, the matrix $I_n - BG^{\mathbb{Q}}B'$ is nonsingular, which allows us to conclude that $A = O_{n\times 1}$.

If $G^{\mathbb{Q}}$ contains even a single unit root, then the matrix $I_n - BG^{\mathbb{Q}}B'$ is singular and we cannot conclude that A equals the zero vector. This is the local identification issue pointed out in Hamilton and Wu (2012), and the reason we require $G^{\mathbb{Q}}$ to have eigenvalues within the unit circle. A similar issue arises when we impose identification restrictions on $K^{\mathbb{P}}$ instead of $K^{\mathbb{Q}}$ and $G^{\mathbb{P}}$ has a unit root.

F Proof of Canonical JSZ Model Identification

Consider a Gaussian ATSM with short rate and risk-neutral dynamics specified as

$$\begin{aligned} r_t &= \delta + \beta' f_t \\ f_{t+1} &= K^{\mathbb{Q}} + G^{\mathbb{Q}} f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

We assume that the eigenvalues of $G^{\mathbb{Q}}$ all lie on or within the unit circle. In addition, we assume, for the sake of simplicity, that the eigenvalues of $G^{\mathbb{Q}}$ are all real and distinct[12].

Below we show that this ATSM is observationally equivalent to a Gaussian ATSM satisfying the restrictions of the JSZ canonical model, and that this canonical model is identified against invariant affine transformations.

1) **Equivalence of True Model to Canonical Form**
Since $G^{\mathbb{Q}}$ has real and distinct eigenvalues, it has eigendecomposition

$$G^{\mathbb{Q}} = P \cdot J^{\mathbb{Q}} \cdot P^{-1}$$

for some nonsingular matrix P and diagonal matrix $J^{\mathbb{Q}}$ whose diagonal elements, collected in the n-dimensional vector

$$\lambda^{\mathbb{Q}} = (\lambda_1^{\mathbb{Q}}, \cdots, \lambda_n^{\mathbb{Q}}),$$

are the ordered eigenvalues of $G^{\mathbb{Q}}$, that is,

$$\lambda_1^{\mathbb{Q}} > \cdots > \lambda_n^{\mathbb{Q}}.$$

If there exists an eigenvalue on the unit circle, we order it first. As in the Dai-Singleton model, we proceed step by step.

Using P, we first define the rotation

$$f_t^{(1)} = P^{-1} f_t$$

of the true model. Under $f_t^{(1)}$, the short rate and risk-neutral dynamics are given

[12]The case of complex and possibly non-distinct eigenvalues is studied in Joslin, Singleton, and Zhu (2011). Despite the added generality of complex and non-distinct eignevalues, the case of real and distinct eigenvalues is a useful simplification, used most notably in Bauer and Rudebusch (2020).

by

$$\begin{aligned} r_t &= \delta + \beta^{(1)\prime} f_t^{(1)} \\ f_{t+1}^{(1)} &= K^{\mathbb{Q}(1)} + J^{\mathbb{Q}} \cdot f_t^{(1)} + \Sigma^{(1)} \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where

$$\begin{aligned} \beta^{(1)} &= P'\beta \\ K^{\mathbb{Q}(1)} &= P^{-1}K^{\mathbb{Q}} \\ \Sigma^{(1)} &= P^{-1}\Sigma. \end{aligned}$$

Since a diagonal matrix is technically in Jordan form, it remains to normalize the loadings β and the intercept δ.

Suppose that the ith element of $\beta^{(1)}$ is equal to 0. Then, due to the diagonal nature of $J^{\mathbb{Q}}$, we can remove the factor $f_{it}^{(1)}$ and reformulate the short rate and risk-neutral factor dynamics. Therefore, we may assume without loss of generality that every element of $\beta^{(1)}$ is non-zero. Now define the rotation

$$f_t^{(2)} = \begin{pmatrix} \beta_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \beta_n^{(1)} \end{pmatrix} f_t^{(1)} = \begin{pmatrix} \beta_1^{(1)} f_{1t}^{(1)} \\ \vdots \\ \beta_n^{(1)} f_{1t}^{(1)} \end{pmatrix}.$$

Then,

$$r_t = \delta + \beta^{(1)\prime} \begin{pmatrix} \frac{1}{\beta_1^{(1)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\beta_n^{(1)}} \end{pmatrix} \cdot f_t^{(2)} = \delta + \iota' f_t^{(2)}$$

$$f_{t+1}^{(2)} = K^{\mathbb{Q}(2)} + J^{\mathbb{Q}} \cdot f_t^{(2)} + \Sigma^{(2)} \cdot v_{t+1}^{\mathbb{Q}},$$

where

$$K^{\mathbb{Q}(2)} = \begin{pmatrix} \beta_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \beta_n^{(1)} \end{pmatrix} \cdot K^{\mathbb{Q}(1)}, \quad \Sigma^{(2)} = \begin{pmatrix} \beta_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \beta_n^{(1)} \end{pmatrix} \cdot \Sigma^{(1)}.$$

Finally, define

$$A = \begin{pmatrix} -\delta + \sum_{i=2}^{n} \left(1 - \lambda_i^{\mathbb{Q}}\right)^{-1} K_i^{\mathbb{Q}(2)} \\ -\left(1 - \lambda_2^{\mathbb{Q}}\right)^{-1} K_2^{\mathbb{Q}(2)} \\ \vdots \\ -\left(1 - \lambda_n^{\mathbb{Q}}\right)^{-1} K_n^{\mathbb{Q}(2)} \end{pmatrix},$$

where the reciprocal of $1 - \lambda_i^{\mathbb{Q}}$ is well-defined for any $2 \leq i \leq n$ due to the assumption of distinct eigenvalues. Now consider the translation

$$f_t^{(3)} = A + f_t^{(2)}.$$

Note that

$$\delta + \iota' A = \delta + \left(-\delta + \sum_{i=2}^{n} \left(1 - \lambda_i^{\mathbb{Q}}\right)^{-1} K_i^{\mathbb{Q}(2)}\right) - \sum_{i=2}^{n} \left(1 - \lambda_i^{\mathbb{Q}}\right)^{-1} K_i^{\mathbb{Q}(2)} = 0,$$

and

$$\begin{aligned} \left(I_n - J^{\mathbb{Q}}\right) A + K^{\mathbb{Q}(2)} &= \begin{pmatrix} \left(1 - \lambda_1^{\mathbb{Q}}\right) A_1 + K_1^{\mathbb{Q}(2)} \\ \vdots \\ \left(1 - \lambda_n^{\mathbb{Q}}\right) A_n + K_n^{\mathbb{Q}(2)} \end{pmatrix} \\ &= \begin{pmatrix} K_1^{\mathbb{Q}(2)} - \left(1 - \lambda_1^{\mathbb{Q}}\right) \delta + \sum_{i=2}^{n} \frac{1 - \lambda_1^{\mathbb{Q}}}{1 - \lambda_i^{\mathbb{Q}}} K_i^{\mathbb{Q}(2)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \end{aligned}$$

Defining

$$k_\infty^{\mathbb{Q}} = K_1^{\mathbb{Q}(2)} - \left(1 - \lambda_1^{\mathbb{Q}}\right) \delta + \sum_{i=2}^{n} \frac{1 - \lambda_1^{\mathbb{Q}}}{1 - \lambda_i^{\mathbb{Q}}} K_i^{\mathbb{Q}(2)},$$

it follows that

$$\begin{aligned} r_t &= \iota' f_t^{(3)} \\ f_{t+1}^{(3)} &= \begin{pmatrix} k_\infty^{\mathbb{Q}} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \lambda_1^{\mathbb{Q}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{\mathbb{Q}} \end{pmatrix} \cdot f_t^{(3)} + \Sigma^{\mathbb{Q}} \cdot v_{t+1}^{\mathbb{Q}}, \end{aligned}$$

where $\Sigma^{\mathbb{Q}}$ is the Cholesky factor of $\Sigma^{(2)}\Sigma^{(2)\prime}$. This model satisfies the constraints of the canonical JSZ model, so that the risk-netural dynamics are summarized by the $n+1+\frac{n(n+1)}{2}$ parameters contained in

$$k_\infty^{\mathbb{Q}}, \lambda^{\mathbb{Q}}, \Sigma^{\mathbb{Q}}.$$

2) **Uniqueness of Canonical Form**

Now we will show that the JSZ canonical form derived above is identified against invariant affine transformations. Suppose the short rate and risk-neutral dynamics are given in the JSZ canonical form under the factors f_t:

$$\begin{aligned} r_t &= \iota' f_t \\ f_{t+1} &= \underbrace{\begin{pmatrix} k_\infty^{\mathbb{Q}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{K^{\mathbb{Q}}} + \underbrace{\begin{pmatrix} \lambda_1^{\mathbb{Q}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^{\mathbb{Q}} \end{pmatrix}}_{G^{\mathbb{Q}}} \cdot f_t + \Sigma \cdot v_{t+1}^{\mathbb{Q}}. \end{aligned}$$

Let X_t be an invariant affine transformation of f_t:

$$X_t = A + B f_t,$$

and suppose the short rate and risk neutral dynamics

$$\begin{aligned} r_t &= \delta_X + \beta_X' f_t \\ X_{t+1} &= K_X^{\mathbb{Q}} + G_X^{\mathbb{Q}} X_t + \Sigma_X \cdot v_{t+1}^{\mathbb{Q}} \end{aligned}$$

under X_t also satisfy the JSZ restrictions.

First, we show that $B = I_n$. From our derivation above

$$G_X^{\mathbb{Q}} = B G^{\mathbb{Q}} B^{-1},$$

and $G^{\mathbb{Q}}$ is a diagonal matrix, so this tells us that the right hand side is the eigendecomposition, which is a special case of the Jordan decomposition, of $G_X^{\mathbb{Q}}$. However, since $G_X^{\mathbb{Q}}$ is in Jordan form, we can see that

$$G_X^{\mathbb{Q}} = I_n \cdot G_X^{\mathbb{Q}} \cdot I_n^{-1}$$

is also a Jordan decomposition of $G_X^{\mathbb{Q}}$. Furthermore, because the Jordan blocks are ordered in both $G^{\mathbb{Q}}$ and $G_X^{\mathbb{Q}}$, this decomposition is unique even against the ordering of the blocks. It follows that $G_X^{\mathbb{Q}} = G^{\mathbb{Q}}$, so that

$$G^{\mathbb{Q}} \cdot B = B \cdot G^{\mathbb{Q}}.$$

Let the (i,j)th element of B be denoted B_{ij}. Then, the above equation implies that

$$\lambda_i^{\mathbb{Q}} B_{ij} = B_{ij} \lambda_j^{\mathbb{Q}}$$

for any $1 \leq i,j \leq m$. If $i \neq j$, then because $\lambda_i^{\mathbb{Q}} \neq \lambda_j^{\mathbb{Q}}$, B_{ij} must be equal to 0. By implication, B is a diagonal matrix with diagonal entries equal to $B_{11}, \cdots, B_{mm}$.

In addition, the condition that

$$\iota' = \beta' = \beta_X' B = \iota' B$$

also tells us that each $B_{ii} = 1$. Therefore, $B = I_n$.

It remains to show that $A = O_{n \times 1}$. Note first that

$$0 = \delta_X = -\iota' A,$$

so that the elements of A must sum to 0. Let A_i be the ith element of A. Since

$$K_X^{\mathbb{Q}} = \left(I_n - G^{\mathbb{Q}}\right) A + K^{\mathbb{Q}}$$

and the last $n-1$ elements of $K_X^{\mathbb{Q}}$ and $K^{\mathbb{Q}}$ are all equal to 0, we can see that

$$\left(1 - \lambda_i^{\mathbb{Q}}\right) A_i = 0$$

for any $2 \leq i \leq n$. Since the second eigenvalue onward lies within the unit circle, $1 - \lambda_i^{\mathbb{Q}} \neq 0$ and $A_i = 0$. Since the elements of A must sum to 0, $A_1 = 0$ as well. Therefore, $A = O_{n \times 1}$.

G Why Minimum Chi-Square Estimation Works

Suppose that there exist a set of reduced form parameters π and the corresponding structural parameters θ. Suppose the reduced form parameters and structural parameters are mapped via the function $g(\cdot)$ on the structural parameter space Θ; that is, given the structural parameters θ, the reduced form parameters are given as

$$\pi = g(\theta).$$

Since the log likelihood depends only on the reduced form parameters π, we can formulate the sample log likelihood $V_T(\cdot)$ as a function on the reduced form parameter space Π. We can now make the standard assumption that $V_T(\cdot)$ converges uniformly in probability to a population log likelihood $V(\cdot)$ that is uniquely minimized at the true reduced form parameters π_0. If Π is compact, then by Newey and MacFadden (1994), the MLE $\hat{\pi}$ of the reduced form parameters is consistent for π_0.

Under a fully identified model, it must be the case that

$$\pi_0 = g(\theta) \quad \text{if and only if} \quad \theta = \theta_0.$$

In other words, the true reduced form parameters, which determine the value of the likelihood, are observed under only one set of structural parameters, namely the true structural parameters. Note that this requires the number of reduced form parameters, k, to be at least as large as the number of structural parameters, m. This is the order condition of identification.

The MSCE $\hat{\theta}$ of θ can be found by solving the following minimization problem, where W is some weighting matrix that is positive definite-valued:

$$\min_{\theta \in \Theta} S_T(\theta) = (\hat{\pi} - g(\theta))' W (\hat{\pi} - g(\theta)).$$

Note that this reduces to solving the equation $\hat{\pi} = g(\theta)$ when $m = k$, so that the model is just identified. The first order condition for minimization implies that $\hat{\theta}$ satisfies

$$Dg(\hat{\theta})' W \left[\hat{\pi} - g(\hat{\theta})\right] = O_{m \times 1},$$

where $Dg(\cdot)$ is the $k \times m$ matrix-valued Jacobian of $g(\cdot)$.

In the case that $g(\cdot)$ is continuous and Θ is compact, the MSCE $\hat{\theta}$ is consistent for the true structural parameters θ_0. This can be seen easily from Newey and MacFadden (1994)'s result on extremum estimators: the parameter space Θ is compact, and by the continuity of $g(\cdot)$ and the consistency of $\hat{\pi}$, $S_T(\cdot)$ converges uniformly in probability to

$S(\cdot)$ defined as

$$S(\theta) = (\pi_0 - g(\theta))' W (\pi_0 - g(\theta)).$$

Finally, since θ_0 is the unique solution to the equation $\pi_0 = g(\theta)$, $S(\theta) = 0$ if and only if $\theta = \theta_0$, meaning that $S(\cdot)$ is uniquely minimized at θ_0. Therefore,

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

for any choice of weights W.

Hamilton and Wu (2012) also show that, if W is chosen appropriately, the MSCE estimator of the structural parameters is as efficient as directly maximizing the log-likelihood $V_T \circ g$ with respect to θ. This follows from a standard stochastic MVT expansion. First, assuming that the MLE $\hat{\pi}$ satisfies the usual regularity conditions, it is asymptotically normal with asymptotic variance equal to the inverse information matrix:

$$\sqrt{T}(\hat{\pi} - \pi_0) \xrightarrow{d} \mathcal{N}\left[O_{k\times 1}, I(\pi_0)^{-1}\right],$$

where the information matrix at π is given as

$$I(\pi) = DV(\pi)' DV(\pi).$$

Define the function $Q_T(\cdot)$ on Θ as

$$Q_T(\theta) = Dg(\hat{\theta})' W [\hat{\pi} - g(\theta)].$$

Then, the stochastic MVT tells us that

$$Q_T(\hat{\theta}) = Q_T(\theta_0) + DQ_T(\theta_0)' \left(\hat{\theta} - \theta_0\right) + o_p(1),$$

where $DQ_T(\cdot)$ is the $m \times m$ matrix-valued Jacobian defined as

$$DQ_T(\theta) = -Dg(\hat{\theta})' W Dg(\theta).$$

The left hand side is equal to the zero vector as per the f.o.c. for minimization, so

$$\sqrt{T}\left(\hat{\theta} - \theta_0\right) = -\left[DQ_T(\theta_0)'\right]^{-1} \sqrt{T} Q_T(\theta_0).$$

Since

$$Q_T(\theta_0) = Dg(\hat{\theta})' W [\hat{\pi} - g(\theta_0)] = Dg(\hat{\theta})' W [\hat{\pi} - \pi_0],$$

we can see that

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)=\left(Dg(\hat{\theta})'WDg(\theta_0)\right)^{-1}Dg(\hat{\theta})'W\cdot\sqrt{T}\left(\hat{\pi}-\pi_0\right).$$

Let the sample weight matrix W converge in probability to its population value W_0. Under standard continuous differentiability and boundedness assumptions,

$$\left(Dg(\hat{\theta})'WDg(\theta_0)\right)^{-1}Dg(\hat{\theta})'W\xrightarrow{p}\left(Dg(\theta_0)'W_0Dg(\theta_0)\right)^{-1}Dg(\theta_0)'W_0$$

by the continuous mapping theorem. Therefore, by Slutsky's theorem,

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)\xrightarrow{p}\mathcal{N}\left[O_{m\times 1},\quad\left[Dg(\theta_0)'W_0Dg(\theta_0)\right]^{-1}Dg(\theta_0)'W_0I(\pi_0)^{-1}W_0Dg(\theta_0)\left[Dg(\theta_0)'W_0Dg(\theta_0)\right]^{-1}\right].$$

Note that the full-information MLE $\tilde{\theta}$ of the structural parameters maximizes the log-likelihood

$$\tilde{V}_T(\theta)=V_T(g(\theta)).$$

with respect to θ. Therefore, defining the population full-information log-likelihood as

$$\tilde{V}(\theta)=V(g(\theta)),$$

the asymptotic variance of $\tilde{\theta}$ is the inverse information matrix $\tilde{I}(\theta)$ evaluated at the true value θ_0, where

$$\begin{aligned}\tilde{I}(\theta)&=D\tilde{V}(\theta)'D\tilde{V}(\theta)\\&=Dg(\theta)'DV(g(\theta))'DV(g(\theta))Dg(\theta).\end{aligned}$$

Therefore, the asymptotic variance of $\tilde{\theta}$ is

$$\tilde{I}(\theta_0)^{-1}=\left[Dg(\theta_0)'I(\pi_0)Dg(\theta_0)\right]^{-1},$$

where we used the fact that $g(\theta_0)=\pi_0$. It follows that, if we choose

$$W=I(\hat{\pi}),$$

then

$$\sqrt{T}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left[O_{m\times 1},\quad\left[Dg(\theta_0)'I(\pi_0)Dg(\theta_0)\right]^{-1}\right].$$

This means that $\hat{\theta}$ is asymptotically as efficient as $\tilde{\theta}$ under the choice $W=I(\hat{\pi})$ of weights.

By the information matrix equality,

$$I(\hat{\pi}) = -\frac{\partial V(\hat{\pi})}{\partial \pi \partial \pi'}.$$

H Consistency of ACM Estimators

Let $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{Z}\}$ be the filtration representing the flow of information in the economy. Here we show that the estimators of the model parameters obtained via the ACM three-step method are consistent under the following assumptions:

A1. Joint Normality of Excess Bond Returns and Risk Factors
The random vector $(exr_{t+1}^{(\tau)}, v_{t+1}')'$ is normally distributed conditional on information up to time t.

A2. Homoskedastic Return Pricing Errors
The return pricing error has constant variance and no cross-sectional correlation, so that the variance of e_t is given as $\sigma^2 \cdot I_m$. We assume variance ergodicity, so that

$$\frac{1}{T}\sum_{t=1}^{T} e_t e_t' \xrightarrow{p} \sigma^2 \cdot I_m.$$

A3. Idiosyncraticity of Short Rate Pricing Errors
$u_t^{(1)}$ is independent of $\mathcal{F}_{t-1}$ and f_t.

A4. Time-Invariance of Beta Term
The beta term $\beta_t^{(\tau)}$ defined as

$$\beta_t^{(\tau)\prime} = \mathrm{Cov}_t\left(exr_{t+1}^{(\tau)}, v_{t+1}\right)\Sigma^{-1}$$

is time-invariant.

A5. Stationarity and Ergodicity of Factors
At any time t, the conditional distribution of v_t given $\mathcal{F}_{t-1}$ is

$$v_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}\left[O_{n\times 1}, I_n\right].$$

We assume variance ergodicity for the innovation process:

$$\frac{1}{T}\sum_{t=1}^{T} v_t v_t' \xrightarrow{p} I_n.$$

The eigenvalues of G are all contained within the unit circle, and the factor process

is L^2-bounded. By implication, the factors f_t are weakly stationary with a causal linear process representation with innovation process $\{v_t\}_{t\in\mathbb{Z}}$.

It is also assumed that $\{f_t\}_{t\in\mathbb{Z}}$ is mean and variance ergodic, so that

$$\frac{1}{T}\sum_{t=1}^{T} f_t \xrightarrow{p} \mu_F := (I_n - G)^{-1} K$$

$$\frac{1}{T}\sum_{t=1}^{T} f_t f_t' \xrightarrow{p} \Omega_F,$$

where μ_F is the unconditional mean of f_t and Ω_F a positive definite matrix defined as the sum of the unconditional variance of f_t and $\mu_F \mu_F'$.

We denote true parameters with 0 subscripts. Below we assume that the factors f_t are perfectly observable, so that we do not have to worry about the generated regressor problem. If consistent estimates of the factors are used instead of the true latent factors, then we need only refer back to appendix D.

First, we study some properties of the factor innovation process $\{v_t\}_{t\in\mathbb{Z}}$, the short rate pricing error process $\{u_t^{(1)}\}_{t\in\mathbb{Z}}$, and the return pricing error process $\{e_t\}_{t\in\mathbb{Z}}$. Since

$$\begin{aligned}\mathbb{E}\left[v_t \mid \mathcal{F}_{t-1}\right] &= O_{n\times 1} \\ \mathbb{E}\left[f_{t-1} v_t' \mid \mathcal{F}_{t-1}\right] = f_{t-1} \cdot \mathbb{E}\left[v_t' \mid \mathcal{F}_{t-1}\right] &= O_{n\times n},\end{aligned}$$

the elements of $\{v_t\}_{t\in\mathbb{Z}}$ and $\{f_{t-1}v_t'\}_{t\in\mathbb{Z}}$ are martingale difference sequences with respect to the filtration $\mathcal{F}$. They are also L^2-bounded by the variance stationarity of both the factors and innovation variances, so by the MDS WLLN,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} v_t &\xrightarrow{p} O_{n\times 1} \\ \frac{1}{T}\sum_{t=1}^{T} f_{t-1} v_t' &\xrightarrow{p} O_{n\times n}.\end{aligned}$$

On the other hand, by design,

$$\mathbb{E}\left[e_t \mid \mathcal{F}_{t-1}\right] = O_{m\times 1}$$

and

$$\beta_0^{(\tau)\prime} \Sigma_0 \cdot \mathbb{E}\left[v_t e_t^{(\tau)} \mid \mathcal{F}_{t-1}\right] = O_{n\times 1}.$$

Stacking these observations by maturity shows us that

$$\beta_0 \Sigma_0 \cdot \mathbb{E}\left[v_t e_t' \mid \mathcal{F}_{t-1}\right] = O_{n\times m}.$$

Premultiplying both sides by β_0' and then $\Sigma_0^{-1}(\beta_0'\beta_0)^{-1}$ now shows us that

$$\mathbb{E}\left[v_t e_t' \mid \mathcal{F}_{t-1}\right] = O_{n\times m}.$$

By implication, the elements of $\{e_t\}_{t\in\mathbb{Z}}$ and $\{v_t e_t'\}_{t\in\mathbb{Z}}$ are martingale difference sequences with respect to the filtration $\mathcal{F}$. Since $f_{t-1} \in \mathcal{F}_{t-1}$,

$$\mathbb{E}\left[f_{t-1} e_t' \mid \mathcal{F}_{t-1}\right] = f_{t-1} \cdot \mathbb{E}\left[e_t' \mid \mathcal{F}_{t-1}\right] = O_{n\times m},$$

which shows us that the elements of $\{f_{t-1}e_t\}_{t\in\mathbb{Z}}$ are also martingale difference sequences with respect to the filtration $\mathcal{F}$. These three processes are also L^2-bounded, so by the MDS WLLN,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} e_t &\xrightarrow{p} O_{m\times 1} \\
\frac{1}{T}\sum_{t=1}^{T} v_t e_t' &\xrightarrow{p} O_{n\times m} \\
\frac{1}{T}\sum_{t=1}^{T} f_{t-1} e_t' &\xrightarrow{p} O_{n\times m}.
\end{aligned}$$

Finally, we can also see that

$$\begin{aligned}
\mathbb{E}\left[u_t^{(1)} \mid f_t, \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[u_t^{(1)}\right] = 0 \\
\mathbb{E}\left[f_t u_t^{(1)} \mid f_t, \mathcal{F}_{t-1}\right] &= f_t \cdot \mathbb{E}\left[u_t^{(1)} \mid f_t, \mathcal{F}_{t-1}\right] = f_t \cdot \mathbb{E}\left[u_t^{(1)}\right] = O_{n\times 1}.
\end{aligned}$$

This implies that both $\{u_t^{(1)}\}_{t\in\mathbb{Z}}$ and $\{f_t u_t^{(1)}\}_{t\in\mathbb{Z}}$ are martingale difference sequences with respect to

$$\mathcal{F}^e = \{\mathcal{F}_t \bigvee \sigma f_{t+1} \mid t \in \mathbb{Z}\}.$$

The two processes are clearly L^2-bounded, so by the MDS WLLN,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} u_t^{(1)} &\xrightarrow{p} 0 \\
\frac{1}{T}\sum_{t=1}^{T} f_t u_t^{(1)} &\xrightarrow{p} O_{n\times 1}.
\end{aligned}$$

We now deal with the estimators of the parameters governing the $\mathbb{P}$-dynamics and short rate dynamics. Since

$$F = \begin{pmatrix} \iota_T & F_{-1} \end{pmatrix} \begin{pmatrix} K_0' \\ G_0' \end{pmatrix} + V \cdot \Sigma_0',$$

we have

$$\begin{aligned} \begin{pmatrix} \hat{K} & \hat{G} \end{pmatrix} &= \begin{pmatrix} K_0 & G_0 \end{pmatrix} + \frac{1}{T} \Sigma_0 \cdot V' \begin{pmatrix} \iota_T & F_{-1} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{T}\sum_{t=1}^T f_{t-1}' \\ \frac{1}{T}\sum_{t=1}^T f_{t-1} & \frac{1}{T}\sum_{t=1}^T f_{t-1} f_{t-1}' \end{pmatrix}^{-1} \\ &= \begin{pmatrix} K_0 & G_0 \end{pmatrix} + \Sigma_0 \cdot \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T v_t & \frac{1}{T}\sum_{t=1}^T v_t f_{t-1}' \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{T}\sum_{t=1}^T f_{t-1}' \\ \frac{1}{T}\sum_{t=1}^T f_{t-1} & \frac{1}{T}\sum_{t=1}^T f_{t-1} f_{t-1}' \end{pmatrix}^{-1}. \end{aligned}$$

By the assumption of men and variance ergodicity,

$$\begin{pmatrix} 1 & \frac{1}{T}\sum_{t=1}^T f_{t-1}' \\ \frac{1}{T}\sum_{t=1}^T f_{t-1} & \frac{1}{T}\sum_{t=1}^T f_{t-1} f_{t-1}' \end{pmatrix} \xrightarrow{p} Q := \begin{pmatrix} 1 & \mu_F' \\ \mu_F & \Omega_F \end{pmatrix},$$

where Q is a positive definite matrix. Furthermore, we saw above that

$$\frac{1}{T}\sum_{t=1}^T v_t, \quad \frac{1}{T}\sum_{t=1}^T v_t f_{t-1}' = o_p(1).$$

It follows that

$$\begin{pmatrix} \hat{K} & \hat{G} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} K_0 & G_0 \end{pmatrix}.$$

Define

$$\tilde{V} = F - \iota_T \cdot \hat{K}' - F_{-1} \hat{G}'.$$

Then,

$$\tilde{V} = \iota_T \left(K_0 - \hat{K} \right)' + F_{-1} \left(G_0 - \hat{G} \right)' + V \cdot \Sigma_0',$$

which implies that

$$\frac{1}{T} \left\| \tilde{V} - V \cdot \Sigma_0' \right\|^2 \leq 2 \left| K_0 - \hat{K} \right|^2 + 2 \frac{1}{\sqrt{T}} \cdot \left\| \frac{1}{\sqrt{T}} F_{-1} \right\|^2 \cdot \left\| G_0 - \hat{G} \right\|^2$$

Since $\left\|\frac{1}{\sqrt{T}}F_{-1}\right\|^2$ is $O_p(1)$ and the rest of the terms are $o_p(1)$, we can see that

$$\frac{1}{T}\left\|\tilde{V} - V \cdot \Sigma_0'\right\|^2 = o_p(1),$$

or equivalently,

$$\frac{1}{\sqrt{T}}\left(\tilde{V} - V \cdot \Sigma_0'\right) = o_p(1).$$

It follows that

$$\frac{1}{T}\tilde{V}'\tilde{V} - \Sigma_0\left(\frac{1}{T}V'V\right)\Sigma_0' = o_p(1)$$

as well, and since $\frac{1}{T}V'V \xrightarrow{p} I_n$, it follows that

$$\hat{\Omega} = \frac{1}{T}\tilde{V}'\tilde{V} \xrightarrow{p} \Sigma_0\Sigma_0' = \Omega_0.$$

It follows that, defining

$$\hat{V} = \tilde{V} \cdot \hat{\Sigma}_0^{-1\prime}$$

we have

$$\frac{1}{\sqrt{T}}\left(\hat{V} - V\right) = \frac{1}{\sqrt{T}}\left(\tilde{V} - V \cdot \Sigma_0'\right)\hat{\Sigma}_0^{-1\prime} + \frac{1}{\sqrt{T}}V\left(\Sigma_0 - \hat{\Sigma}_0\right)'\hat{\Sigma}_0^{-1\prime} = o_p(1).$$

The generated factor innovations are consistent for the true innovations. By implication,

$$\begin{aligned} &\frac{1}{T}\hat{V}'\iota_T \xrightarrow{p} O_{n\times 1}, \\ \frac{1}{T}\hat{V}'V, \quad &\frac{1}{T}\hat{V}'\hat{V} \xrightarrow{p} I_n \end{aligned}$$

and

$$\frac{1}{T}F_{-1}'\hat{V} - \frac{1}{T}F_{-1}'V = \left(\frac{1}{\sqrt{T}}F_{-1}\right)'\left[\frac{1}{\sqrt{T}}\left(\hat{V} - V\right)\right] = o_p(1).$$

As for the short rate parameters, since

$$r = \begin{pmatrix}\iota_T & F\end{pmatrix}\begin{pmatrix}\delta_{00} \\ \delta_{10}\end{pmatrix} + u^{(1)}$$

where $u^{(1)} = (u_1^{(1)}, \cdots, u_T^{(1)})'$, we have

$$\begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \end{pmatrix} = \begin{pmatrix} \delta_{00} \\ \delta_{10} \end{pmatrix} + \left[\frac{1}{T} \begin{pmatrix} \iota_T' \\ F' \end{pmatrix} \begin{pmatrix} \iota_T & F \end{pmatrix} \right]^{-1} \frac{1}{T} \begin{pmatrix} \iota_T' \\ F' \end{pmatrix} u^{(1)}.$$

Here,

$$\frac{1}{T} \begin{pmatrix} \iota_T' \\ F' \end{pmatrix} \begin{pmatrix} \iota_T & F \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{T}\sum_{t=1}^T f_t' \\ \frac{1}{T}\sum_{t=1}^T f_t & \frac{1}{T}\sum_{t=1}^T f_t f_t' \end{pmatrix} \xrightarrow{p} Q = \begin{pmatrix} 1 & \mu_F' \\ \mu_F & \Omega_F \end{pmatrix},$$

and

$$\frac{1}{T} \begin{pmatrix} \iota_T' \\ F' \end{pmatrix} u^{(1)} = \begin{pmatrix} \frac{1}{T}\sum_{t=1}^T u_t^{(1)} \\ \frac{1}{T}\sum_{t=1}^T f_t u_t^{(1)} \end{pmatrix} = o_p(1),$$

so it follows that

$$\begin{pmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \delta_{00} \\ \delta_{10} \end{pmatrix}.$$

Now we move onto the second step. Recall that the OLS estimator of $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is defined as

$$\begin{pmatrix} \hat{\mathbf{a}} & \hat{\mathbf{b}} & \hat{\mathbf{c}} \end{pmatrix} = exr' \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \left[\begin{pmatrix} \iota_T' \\ F_{-1}' \\ \hat{V}' \end{pmatrix} \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \right]^{-1}.$$

Using the fact that

$$exr' = \begin{pmatrix} \mathbf{a}_0 & \mathbf{b}_0 & \mathbf{c}_0 \end{pmatrix} \begin{pmatrix} \iota_T' \\ F_{-1}' \\ V' \end{pmatrix} + E',$$

we can see that

$$\begin{pmatrix} \hat{\mathbf{a}} & \hat{\mathbf{b}} & \hat{\mathbf{c}} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_0 & \mathbf{b}_0 & \mathbf{c}_0 \end{pmatrix} \left[\begin{pmatrix} \iota_T' \\ F_{-1}' \\ V' \end{pmatrix} \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \right] \cdot \left[\begin{pmatrix} \iota_T' \\ F_{-1}' \\ \hat{V}' \end{pmatrix} \begin{pmatrix} \iota_T & F_{-1} & \hat{V} \end{pmatrix} \right]^{-1}$$

$$+E'\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix}\left[\begin{pmatrix}\iota_T' \\ F_{-1}' \\ \hat{V}'\end{pmatrix}\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix}\right]^{-1}.$$

From the results we derived above concerning $\hat{V}$, we can see that

$$\hat{Q}_T^e := \frac{1}{T}\begin{pmatrix}\iota_T' \\ F_{-1}' \\ \hat{V}'\end{pmatrix}\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix} = \begin{pmatrix}1 & \frac{1}{T}\iota_T' F_{-1} & \frac{1}{T}\iota_T'\hat{V} \\ \frac{1}{T}F_{-1}'\iota_T & \frac{1}{T}F_{-1}'F_{-1} & \frac{1}{T}F_{-1}'\hat{V} \\ \frac{1}{T}\hat{V}'\iota_T & \frac{1}{T}\hat{V}'F_{-1} & \frac{1}{T}\hat{V}'\hat{V}\end{pmatrix}$$
$$\xrightarrow{p} Q^e := \begin{pmatrix}1 & \mu_F' & O_{1\times n} \\ \mu_F & \Omega_F & O_{n\times n} \\ O_{n\times 1} & O_{n\times n} & I_n\end{pmatrix},$$

where Q^e is positive definite. Similarly, defining

$$Q_T^e = \frac{1}{T}\begin{pmatrix}\iota_T' \\ F_{-1}' \\ V'\end{pmatrix}\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix},$$

we have

$$Q_T^e \xrightarrow{p} Q^e.$$

Finally, we showed above that

$$\frac{1}{T}E'\iota_T = \frac{1}{T}\sum_{t=1}^{T} e_t \xrightarrow{p} O_{m\times 1}$$
$$\frac{1}{T}F_{-1}'E = \frac{1}{T}\sum_{t=1}^{T} f_{t-1}e_t' \xrightarrow{p} O_{n\times m}$$
$$\frac{1}{T}E'V = \frac{1}{T}\sum_{t=1}^{T} e_t v_t' \xrightarrow{p} O_{m\times n}$$
$$\frac{1}{T}E'E = \frac{1}{T}\sum_{t=1}^{T} e_t e_t' \xrightarrow{p} \sigma_0^2 \cdot I_m.$$

Additionally, we have

$$\frac{1}{T}E'\hat{V} - \frac{1}{T}E'V = \left(\frac{1}{\sqrt{T}}E\right)'\left[\frac{1}{\sqrt{T}}\left(\hat{V}-V\right)\right].$$

Here, $\frac{1}{\sqrt{T}}E = O_p(1)$, so

$$\frac{1}{T}E'\hat{V} - \frac{1}{T}E'V = o_p(1),$$

which implies that

$$\frac{1}{T}E'\hat{V} \xrightarrow{p} O_{n\times n}.$$

Putting all these results together, we have

$$\frac{1}{T}E'\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix} = o_p(1).$$

Therefore,

$$\begin{aligned}\begin{pmatrix}\hat{\mathbf{a}} & \hat{\mathbf{b}} & \hat{\mathbf{c}}\end{pmatrix} &= \begin{pmatrix}\mathbf{a}_0 & \mathbf{b}_0 & \mathbf{c}_0\end{pmatrix} Q_T^e \left(\hat{Q}_T^e\right)^{-1} + \frac{1}{T}E'\begin{pmatrix}\iota_T & F_{-1} & \hat{V}\end{pmatrix}\left(\hat{Q}_T^e\right)^{-1} \\ &\xrightarrow{p} \begin{pmatrix}\mathbf{a}_0 & \mathbf{b}_0 & \mathbf{c}_0\end{pmatrix}.\end{aligned}$$

Our estimator of the return pricing error variance is defined as

$$\hat{\sigma}^2 = \frac{1}{mT}\operatorname{tr}\left(\hat{E}'\hat{E}\right)$$

A process similar to the proof of the consistency of $\hat{\Omega}$ shows us that

$$\hat{\sigma}^2 \xrightarrow{p} \frac{1}{m}\operatorname{tr}\left(\sigma_0^2 \cdot I_m\right) = \sigma_0^2.$$

This proves the consistency of the second-step estimators.

The third step now involves recovering the structural paramters from the reduced-form parameters $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$. In this sense, it is similar to the MCSE step of the HW model. We can immediately see that our estimate of β is consistent:

$$\hat{\beta} = \hat{\mathbf{c}}\hat{\Sigma}^{-1} \xrightarrow{p} \mathbf{c}_0\Sigma_0^{-1} = \beta_0 = \begin{pmatrix}\beta_0^{(\tau_1)\prime} \\ \vdots \\ \beta_0^{(\tau_m)\prime}\end{pmatrix}.$$

By implication, our estimate of B is also consistent:

$$\hat{B} = \begin{pmatrix} \text{vec}\left(\hat{\beta}_1 \hat{\beta}_1'\right) \\ \vdots \\ \text{vec}\left(\hat{\beta}_m \hat{\beta}_m'\right) \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \text{vec}\left(\beta_0^{(\tau_1)} \beta_0^{(\tau_1)\prime}\right) \\ \vdots \\ \text{vec}\left(\beta_0^{(\tau_m)} \beta_0^{(\tau_m)\prime}\right) \end{pmatrix} = B_0.$$

Finally, our estimators of the market price of risk parameters are consistent:

$$\begin{aligned} \hat{\lambda} &= \left(\hat{\beta}'\hat{\beta}\right)^{-1} \hat{\beta}' \left[\hat{\mathbf{a}} + \frac{1}{2}\left(\hat{B} \cdot \text{vec}\left(\hat{\Omega}\right) + \hat{\sigma}^2 \cdot \iota_m\right)\right] \\ &\xrightarrow{p} \left(\beta_0'\beta_0\right)^{-1} \beta_0' \left[\mathbf{a}_0 + \frac{1}{2}\left(B_0 \cdot \text{vec}\left(\Omega_0\right) + \sigma_0^2 \cdot \iota_m\right)\right] = \lambda_0 \end{aligned}$$

$$\begin{aligned} \hat{\Lambda} &= \left(\hat{\beta}'\hat{\beta}\right)^{-1} \hat{\beta}'\hat{\mathbf{b}} \\ &\xrightarrow{p} \left(\beta_0'\beta_0\right)^{-1} \beta_0'\mathbf{b}_0 = \Lambda_0 \end{aligned}$$

by the continuous mapping theorem and the consistency of all parameters involved.

I Derivation of Approximate Forward Rate Formula in the Wu-Xia SRTSM

Recall that, for any $h > 0$, the shadow rate dynamics and risk-neutral dynamics imply that

$$s_{t+h} = \delta + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \right] K^{\mathbb{Q}} + \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \Sigma \cdot v_{t+h-j}^{\mathbb{Q}} \right] + \beta' \left(G^{\mathbb{Q}} \right)^h f_t.$$

We already saw that

$$\mathbb{E}_t^{\mathbb{Q}} [s_{t+h}] = \bar{a}(h) + b(h)' f_t$$

$$\mathrm{Var}_t^{\mathbb{Q}} (s_{t+h}) = \left(\sigma^{\mathbb{Q}}(h) \right)^2.$$

Define

$$a(h) = \bar{a}(h) - \frac{1}{2} \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \right] \Sigma\Sigma' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}} \right)^j \right]' \beta,$$

and note that

$$\frac{1}{2} \left[\mathrm{Var}_t^{\mathbb{Q}} \left(\sum_{j=1}^{h} s_{t+j} \right) - \mathrm{Var}_t^{\mathbb{Q}} \left(\sum_{j=1}^{h-1} s_{t+j} \right) \right] = \frac{1}{2} \left(\mathrm{Var}_t^{\mathbb{Q}} (s_{t+h}) + 2 \cdot \sum_{j=1}^{h-1} \mathrm{Cov}_t^{\mathbb{Q}} (s_{t+h},\, s_{t+j}) \right).$$

For any $1 \leq j \leq h$,

$$\mathrm{Cov}_t^{\mathbb{Q}} (s_{t+h}, s_{t+j}) = \sum_{i=0}^{h-1} \sum_{k=0}^{j-1} \beta' \left(G^{\mathbb{Q}} \right)^i \Sigma \cdot \mathrm{Cov}_t^{\mathbb{Q}} \left(v_{t+h-i}^{\mathbb{Q}}, v_{t+j-k}^{\mathbb{Q}} \right) \cdot \Sigma' \left(G^{\mathbb{Q}\prime} \right)^k \beta$$

$$= \sum_{i=0}^{j-1} \beta' \left(G^{\mathbb{Q}} \right)^{h-j+i} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^i \beta,$$

so we have

$$2 \cdot \sum_{j=1}^{h-1} \mathrm{Cov}_t^{\mathbb{Q}} (s_{t+h},\, s_{t+j}) = 2 \cdot \sum_{j=1}^{h-1} \sum_{i=0}^{j-1} \beta' \left(G^{\mathbb{Q}} \right)^{h-j+i} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^i \beta$$

$$\begin{aligned} &= 2\beta' \left(G^{\mathbb{Q}} \right)^{h-1} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^0 \beta \\ &\quad + 2\beta' \left(G^{\mathbb{Q}} \right)^{h-2} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^0 \beta + 2\beta' \left(G^{\mathbb{Q}} \right)^{h-1} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^1 \beta \\ &\quad + \cdots + 2\beta' \left(G^{\mathbb{Q}} \right)^1 \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^0 \beta + \cdots + 2\beta' \left(G^{\mathbb{Q}} \right)^{h-1} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime} \right)^{h-2} \beta \end{aligned}$$

$$
\begin{aligned}
&= \beta' \left(G^{\mathbb{Q}}\right)^{h-1} \Sigma\Sigma' \left[\sum_{j=0}^{h-2} \left(G^{\mathbb{Q}\prime}\right)^{j}\right] \beta + \cdots + \beta' \left(G^{\mathbb{Q}}\right)^{0} \Sigma\Sigma' \left[\sum_{j=1}^{h-1} \left(G^{\mathbb{Q}\prime}\right)^{j}\right] \beta \\
&= \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^{j}\right] \Sigma\Sigma' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}\prime}\right)^{j}\right] \beta - \sum_{j=0}^{h-1} \beta' \left(G^{\mathbb{Q}}\right)^{j} \Sigma\Sigma' \left(G^{\mathbb{Q}\prime}\right)^{j} \beta \\
&= \beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^{j}\right] \Sigma\Sigma' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}\prime}\right)^{j}\right] \beta - \mathrm{Var}_t^{\mathbb{Q}}\left(s_{t+h}\right).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} s_{t+j}\right) - \mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} s_{t+j}\right)\right] &= \frac{1}{2}\beta' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^{j}\right] \Sigma\Sigma' \left[\sum_{j=0}^{h-1} \left(G^{\mathbb{Q}}\right)^{j}\right]' \beta \\
&= \bar{a}(h) - a(h).
\end{aligned}
$$

In the Wu-Xia model, the h-period ahead forward rate is given as

$$
f_t^{(h)} = \mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] - \frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} r_{t+j}\right) - \mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} r_{t+j}\right)\right].
$$

To compute the forward rate, we must now express the expectation and variances above in terms of the moments of the shadow rate. We proceed in steps:

Expectations

We already saw when approximating the forward rate for the Ichiue and Ueno (2013) SRTSM that

$$
\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] = \underline{r} + \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{\bar{a}(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right),
$$

where the function $g : \mathbb{R} \to \mathbb{R}$ is defined as

$$
g(x) = x \cdot \Phi(x) + \phi(x)
$$

for any $x \in \mathbb{R}$. For notational brevity, denote

$$
z_t(j) = \frac{\bar{a}(j) + b(j)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(j)}
$$

for any $1 \leq j \leq h$.

Variances

Moving onto the variance terms, as before we have

$$\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}^2\right] = \mathbb{E}_t^{\mathbb{Q}}\left[s_{t+j}^2 \cdot I_{\{s_{t+j}\geq \underline{r}\}}\right] + \underline{r}^2 \cdot \mathbb{Q}_t\left(s_{t+j} < \underline{r}\right).$$

As above,

$$\begin{aligned}
\mathbb{E}_t^{\mathbb{Q}}\left[s_{t+j}^2 \cdot I_{\{s_{t+j}\geq \underline{r}\}}\right] &= \int_{-z_t(j)}^{\infty} \left(\sigma^{\mathbb{Q}}(j)x + \bar{a}(j) + b(j)'f_t\right)^2 \cdot \phi(x)dx \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \cdot \int_{-z_t(j)}^{\infty} x^2 \cdot \phi(x)dx + \left(\bar{a}(j) + b(j)'f_t\right)^2 \cdot \Phi(z_t(j)) \\
&\quad + 2 \cdot \left(\bar{a}(j) + b(j)'f_t\right)\sigma^{\mathbb{Q}}(j) \cdot \int_{-z_t(j)}^{\infty} x \cdot \phi(x)dx \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) - z_t(j) \cdot \phi(z_t(j))\right] + \left(\bar{a}(j) + b(j)'f_t\right)^2 \cdot \Phi(z_t(j)) \\
&\quad + 2 \cdot \left(\bar{a}(j) + b(j)'f_t\right)\sigma^{\mathbb{Q}}(j) \cdot \phi(z_t(j)),
\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}^2\right] &= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) - z_t(j) \cdot \phi(z_t(j))\right] + \left(\bar{a}(j) + b(j)'f_t\right)^2 \cdot \Phi(z_t(j)) \\
&\quad + 2 \cdot \left(\bar{a}(j) + b(j)'f_t\right)\sigma^{\mathbb{Q}}(j) \cdot \phi(z_t(j)) + \underline{r}^2 \cdot \left[1 - \Phi(z_t(j))\right] \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) + \left(z_t(j) + \frac{\underline{r}}{\sigma^{\mathbb{Q}}(j)}\right)^2 \cdot \Phi(z_t(j)) + 2\frac{\underline{r}}{\sigma^{\mathbb{Q}}(j)} \cdot \phi(z_t(j)) + z_t(j) \cdot \phi(z_t(j))\right] \\
&\quad + \underline{r}^2\left[1 - \Phi(z_t(j))\right] \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) + \left(z_t(j) + 2 \cdot \frac{\underline{r}}{\sigma^{\mathbb{Q}}(j)}\right) \cdot g(z_t(j))\right] + \underline{r}^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{Var}_t^{\mathbb{Q}}\left(r_{t+j}\right) &= \mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}^2\right] - \left(\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}\right]\right)^2 \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) + \left(z_t(j) + 2 \cdot \frac{\underline{r}}{\sigma^{\mathbb{Q}}(j)}\right) \cdot g(z_t(j))\right] + \underline{r}^2 - \left[\underline{r} + \sigma^{\mathbb{Q}}(j) \cdot g(z_t(j))\right]^2 \\
&= \left(\sigma^{\mathbb{Q}}(j)\right)^2 \left[\Phi(z_t(j)) + z_t(j) \cdot g(z_t(j)) - g(z_t(j))^2\right].
\end{aligned}$$

Wu and Xia choose to approximate the above variance as

$$\mathrm{Var}_t^{\mathbb{Q}}\left(r_{t+j}\right) \approx \left(\sigma^{\mathbb{Q}}(j)\right)^2 \cdot \Phi(z_t(j)) = \mathrm{Var}_t^{\mathbb{Q}}\left(s_{t+j}\right) \cdot \mathbb{Q}_t\left(s_{t+j} \geq \underline{r}\right).$$

They show that the absolute approximation error

$$\left|\left(\sigma^{\mathbb{Q}}(j)\right)^2\left[z_t(j)\cdot g(z_t(j))-g(z_t(j))^2\right]\right|$$

is bounded above by $\left(\sigma^{\mathbb{Q}}(j)\right)^2\phi(0)^2$, a small value.

Covariances

It remains to study the covariance terms. Choose any $1\leq j\neq k\leq h$, and note that

$$\mathrm{Cov}_t^{\mathbb{Q}}\left(r_{t+j},r_{t+k}\right)=\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}r_{t+k}\right]-\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}\right]\cdot\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+k}\right].$$

To compute the joint moment $\mathbb{E}_t^{\mathbb{Q}}\left[r_{t+j}r_{t+k}\right]$, define

$$\tilde{s}_{t+j}=\frac{s_{t+j}-\bar{a}(j)-b(j)'f_t}{\sigma^{\mathbb{Q}}(j)}\quad\text{and}$$

$$\tilde{r}_{t+j}=\frac{r_{t+j}-\bar{a}(j)-b(j)'f_t}{\sigma^{\mathbb{Q}}(j)}=\max\left(\tilde{s}_{t+j},\,-z_t(j)\right),$$

and likewise for $\tilde{s}_{t+k}$ and $\tilde{r}_{t+k}$. Here,

$$\begin{pmatrix}\tilde{s}_{t+j}\\ \tilde{s}_{t+k}\end{pmatrix}\mid\mathcal{F}_t\sim\mathcal{N}\left[O_{2\times1},\,\begin{pmatrix}1 & \rho_t(j,k)\\ \rho_t(j,k) & 1\end{pmatrix}\right],$$

under the risk-neutral measure, where $\rho_t(j,k)$ is the correlation coefficient between s_{t+j} and s_{t+k}, so the result on truncated bivariate normally distributed random vectors in the appendix shows us that

$$\begin{aligned}\mathbb{E}_t^{\mathbb{Q}}\left[\tilde{r}_{t+j}\tilde{r}_{t+k}\right]=&\,\rho_t(j,k)\cdot F(z_t(j),z_t(k);\rho_t(j,k))\\&+(1-\rho_t(j,k)^2)\cdot f(z_t(j),z_t(k);\rho_t(j,k))+z_t(j)z_t(k)\cdot F(-z_t(j),-z_t(k);\rho_t(j,k))\\&-z_t(k)\cdot h(z_t(j),z_t(k);\rho_t(j,k))-z_t(j)\cdot h(z_t(k),z_t(j);\rho_t(j,k)),\end{aligned}$$

where we define

$$f(x_1,x_2;\rho)=\frac{1}{2\pi}\left(1-\rho^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left(x_1^2+x_2^2-2\rho x_1x_2\right)\right)$$

$$h(x_1,x_2;\rho)=\phi(x_1)\cdot\Phi\left(\frac{\rho x_1-x_2}{\sqrt{1-\rho^2}}\right)$$

$$F(x_1,x_2;\rho)=\int_{-\infty}^{\alpha_1}\int_{-\infty}^{\alpha_2}f(x_1,x_2;\rho)dx_2dx_1$$

for any $x_1, x_2 \in \mathbb{R}$ and $\rho \in (-1, 1)$.

On the other hand,

$$\begin{aligned}
\mathbb{E}_t^{\mathbb{Q}}[\tilde{r}_{t+j}] &= \frac{1}{\sigma^{\mathbb{Q}}(j)}\left[\mathbb{E}_t^{\mathbb{Q}}[r_{t+j}] - \bar{a}(j) - b(j)' f_t\right] \\
&= \frac{1}{\sigma^{\mathbb{Q}}(j)}\left[\underline{r} - \bar{a}(j) - b(j)' f_t + \sigma^{\mathbb{Q}}(j) \cdot g(z_t(j))\right] \\
&= g(z_t(j)) - z_t(j) \\
&= \phi(z_t(j)) - z_t(j) \cdot (1 - \Phi(z_t(j))) = g(-z_t(j)),
\end{aligned}$$

and likewise for $\mathbb{E}_t^{\mathbb{Q}}[\tilde{r}_{t+k}]$. It follows that

$$\begin{aligned}
\mathrm{Cov}_t^{\mathbb{Q}}(r_{t+j},\, r_{t+k}) &= \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)\left[\mathbb{E}_t^{\mathbb{Q}}[\tilde{r}_{t+j}\tilde{r}_{t+k}] - \mathbb{E}_t^{\mathbb{Q}}[\tilde{r}_{t+j}] \cdot \mathbb{E}_t^{\mathbb{Q}}[\tilde{r}_{t+k}]\right] \\
&= \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)\rho_t(j,k) \cdot F(z_t(j), z_t(k); \rho_t(j,k)) \\
&\quad + \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)\left[(1 - \rho_t(j,k)^2) \cdot f(z_t(j), z_t(k); \rho_t(j,k)) + z_t(j) z_t(k) \cdot F(-z_t(j), -z_t(k); \rho_t(j,k))\right] \\
&\quad - \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)\left[z_t(k) \cdot h(z_t(j), z_t(k); \rho_t(j,k)) + z_t(j) \cdot h(z_t(k), z_t(j); \rho_t(j,k))\right] \\
&\quad - \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k) g(-z_t(j)) g(-z_t(k)).
\end{aligned}$$

Wu and Xia choose to approximate the above covariance as

$$\begin{aligned}
\mathrm{Cov}_t^{\mathbb{Q}}(r_{t+j},\, r_{t+k}) &\approx \sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)\rho_t(j,k) \cdot F(z_t(j), z_t(k); \rho_t(j,k)) \\
&= \mathrm{Cov}_t^{\mathbb{Q}}(s_{t+j},\, s_{t+k}) \cdot \mathbb{Q}_t(s_{t+j} \geq \underline{r},\, s_{t+k} \geq \underline{r}).
\end{aligned}$$

Defining

$$\begin{aligned}
D(\alpha_1, \alpha_2; \rho) = &\left(1 - \rho^2\right) \cdot f(\alpha_1, \alpha_2; \rho) + \alpha_1 \alpha_2 \cdot F(-\alpha_1, -\alpha_2; \rho) \\
&- \alpha_1 \cdot h(\alpha_1, \alpha_2; \rho) - \alpha_2 \cdot h(\alpha_2, \alpha_1; \rho) \\
&- g(-\alpha_1) g(-\alpha_2),
\end{aligned}$$

the absolute approximation error can be written as

$$\left|\sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k) \cdot D(-z_t(j), -z_t(k); \rho_t(j,k))\right|.$$

Wu and Xia show that the absolute value of this quantity is bounded above by $\sigma^{\mathbb{Q}}(j)\sigma^{\mathbb{Q}}(k)(1 - \rho_t(j,k)^2) \cdot \phi(0)^2$, a very small amount.

They also claim that, given the persistence of the shadow rate, we can approximate the conditional probability

$$\mathbb{Q}_t(s_{t+j} \geq \underline{r} \mid s_{t+k} \geq \underline{r}) = 1.$$

It follows that

$$\mathrm{Cov}_t^{\mathbb{Q}}\left(r_{t+h},\, r_{t+h-i}\right) \approx \mathrm{Cov}_t^{\mathbb{Q}}\left(s_{t+h},\, s_{t+h-i}\right) \cdot \mathbb{Q}_t\left(s_{t+h} \geq \underline{r}\right)$$

for any $0 \leq i \leq h-1$.

Approximation to the Foward Rate

Putting all these results together,

$$\begin{aligned}
\frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} r_{t+j}\right) - \mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} r_{t+j}\right)\right] &= \frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(r_{t+h}\right) + 2 \cdot \sum_{i=0}^{h-1} \mathrm{Cov}_t^{\mathbb{Q}}\left(r_{t+h},\, r_{t+h-i}\right)\right] \\
&\approx \frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(s_{t+h}\right) + 2 \cdot \sum_{i=0}^{h-1} \mathrm{Cov}_t^{\mathbb{Q}}\left(s_{t+h},\, s_{t+h-i}\right)\right] \cdot \mathbb{Q}_t\left(s_{t+h} \geq \underline{r}\right) \\
&= \frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} s_{t+j}\right) - \mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} s_{t+j}\right)\right] \cdot \mathbb{Q}_t\left(s_{t+h} \geq \underline{r}\right) \\
&= \Phi\left(\frac{\bar{a}(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) \cdot \left(\bar{a}(h) - a(h)\right),
\end{aligned}$$

and we have

$$\begin{aligned}
f_t^{(h)} &= \mathbb{E}_t^{\mathbb{Q}}\left[r_{t+h}\right] - \frac{1}{2}\left[\mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h} r_{t+j}\right) - \mathrm{Var}_t^{\mathbb{Q}}\left(\sum_{j=1}^{h-1} r_{t+j}\right)\right] \\
&\approx \underline{r} + \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{\bar{a}(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) + \Phi\left(\frac{\bar{a}(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) \cdot \left(a(h) - \bar{a}(h)\right).
\end{aligned}$$

Finally, defining the function $\bar{g} : \mathbb{R} \to \mathbb{R}$ as

$$\bar{g}(a) = \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{a + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right)$$

for any $a \in \mathbb{R}$, a first order Taylor approximation of $\bar{g}(a(h))$ around $\bar{a}(h)$ tells us that

$$\begin{aligned}
\bar{g}(a(h)) &\approx \bar{g}(\bar{a}(h)) + \frac{\partial \bar{g}(\bar{a}(h))}{\partial a} \cdot \left(a(h) - \bar{a}(h)\right) \\
&= \bar{g}(\bar{a}(h)) + \Phi\left(\frac{\bar{a}(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right) \cdot \left(a(h) - \bar{a}(h)\right) = f_t^{(h)} - \underline{r}.
\end{aligned}$$

We are thus left with the approximation

$$\begin{aligned} f_t^{(h)} &\approx \bar{g}(a(h)) \\ &= \underline{r} + \sigma^{\mathbb{Q}}(h) \cdot g\left(\frac{a(h) + b(h)' f_t - \underline{r}}{\sigma^{\mathbb{Q}}(h)}\right). \end{aligned}$$

J Moments of Bivariate Truncated Normal Random Vectors

Consider a bivariate random vector (X_1, X_2) with the following distribution:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[O_{2\times 1},\ \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right],$$

so that $\rho \in (-1, 1)$ is the correlation coefficient of X_1 and X_2. We want to find the truncated moments

$$\mathbb{E}\left[X_1 X_2 \cdot I_{\{X_1 \geq \alpha_1,\ X_2 \geq \alpha_2\}}\right], \quad \mathbb{E}\left[X_1 \cdot I_{\{X_1 \geq \alpha_1,\ X_2 < \alpha_2\}}\right], \quad \text{and} \quad \mathbb{P}(X_1 < \alpha_1,\ X_2 < \alpha_2).$$

First, note that the density of (X_1, X_2) with respect to the Lebesgue measure is given as

$$\begin{aligned} f(x_1, x_2; \rho) &= \frac{1}{2\pi} \left| \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right) \\ &= \frac{1}{2\pi} \left(1 - \rho^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(x_1^2 + x_2^2 - 2\rho x_1 x_2 \right) \right). \end{aligned}$$

Note that, letting $\phi : \mathbb{R} \to (0, +\infty)$ be the standard normal density, we can rewrite the joint density in terms of the product of two standard normal densities:

$$\begin{aligned} f(x_1, x_2; \rho) &= \frac{1}{2\pi} \left(1 - \rho^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2(1-\rho^2)} (x_1^2 + x_2^2 - 2\rho x_1 x_2) \right) \\ &= \left(1 - \rho^2\right)^{-\frac{1}{2}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x_1^2 \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2(1-\rho^2)} \left(x_2 - \rho x_1 \right)^2 \right) \\ &= \left(1 - \rho^2\right)^{-\frac{1}{2}} \phi(x_1) \cdot \phi\left(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}} \right). \end{aligned}$$

Due to the symmetry of the above operation, we also have

$$f(x_1, x_2; \rho) = \left(1 - \rho^2\right)^{-\frac{1}{2}} \phi(x_2) \cdot \phi\left(\frac{x_1 - \rho x_2}{\sqrt{1-\rho^2}} \right).$$

The following derivative will prove useful later on:

$$\frac{\partial f(x_1, x_2; \rho)}{\partial x_1} = \left[-x_1 + \frac{\rho(x_2 - \rho x_1)}{1 - \rho^2} \right] \cdot f(x_1, x_2; \rho)$$

$$= \frac{\rho x_2 - x_1}{1-\rho^2} \cdot f(x_1, x_2; \rho).$$

By implication,

$$x_1 \cdot f(x_1, x_2; \rho) = \rho x_2 \cdot f(x_1, x_2; \rho) - (1-\rho^2) \cdot \frac{\partial f(x_1, x_2; \rho)}{\partial x_1}.$$

For later use, we define

$$h(x_1, x_2; \rho) = \phi(x_1) \cdot \Phi\left(\frac{\rho x_1 - x_2}{\sqrt{1-\rho^2}}\right),$$

where $\Phi : \mathbb{R} \to [0,1]$ is the standard normal cdf. Note that

$$\int_{\alpha_1}^{\infty} f(x_1, \alpha_2; \rho) dx_1 = \phi(\alpha_2) \cdot \int_{\alpha_1}^{\infty} \frac{1}{\sqrt{1-\rho^2}} \phi\left(\frac{x_1 - \rho\alpha_2}{\sqrt{1-\rho^2}}\right) dx_1;$$

because $\frac{1}{\sqrt{1-\rho^2}}\phi\left(\frac{x_1-\rho\alpha_2}{\sqrt{1-\rho^2}}\right)$ is the density of a normally distributed variable with mean $\rho\alpha_2$ and variance $(1-\rho^2)$ evaluated at x_1, the integral above can be written as

$$\int_{\alpha_1}^{\infty} f(x_1, \alpha_2; \rho) dx_1 = \phi(\alpha_2) \cdot \Phi\left(\frac{\rho\alpha_2 - \alpha_1}{\sqrt{1-\rho^2}}\right) = h(\alpha_2, \alpha_1; \rho).$$

By symmetry, we also have

$$\int_{\alpha_2}^{\infty} f(\alpha_1, x_2; \rho) dx_2 = h(\alpha_1, \alpha_2; \rho).$$

Finally, we define the cdf $F(\cdot, \cdot; \rho)$ of (X_1, X_2) as

$$F(\alpha_1, \alpha_2; \rho) = \mathbb{P}(X_1 \leq \alpha_1,\ X_2 \leq \alpha_2) = \int_{-\infty}^{\alpha_1} \int_{-\infty}^{\alpha_2} f(x_1, x_2; \rho) dx_2 dx_1.$$

We first compute the joint moment $\mathbb{E}\left[X_1 X_2 \cdot I_{\{X_1 \geq \alpha_1,\ X_2 \geq \alpha_2\}}\right]$. By definition,

$$\begin{aligned}\mathbb{E}\left[X_1 X_2 \cdot I_{\{X_1 \geq \alpha_1,\ X_2 \geq \alpha_2\}}\right] &= \int_{\alpha_1}^{\infty} \int_{\alpha_2}^{\infty} x_1 x_2 \cdot f(x_1, x_2; \rho) dx_2 dx_1 \\ &= \left(1-\rho^2\right)^{-\frac{1}{2}} \cdot \int_{\alpha_1}^{\infty} x_1 \cdot \phi(x_1) \left(\int_{\alpha_2}^{\infty} x_2 \cdot \phi\left(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}}\right) dx_2\right) dx_1.\end{aligned}$$

For any fixed $x_1 \in \mathbb{R}$, by a change of variables we have

$$\begin{aligned}\int_{\alpha_2}^{\infty} x_2 \cdot \phi\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right) dx_2 &= \sqrt{1-\rho^2} \cdot \int_{\frac{\alpha_2-\rho x_1}{\sqrt{1-\rho^2}}}^{\infty} \left(\sqrt{1-\rho^2}\cdot z + \rho x_1\right) \cdot \phi(z) dz \\ &= (1-\rho^2) \cdot \int_{\frac{\alpha_2-\rho x_1}{\sqrt{1-\rho^2}}}^{\infty} z \cdot \phi(z) dz + \rho\left(1-\rho^2\right)^{\frac{1}{2}} x_1 \cdot \int_{\frac{\alpha_2-\rho x_1}{\sqrt{1-\rho^2}}}^{\infty} \phi(z) dz \\ &= \left(1-\rho^2\right) \cdot \phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) + \rho\left(1-\rho^2\right)^{\frac{1}{2}} x_1 \cdot \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right).\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}\left[X_1 X_2 \cdot I_{\{X_1 \geq \alpha_1,\, X_2 \geq \alpha_2\}}\right] \\ &= \left(1-\rho^2\right)^{\frac{1}{2}} \cdot \int_{\alpha_1}^{\infty} x_1 \cdot \phi(x_1) \phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) dx_1 + \rho \cdot \int_{\alpha_1}^{\infty} x_1^2 \cdot \phi(x_1) \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) dx_1 \\ &= \left(1-\rho^2\right) \cdot \int_{\alpha_1}^{\infty} x_1 \cdot f(x_1, \alpha_2; \rho) dx_1 + \rho \cdot \int_{\alpha_1}^{\infty} x_1^2 \cdot \phi(x_1) \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) dx_1.\end{aligned}$$

Inspecting the first term, we have

$$\begin{aligned}\int_{\alpha_1}^{\infty} x_1 \cdot f(x_1, x_2; \rho) dx_1 &= \rho\alpha_2 \cdot \int_{\alpha_1}^{\infty} f(x_1, \alpha_2; \rho) dx_1 - (1-\rho^2) \cdot [f(x_1, \alpha_2; \rho)]_{\alpha_1}^{\infty} \\ &= \rho\alpha_2 \cdot h(\alpha_2, \alpha_1; \rho) + (1-\rho^2) \cdot f(\alpha_1, \alpha_2; \rho).\end{aligned}$$

On the other hand, because

$$\begin{aligned}\frac{\partial}{\partial x_1}\left(-x_1\phi(x_1) \cdot \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right)\right) &= -\phi(x_1) \cdot \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) \\ &\quad + x_1^2\phi(x_1) \cdot \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) - \frac{\rho}{\sqrt{1-\rho^2}} \cdot x_1\phi(x_1) \cdot \phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right),\end{aligned}$$

integration by parts shows us that the second term becomes

$$\begin{aligned}&\int_{\alpha_1}^{\infty} x_1^2 \cdot \phi(x_1) \Phi\left(\frac{\rho x_1-\alpha_2}{\sqrt{1-\rho^2}}\right) dx_1 \\ &= \alpha_1 \cdot h(\alpha_1, \alpha_2; \rho) + \int_{\alpha_1}^{\infty} h(x_1, \alpha_2; \rho) dx_1 + \rho \cdot \int_{\alpha_1}^{\infty} x_1 \cdot f(x_1, \alpha_2; \rho) dx_1.\end{aligned}$$

It follows that

$$\mathbb{E}\left[X_1X_2\cdot I_{\{X_1\geq\alpha_1,\,X_2\geq\alpha_2\}}\right]$$

$$=\int_{\alpha_1}^{\infty}x_1\cdot f(x_1,\alpha_2;\rho)dx_1+\rho\alpha_1\cdot h(\alpha_1,\alpha_2;\rho)+\rho\cdot\int_{\alpha_1}^{\infty}h(x_1,\alpha_2;\rho)dx_1$$

$$=\rho\alpha_2\cdot h(\alpha_2,\alpha_1;\rho)+(1-\rho^2)\cdot f(\alpha_1,\alpha_2;\rho)+\rho\alpha_1\cdot h(\alpha_1,\alpha_2;\rho)+\rho\cdot\int_{\alpha_1}^{\infty}h(x_1,\alpha_2;\rho)dx_1.$$

Finally, by a linear change of variables,

$$\begin{aligned}\int_{\alpha_1}^{\infty}h(x_1,\alpha_2;\rho)dx_1&=\int_{\alpha_1}^{\infty}\int_{\alpha_2}^{\infty}f(x_1,x_2;\rho)dx_2dx_1\\&=\int_{-\infty}^{-\alpha_1}\int_{-\infty}^{-\alpha_2}f(-x_1,-x_2;\rho)dx_2dx_1\\&=\int_{-\infty}^{-\alpha_1}\int_{-\infty}^{-\alpha_2}f(x_1,x_2;\rho)dx_2dx_1=F(-\alpha_1,-\alpha_2;\rho),\end{aligned}$$

so we have

$$\mathbb{E}\left[X_1X_2\cdot I_{\{X_1\geq\alpha_1,\,X_2\geq\alpha_2\}}\right]$$

$$=\rho\left[\alpha_2\cdot h(\alpha_2,\alpha_1;\rho)+\alpha_1\cdot h(\alpha_1,\alpha_2;\rho)+F(-\alpha_1,-\alpha_2;\rho)\right]+(1-\rho^2)\cdot f(\alpha_1,\alpha_2;\rho).$$

We now move onto the moment $\mathbb{E}\left[X_1\cdot I_{\{X_1\geq\alpha_1,\,X_2<\alpha_2\}}\right]$. By definition,

$$\begin{aligned}\mathbb{E}\left[X_1\cdot I_{\{X_1\geq\alpha_1,\,X_2<\alpha_2\}}\right]&=\int_{\alpha_1}^{\infty}\int_{-\infty}^{\alpha_2}x_1\cdot f(x_1,x_2;\rho)dx_2dx_1\\&=\int_{\alpha_1}^{\infty}x_1\cdot\phi(x_1)\left(\int_{-\infty}^{\alpha_2}\frac{1}{\sqrt{1-\rho^2}}\phi\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)dx_2\right)dx_1.\end{aligned}$$

For any given $x_1\in\mathbb{R}$,

$$\int_{-\infty}^{\alpha_2}\frac{1}{\sqrt{1-\rho^2}}\phi\left(\frac{x_2-\rho x_1}{\sqrt{1-\rho^2}}\right)dx_2$$

is equivalent to $\mathbb{P}(Z\leq\alpha_2)$, where $Z\sim\mathcal{N}\left[\rho x_1,\,1-\rho^2\right]$, so it follows that

$$\mathbb{E}\left[X_1\cdot I_{\{X_1\geq\alpha_1,\,X_2<\alpha_2\}}\right]=\int_{\alpha_1}^{\infty}x_1\cdot\phi(x_1)\Phi\left(\frac{\alpha_2-\rho x_1}{\sqrt{1-\rho^2}}\right)dx_1.$$

Integration by parts now reveals that

$$\begin{aligned}\mathbb{E}\left[X_1 \cdot I_{\{X_1 \geq \alpha_1,\, X_2 < \alpha_2\}}\right] &= \phi(\alpha_1) \cdot \Phi\left(\frac{\alpha_2 - \rho\alpha_1}{\sqrt{1-\rho^2}}\right) - \frac{\rho}{\sqrt{1-\rho^2}} \cdot \int_{\alpha_1}^{\infty} \phi(x_1)\phi\left(\frac{\alpha_2 - \rho x_1}{\sqrt{1-\rho^2}}\right) dx_1 \\ &= h(-\alpha_1, -\alpha_2; \rho) - \rho \cdot \int_{\alpha_1}^{\infty} f(x_1, \alpha_2; \rho) dx_1 \\ &= h(-\alpha_1, -\alpha_2; \rho) - \rho \cdot h(\alpha_2, \alpha_1; \rho).\end{aligned}$$

By symmetry,

$$\mathbb{E}\left[X_2 \cdot I_{\{X_1 < \alpha_1,\, X_2 \geq \alpha_2\}}\right] = h(-\alpha_2, -\alpha_1; \rho) - \rho \cdot h(\alpha_1, \alpha_2; \rho).$$

Therefore, defining the truncated normal variables

$$Y_1 = \max(X_1, \alpha_1) \quad \text{and} \quad Y_2 = \max(X_2, \alpha_2),$$

we have

$$\begin{aligned}\mathbb{E}[Y_1 Y_2] &= \mathbb{E}\left[X_1 X_2 \cdot I_{\{X_1 \geq \alpha_1,\, X_2 \geq \alpha_2\}}\right] + \alpha_2 \cdot \mathbb{E}\left[X_1 \cdot I_{\{X_1 \geq \alpha_1,\, X_2 < \alpha_2\}}\right] \\ &\quad + \alpha_1 \cdot \mathbb{E}\left[X_2 \cdot I_{\{X_1 < \alpha_1,\, X_2 \geq \alpha_2\}}\right] + \alpha_1\alpha_2 \cdot \mathbb{P}(X_1 < \alpha_1,\, X_2 < \alpha_2) \\ &= \rho\left[\alpha_2 \cdot h(\alpha_2, \alpha_1; \rho) + \alpha_1 \cdot h(\alpha_1, \alpha_2; \rho) + F(-\alpha_1, -\alpha_2; \rho)\right] + (1-\rho^2) \cdot f(\alpha_1, \alpha_2; \rho) \\ &\quad + \alpha_2 \cdot \left[h(-\alpha_1, -\alpha_2; \rho) - \rho \cdot h(\alpha_2, \alpha_1; \rho)\right] + \alpha_1 \cdot \left[h(-\alpha_2, -\alpha_1; \rho) - \rho \cdot h(\alpha_1, \alpha_2; \rho)\right] \\ &\quad + \alpha_1\alpha_2 \cdot F(\alpha_1, \alpha_2; \rho) \\ &= \rho \cdot F(-\alpha_1, -\alpha_2; \rho) + (1-\rho^2) \cdot f(\alpha_1, \alpha_2; \rho) + \alpha_2 \cdot h(-\alpha_1, -\alpha_2; \rho) \\ &\quad + \alpha_1 \cdot h(-\alpha_2, -\alpha_1; \rho) + \alpha_1\alpha_2 \cdot F(\alpha_1, \alpha_2; \rho).\end{aligned}$$

K Gibbs Sampling Algorithm for FS-ZLB Model

The measurement and transition equations of our model are given as follows:

$$r_t = \beta_{sr}' f_t + \sigma_u \cdot u_t$$

$$\mathcal{Y}_t = \mathcal{A}_{s_t} + \mathcal{B} f_t + \sigma_e \cdot e_t$$

$$f_t = G_{s_t}^{\mathbb{P}} f_{t-1} + \Omega_{s_t} \cdot v_t^{\mathbb{P}}.$$

We follow Hamilton and Wu (2012) and the RY specification of case **P** in Joslin, Singleton, and Zhu (2011) and assume that yields of three maturities, namely the federal funds rate, the 2 year and the 10 year yields, are observed without error. These three yields are now related to the factors as

$$\begin{pmatrix} r_t \\ Y_t(24) \\ Y_t(120) \end{pmatrix} = \begin{pmatrix} 0 \\ a_{s_t}(24)/24 \\ a_{s_t}(120)/120 \end{pmatrix} + \begin{pmatrix} \beta_{sr}' \\ b(24)'/24 \\ b(120)'/120 \end{pmatrix} f_t.$$

Here, since the federal funds rate, being an overnight rate, is close to a zero maturity yield, we approximate the factor loadings β_{sr} as $(1,0,0)'$. We essentially identify the level factor with the short rate. The first element in $b(\tau)'/\tau$ equals 1 for any maturity τ, so the above equation can be reformulated in terms of yield spreads as

$$\underbrace{\begin{pmatrix} Y_t(24) - r_t \\ Y_t(120) - r_t \end{pmatrix}}_{\mathcal{YS}_t^{(1)}} = \underbrace{\begin{pmatrix} a_{s_t}(24)/24 \\ a_{s_t}(120)/120 \end{pmatrix}}_{\mathcal{A}_{s_t}^{(1)}} + \underbrace{\begin{pmatrix} b(24)'/24 \\ b(120)'/120 \end{pmatrix}}_{\mathcal{B}^{(1)}} f_t^{SC},$$

where f_t^{SC} collects the slope and curvature factors S_t, C_t. Inverting this expression shows us that the slope and curvature factors are given as affine functions of the yield spreads $\mathcal{YS}_t^{(1)}$:

$$f_t^{SC} = -\mathcal{B}^{(1)-1}\mathcal{A}_{s_t}^{(1)} + \mathcal{B}^{(1)-1}\mathcal{YS}_t^{(1)}.$$

The rest of the yield spreads are collected as

$$\underbrace{\begin{pmatrix} Y_t(\tau_3) - r_t \\ \vdots \\ Y_t(\tau_m) - r_t \end{pmatrix}}_{\mathcal{YS}_t^{(2)}} = \underbrace{\begin{pmatrix} a_{s_t}(\tau_3)/\tau_3 \\ \vdots \\ a_{s_t}(\tau_m)/\tau_m \end{pmatrix}}_{\mathcal{A}_{s_t}^{(2)}} + \underbrace{\begin{pmatrix} b(\tau_3)'/\tau_3 \\ \vdots \\ b(\tau_m)'/\tau_m \end{pmatrix}}_{\mathcal{B}^{(2)}} f_t^{SC} + \sigma_e \cdot e_t,$$

where e_t is now an $(m-3)$-dimensional random vector of standard normally distribued measurement errors.

Following Hamilton and Wu (2012), we can reexpress the state-space model as a restricted Gaussian VAR. First, note that the factor dynamics are comprised of the regime-dependent level factor dynamics, and the regime-independent dynamics of the rest of the factors:

$$L_t = \bar{r}_{s_t} + \rho_{s_t} L_{t-1} + g_{s_t}^{\mathbb{P}\prime} f_{t-1}^{SC} + \omega_{s_t} \cdot v_{1t}^{\mathbb{P}}$$

$$f_t^{SC} = \tilde{G} f_{t-1}^{SC} + \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}},$$

where $v_t^{\mathbb{P}SC}$ collects the last two elements of the vector of risk factors $v_{SC,t}^{\mathbb{P}}$. Since f_t^{SC} is an affine function of the yield spreads $\mathcal{YS}_t^{(1)}$, we can see that

$$\begin{aligned}
\mathcal{YS}_t^{(1)} &= \mathcal{A}_{s_t}^{(1)} + \mathcal{B}^{(1)} \left(\tilde{G} f_{t-1}^{SC} + \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}} \right) \\
&= \mathcal{A}_{s_t}^{(1)} + \mathcal{B}^{(1)} \tilde{G} \left(-\mathcal{B}^{(1)-1} \mathcal{A}_{s_{t-1}}^{(1)} + \mathcal{B}^{(1)-1} \mathcal{YS}_{t-1}^{(1)} \right) + \mathcal{B}^{(1)} \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}} \\
&= \left(\mathcal{A}_{s_t}^{(1)} - \mathcal{B}^{(1)} \tilde{G} \mathcal{B}^{(1)-1} \cdot \mathcal{A}_{s_{t-1}}^{(1)} \right) + \mathcal{B}^{(1)} \tilde{G} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_{t-1}^{(1)} + \mathcal{B}^{(1)} \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{YS}_t^{(2)} &= \mathcal{A}_{s_t}^{(2)} + \mathcal{B}^{(2)} f_t^{SC} + \sigma_e \cdot e_t \\
&= \left(\mathcal{A}_{s_t}^{(2)} - \mathcal{B}^{(2)} \mathcal{B}^{(1)-1} \mathcal{A}_{s_t}^{(1)} \right) + \mathcal{B}^{(2)} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_t^{(1)} + \sigma_e \cdot e_t.
\end{aligned}$$

Furthermore, identifying the level factor with the short rate and replacing f_{t-1}^{SC} with the affine function of $\mathcal{YS}_{t-1}^{(1)}$ allows us to reformulate the level factor dynamics as

$$r_t = \left(\bar{r}_{s_t} - g_{s_t}^{\mathbb{P}\prime} \mathcal{B}^{(1)-1} \mathcal{A}_{s_{t-1}}^{(1)} \right) + \rho_{s_t} r_{t-1} + g_{s_t}^{\mathbb{P}\prime} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_{t-1}^{(1)} + \omega_{s_t} \cdot v_{1t}^{\mathbb{P}}.$$

The three equations that characterize the model can be written as

$$r_t = \left(\bar{r}_{s_t} - g_{s_t}^{\mathbb{P}\prime} \mathcal{B}^{(1)-1} \mathcal{A}_{s_{t-1}}^{(1)} \right) + \rho_{s_t} r_{t-1} + g_{s_t}^{\mathbb{P}\prime} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_{t-1}^{(1)} + \omega_{s_t} \cdot v_{1t}^{\mathbb{P}} \tag{5.75}$$

$$\mathcal{YS}_t^{(1)} = \left(\mathcal{A}_{s_t}^{(1)} - \mathcal{B}^{(1)} \tilde{G} \mathcal{B}^{(1)-1} \cdot \mathcal{A}_{s_{t-1}}^{(1)} \right) + \mathcal{B}^{(1)} \tilde{G} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_{t-1}^{(1)} + \mathcal{B}^{(1)} \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}} \tag{5.76}$$

$$\mathcal{YS}_t^{(2)} = \left(\mathcal{A}_{s_t}^{(2)} - \mathcal{B}^{(2)} \mathcal{B}^{(1)-1} \mathcal{A}_{s_t}^{(1)} \right) + \mathcal{B}^{(2)} \mathcal{B}^{(1)-1} \cdot \mathcal{YS}_t^{(1)} + \sigma_e \cdot e_t. \tag{5.77}$$

The parameters to be estimated are

$$\theta = \{g_{s_t}^{\mathbb{P}}, \omega_{s_t}, \tilde{G}^{\mathbb{P}}, \Omega_{22}, \kappa^{\mathbb{Q}}, K_{s_t}^{\mathbb{Q}}, \sigma_e, P\}$$

and the regime process $S = \{s_t\}_{1 \leq t \leq T}$.

Estimating the $\mathbb{Q}$-parameter $K_{s_t}^{\mathbb{Q}}$ has traditionally proven difficult, not least because the separate estimation of $K_{s_t}^{\mathbb{Q}}$ and Ω_{s_t}, P means that the latter enter into the model in a non-linear fashion. To circumvent this difficulty, we use the identity

$$\mathcal{A}_{s_t} = -\frac{1}{2} \underbrace{\begin{pmatrix} \frac{1}{\tau_1} \sum_{i=1}^{\tau_1 - 1} b(\tau_1 - i)' \left(\sum_{j=1}^{N} \Omega_j \cdot P_{s_t, j} \right) b(\tau_1 - i) \\ \vdots \\ \frac{1}{\tau_m} \sum_{i=1}^{\tau_m - 1} b(\tau_m - i)' \left(\sum_{j=1}^{N} \Omega_j \cdot P_{s_t, j} \right) b(\tau_m - i) \end{pmatrix}}_{c_{0,s_t}} + \underbrace{\begin{pmatrix} \frac{1}{\tau_1} \sum_{i=1}^{\tau_1 - 1} b(\tau_1 - i) \\ \vdots \\ \frac{1}{\tau_m} \sum_{i=1}^{\tau_m - 1} b(\tau_m - i) \end{pmatrix}}_{c_1} K^{\mathbb{Q}}.$$

$K_{s_t}^{\mathbb{Q}}$ is a function of the intercept term $\mathcal{A}_{s_t}$ and the paramters $\kappa^{\mathbb{Q}}, \Omega_{s_t}$ and P:

$$K_{s_t}^{\mathbb{Q}} = \left(c_1' c_1 \right)^{-1} c_1' \left(\mathcal{A}_{s_t} + \frac{1}{2} c_{0,s_t} \right)$$

In other words, an equivalent form of estimation is to estimate the parameters

$$\theta = \{g_{s_t}^{\mathbb{P}}, \omega_{s_t}, \tilde{G}^{\mathbb{P}}, \Omega_{22}, \kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}, \sigma_e, P\}.$$

This proves much easier, because, as we will see below, $\mathcal{A}_{s_t}$ serves as the intercept term of the reduced form version of the model.

Collect the observed variables as

$$X_t = \begin{pmatrix} r_t \\ \mathcal{YS}_t^{(1)} \\ \mathcal{YS}_t^{(2)} \end{pmatrix}.$$

We first discuss how to sample the parameters aside from $\kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}$ and the regime S. The joint likelihood of the data X and the regime S is given as

$$\begin{aligned} l(X, S \mid \theta) &= \prod_{t=1}^{T} l(X_t, s_t \mid \mathcal{F}_{t-1}, \theta) \\ &= \prod_{t=1}^{T} l(X_t \mid s_t, \mathcal{F}_{t-1}, \theta) \cdot l(s_t \mid \mathcal{F}_{t-1}, \theta) \\ &= \prod_{t=1}^{T} l(X_t \mid s_t, \mathcal{F}_{t-1}, \theta) \cdot P_{s_{t-1}, s_t}, \end{aligned}$$

where $\mathcal{F}_t$ is the σ-algebra generated by the history of X_t and s_t up to time t. We can further decompose each $l(X_t \mid s_t, \mathcal{F}_{t-1}, \theta)$ as

$$\begin{aligned}
l(X_t \mid s_t, \mathcal{F}_{t-1}, \theta) &= l(r_t, \mathcal{YS}_t^{(1)}, \mathcal{YS}_t^{(2)} \mid s_t, \mathcal{F}_{t-1}, \theta) \\
&= l(r_t \mid \mathcal{YS}_t, s_t, \mathcal{F}_{t-1}, \theta) \cdot l(\mathcal{YS}_t^{(2)} \mid \mathcal{YS}_t^{(1)}, s_t, \mathcal{F}_{t-1}, \theta) \cdot l(\mathcal{YS}_t^{(1)} \mid s_t, \mathcal{F}_{t-1}, \theta) \\
&= l(r_t \mid s_t, r_{t-1}, \mathcal{YS}_{t-1}^{(1)}; \{\kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}^{(1)}, \bar{r}_{s_t}, \rho_{s_t}, g_{s_t}^{\mathbb{P}}, \omega_{s_t}^2\}) \\
&\quad \times l(\mathcal{YS}_t^{(2)} \mid s_t, \mathcal{YS}_t^{(1)}; \{\kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}, \sigma_e\}) \\
&\quad \times l(\mathcal{YS}_t^{(1)} \mid s_t, \mathcal{YS}_{t-1}^{(1)}; \{\kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}^{(1)}, \tilde{G}, \Omega_{22}\}).
\end{aligned}$$

Aside from $\kappa^{\mathbb{Q}}$ and $K^{\mathbb{Q}}$, each of the parameters appears only in one likelihood term, and thus only in one of the equations from (5.75) to (5.77). Since each of them are Gaussian regressions, we can either use the Normal-Normal or Inverse Gamma- Inverse Gamma (alternatively, Inverse Wishart- Inverse Wishart) update to recover the full conditional distribution of every one of them. We detail each step below:

Block 1: Sampling $\rho_{s_t}, g_{s_t}^{\mathbb{P}}, \omega_{s_t}^2$

The full conditional distribution of ρ_2 under the prior

$$\rho_2 \sim \mathcal{N}[\bar{\rho}_0, V_{\rho,0}]$$

is given as

$$\rho_2 \mid X, S, \theta \setminus \{\rho_2\} \sim \mathcal{N}[\bar{\rho}_1, V_{\rho,1}]$$

by the Normal-Normal update, where

$$V_{\rho,1} = \left[\frac{1}{V_{\rho,0}} + \frac{1}{\omega_2^2} \sum_{1 \le t \le T, s_t = 2} r_{t-1}^2 \right]^{-1}$$

$$\bar{\rho}_1 = V_{\rho,1} \left[\frac{1}{\omega_2^2} \sum_{1 \le t \le T, s_t = 2} (r_t - \bar{r}_2)\, r_{t-1} + \frac{\bar{\rho}_0}{V_{\rho,0}} \right].$$

The full conditional distribution of $g_1^{\mathbb{P}}$ under the prior

$$g_1^{\mathbb{P}} \sim \mathcal{N}[\bar{g}_0, V_{g,0}]$$

is given as

$$g_1^{\mathbb{P}} \mid Y, S, \theta \setminus \{g_1^{\mathbb{P}}\} \sim \mathcal{N}\left[\bar{g}_1, V_{g,1}\right]$$

by the Normal-Normal update, where

$$V_{g,1} = \left[V_{g,0}^{-1} + \frac{1}{\omega_1^2} \sum_{1 \leq t \leq T, s_t = 1} f_{t-1}^{SC} f_{t-1}^{SC\prime}\right]^{-1}$$

$$\bar{g}_1 = V_{g,1}\left[\frac{1}{\omega_1^2} \sum_{1 \leq t \leq T, s_t = 1} f_{t-1}^{SC} \Delta r_t + V_{g,0}^{-1} \bar{g}_0\right].$$

For any $i = 1, 2$, the full conditional distribution of ω_i^2 under the prior

$$\omega_i^2 \sim \mathcal{IG}\left[\frac{a_{i,0}}{2}, \frac{d_{i,0}}{2}\right]$$

is given as

$$\omega_i^2 \mid X, S, \theta \setminus \{\omega_i^2\} \sim \mathcal{IG}\left[\frac{a_{i,1}}{2}, \frac{d_{i,1}}{2}\right]$$

by the Inverse Gamma- Inverse Gamma update, where

$$\begin{aligned} a_{i,1} &= a_{i,0} + \#\{1 \leq t \leq T \mid s_t = i\} \\ d_{i,1} &= d_{i,0} + \sum_{1 \leq t \leq T, s_t = i} \left(r_t - \bar{r}_i - \rho_i r_{t-1} - g_i^{\mathbb{P}\prime} f_{t-1}^{SC}\right)^2. \end{aligned}$$

Block 2: Sampling $\tilde{G}, \Omega_{22}$

Define $\gamma = \text{vec}\left(\tilde{G}^{\mathbb{P}\prime}\right)$. The likelihood represented by equation (5.76) is equivalent to the likelihood of the equation

$$f_t^{SC} = \tilde{G} f_{t-1}^{SC} + \Omega_{22}^{\frac{1}{2}} \cdot v_{SC,t}^{\mathbb{P}}.$$

Thus, the full conditional distribution of γ under the prior

$$\gamma \sim \mathcal{N}\left[\bar{\gamma}_0, V_{\gamma,0}\right]$$

is given as

$$\gamma \mid X, S, \theta \setminus \{\gamma\} \sim \mathcal{N}\left[\bar{\gamma}_1, V_{\gamma,1}\right]$$

by the Normal-Normal update, where

$$V_{\gamma,1} = \left[V_{\gamma,0}^{-1} + \left(\Omega_{22}^{-1} \bigotimes \sum_{t=1}^{T} f_{t-1}^{SC} f_{t-1}^{SC\prime}\right)\right]^{-1}$$

$$\bar{\gamma}_1 = V_{\gamma,1}\left[\left(\Omega_{22}^{-1} \bigotimes I_2\right) \cdot \text{vec}\left(\sum_{t=1}^{T} f_{t-1}^{SC} f_{t}^{SC\prime}\right) + V_{\gamma,0}^{-1}\bar{\gamma}_0\right].$$

The full conditional distribution of Ω_{22} under the prior

$$\Omega_{22} \sim \mathcal{IW}\left[v_0, \Psi_0\right]$$

is given as

$$\Omega_{22} \mid X, S, \theta \setminus \{\Omega_{22}\} \sim \mathcal{IW}\left[v_1, \Psi_1\right]$$

by the Inverse Wishart- Inverse Wishart update, where

$$v_1 = v_0 + T$$

$$\Psi_1 = \Psi_0 + \sum_{t=1}^{T}\left(f_t^{SC} - \tilde{G}^{\mathbb{P}} f_{t-1}^{SC}\right)\left(f_t^{SC} - \tilde{G}^{\mathbb{P}} f_{t-1}^{SC}\right)'.$$

Block 3: Sampling σ_e

The likelihood represented by equation (5.77) is equivalent to that represented by

$$\mathcal{YS}_t^{(2)} = \mathcal{A}_{s_t}^{(2)} + \mathcal{B}^{(2)} f_t^{SC} + \sigma_e \cdot e_t.$$

Therefore, the full conditional distribution for σ_e^2 under its prior

$$\sigma_e^2 \sim \mathcal{IG}\left[\frac{a_0}{2}, \frac{d_0}{2}\right]$$

is given as

$$\sigma_e^2 \mid X, S, \theta \setminus \{\sigma_e^2\} \sim \mathcal{IG}\left[\frac{a_1}{2}, \frac{d_1}{2}\right]$$

under the Inverse Gamma- Inverse Gamme update, where

$$\begin{aligned} a_1 &= a_0 + T(m-2) \\ d_1 &= d_0 + \sum_{t=1}^{T} \left| \mathcal{Y}\mathcal{S}_t^{(2)} - \mathcal{A}_{s_t}^{(2)} - \mathcal{B}^{(2)} f_t^{SC} \right|^2. \end{aligned}$$

Block 4: Sampling P

For any $i = 1, 2$, letting j denote the regime that is not regime i, full conditional distribution of P_{ii} under the prior

$$P_{ii} \sim beta\left[\alpha_{i,0}, \beta_{i,0}\right]$$

is given as

$$P_{ii} \mid X, S, \theta \setminus \{P_{ii}\} \sim beta\left[\alpha_{i,0} + n_{ii}, \beta_{i,0} + n_{ij}\right],$$

where n_{ii} is the number of times the regime changes from regime i to regime i and n_{ij} is the number of times the regime changes from regime i to regime j.

To estimate $\kappa^{\mathbb{Q}}$ and the regime process S, it is necessary to obtain a tractable expression for the likelihood of the data X. In addition, to estimate $\mathcal{A}_{s_t}$, we must collect the three equations characterizing the system together, since $\mathcal{A}_{s_t}^{(1)}$ appears in all three. To this end, we collect the three equations as

$$\begin{pmatrix} 1 & O_{1\times 2} & O_{1\times(m-2)} \\ O_{2\times 1} & I_2 & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & -\mathcal{B}^{(2)}\mathcal{B}^{(1)-1} & I_{m-2} \end{pmatrix} X_t$$

$$= \underbrace{\begin{pmatrix} \bar{r}_{s_t} \\ O_{m\times 1} \end{pmatrix}}_{\bar{R}_{s_t}} + \begin{pmatrix} -g_{s_t}^{\mathbb{P}\prime}\mathcal{B}^{(1)-1} & O_{1\times(m-2)} \\ -\mathcal{B}^{(1)}\tilde{G}\mathcal{B}^{(1)-1} & O_{2\times(m-2)} \\ O_{(m-2)\times 2} & O_{(m-2)\times(m-2)} \end{pmatrix} \underbrace{\begin{pmatrix} \mathcal{A}_{s_{t-1}}^{(1)} \\ \mathcal{A}_{s_{t-1}}^{(2)} \end{pmatrix}}_{\mathcal{A}_{s_{t-1}}} + \begin{pmatrix} O_{1\times 2} & O_{1\times(m-2)} \\ I_2 & O_{2\times(m-2)} \\ -\mathcal{B}^{(2)}\mathcal{B}^{(1)-1} & I_{m-2} \end{pmatrix} \underbrace{\begin{pmatrix} \mathcal{A}_{s_t}^{(1)} \\ \mathcal{A}_{s_t}^{(2)} \end{pmatrix}}_{\mathcal{A}_{s_t}}$$

$$+ \begin{pmatrix} \rho_{s_t} & g_{s_t}^{\mathbb{P}\prime}\mathcal{B}^{(1)-1} & O_{1\times(m-2)} \\ O_{2\times 1} & \mathcal{B}^{(1)}\tilde{G}\mathcal{B}^{(1)-1} & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & O_{(m-2)\times 2} & O_{(m-2)\times(m-2)} \end{pmatrix} X_{t-1} + \begin{pmatrix} \omega_{s_t} & O_{1\times 2} & O_{1\times(m-2)} \\ O_{2\times 1} & \mathcal{B}^{(1)}\Omega_{22}^{\frac{1}{2}} & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & O_{(m-2)\times 2} & \sigma_e \cdot I_{m-2} \end{pmatrix} \begin{pmatrix} v_{1t}^{\mathbb{P}} \\ v_{SC,t}^{\mathbb{P}} \\ e_t \end{pmatrix}.$$

Since

$$\begin{pmatrix} 1 & O_{1\times 2} & O_{1\times(m-2)} \\ O_{2\times 1} & I_2 & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & -\mathcal{B}^{(2)}\mathcal{B}^{(1)-1} & I_{m-2} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & O_{1\times 2} & O_{1\times(m-2)} \\ O_{2\times 1} & I_2 & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & \mathcal{B}^{(2)}\mathcal{B}^{(1)-1} & I_{m-2} \end{pmatrix},$$

defining

$$D_{1,s_t} = \begin{pmatrix} -g_{s_t}^{\mathbb{P}\prime}\mathcal{B}^{(1)-1} & O_{1\times(m-2)} \\ -\mathcal{B}^{(1)}\tilde{G}\mathcal{B}^{(1)-1} & O_{2\times(m-2)} \\ -\mathcal{B}^{(2)}\tilde{G}\mathcal{B}^{(1)-1} & O_{(m-2)\times(m-2)} \end{pmatrix}$$

$$E = \begin{pmatrix} O_{1\times m} \\ I_m \end{pmatrix}$$

$$C_{1,s_t,s_{t-1}} = \begin{pmatrix} D_{1,s_t}\cdot(2-s_{t-1}) & D_{1,s_t}\cdot(s_{t-1}-1) \end{pmatrix} + \begin{pmatrix} E\cdot(2-s_t) & E\cdot(s_t-1) \end{pmatrix}$$

the reduced form of this system is

$$X_t = \bar{R}_{s_t} + C_{1,s_t,s_{t-1}} \cdot \underbrace{\begin{pmatrix} \mathcal{A}_1 \\ \mathcal{A}_2 \end{pmatrix}}_{\mathcal{A}}$$

$$+ \underbrace{\begin{pmatrix} \rho_{s_t} & g_{s_t}^{\mathbb{P}\prime}\mathcal{B}^{(1)-1} & O_{1\times(m-2)} \\ O_{2\times 1} & \mathcal{B}^{(1)}\tilde{G}\mathcal{B}^{(1)-1} & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & \mathcal{B}^{(2)}\tilde{G}\mathcal{B}^{(1)-1} & O_{(m-2)\times(m-2)} \end{pmatrix}}_{\Phi_{s_t}} X_{t-1} + \underbrace{\begin{pmatrix} \omega_{s_t} & O_{1\times 2} & O_{1\times(m-2)} \\ O_{2\times 1} & \mathcal{B}^{(1)}\Omega_{22}^{\frac{1}{2}} & O_{2\times(m-2)} \\ O_{(m-2)\times 1} & \mathcal{B}^{(2)}\Omega_{22}^{\frac{1}{2}} & \sigma_e \cdot I_{m-2} \end{pmatrix}}_{\Sigma_{s_t}^{\frac{1}{2}}} \underbrace{\begin{pmatrix} v_{1t}^{\mathbb{P}} \\ v_{SC,t}^{\mathbb{P}} \\ e_t \end{pmatrix}}_{u_t}.$$

The time t likelihood given the time t and $t-1$ regimes now becomes

$$l(X_t \mid s_t, s_{t-1}, X_{t-1}, \theta) = \left(\frac{1}{2\pi}\right)^{\frac{m+1}{2}} |\Sigma_{s_t}|^{-\frac{1}{2}}$$

$$\times \exp\Big(-\frac{1}{2}\left(X_t - \bar{R}_{s_t} - C_{1,s_t,s_{t-1}}\mathcal{A} - \Phi_{s_t}X_{t-1}\right)' \Sigma_{s_t}^{-1}\left(X_t - \bar{R}_{s_t} - C_{1,s_t,s_{t-1}}\mathcal{A} - \Phi_{s_t}X_{t-1}\right)\Big),$$

and

$$\begin{aligned} l(X,S \mid \theta) &= \prod_{t=1}^{T} l(X_t \mid s_t, s_{t-1}, X_{t-1}, \theta) \cdot P_{s_{t-1},s_t} \\ &= \left(\frac{1}{2\pi}\right)^{\frac{T(m+1)}{2}} \left|\prod_{t=1}^{T}\Sigma_{s_t}\right|^{-\frac{1}{2}} \cdot \left(\prod_{t=1}^{T} P_{s_{t-1},s_t}\right) \\ &\quad \times \exp\left(-\frac{1}{2}\sum_{t=1}^{T}\left(X_t - \bar{R}_{s_t} - C_{1,s_t,s_{t-1}}\mathcal{A} - \Phi_{s_t}X_{t-1}\right)' \Sigma_{s_t}^{-1}\left(X_t - \bar{R}_{s_t} - C_{1,s_t,s_{t-1}}\mathcal{A} - \Phi_{s_t}X_{t-1}\right)\right) \end{aligned}$$

It follows that $\kappa^{\mathbb{Q}}, \mathcal{A}_{s_t}$ and the regime process S can be sampled as follows:

Block 5: Sampling $\kappa^{\mathbb{Q}}$ and $\mathcal{A}$

Since $\kappa^{\mathbb{Q}}$ is assumed to take finitely many values, we can just take the probability that $\kappa^{\mathbb{Q}}$ equals a certain value as proportional to the likelihood value under this value.

As for $\mathcal{A}$, we can use the fact that it serves as the intercept term in the regression

$$X_t = \bar{R}_{s_t} + C_{1,s_t,s_{t-1}}\mathcal{A} + \Phi_{s_t}X_{t-1} + \Sigma_{s_t}^{\frac{1}{2}} \cdot u_t.$$

It follows that the full conditional distribution of $\mathcal{A}$ under the prior

$$\mathcal{A} \sim \mathcal{N}\left[\bar{\mathcal{A}}_0, V_{\mathcal{A},0}\right]$$

is given as

$$\mathcal{A} \mid X, S, \theta \setminus \{\mathcal{A}\} \sim \mathcal{N}\left[\bar{\mathcal{A}}_1, V_{\mathcal{A},1}\right]$$

by the Normal-Normal update, where

$$V_{\mathcal{A},1} = \left[\sum_{t=1}^{T} C'_{1,s_t,s_{t-1}}\Sigma_{s_t}^{-1}C_{1,s_t,s_{t-1}} + V_{\mathcal{A},0}^{-1}\right]^{-1}$$

$$\bar{\mathcal{A}}_{i,1} = V_{\mathcal{A},1}\left[\sum_{t=1}^{T} C'_{1,s_t,s_{t-1}}\Sigma_{s_t}^{-1}\left(X_t - \bar{R}_{s_t} - \Phi_{s_t}X_{t-1}\right) + V_{\mathcal{A},0}^{-1}\bar{\mathcal{A}}_0\right].$$

Block 6: Sample S

It now remains to sample the regime process. This is done through the standard Hamilton filter combined with the Carter-Kohn backward recursion. By assumption, information on whether the economy is at the zero lower bound at a certain point in time is conveyed through the factors,

Let $\mathbf{X}_t$ be the σ-algebra generated by the history of r_t and ST_t up to time t, and define the quantities

$$\alpha_{t|s} = \begin{pmatrix} \mathbb{P}(s_t = 1 \mid \mathbf{X}_s) \\ \vdots \\ \mathbb{P}(s_t = N \mid \mathbf{X}_s) \end{pmatrix}$$

for any $0 \leq t, s \leq T$. Then, the predictive and filtered probabilities can be computed recursively, and the regime process sampled, as follows:

1) **Step 0: Loading Initial Values**
Prepare initial values by setting

$$\alpha_{0|0} = \begin{pmatrix} \frac{1-P_{22}}{2-P_{11}-P_{22}} \\ \\ \frac{1-P_{11}}{2-P_{11}-P_{22}} \end{pmatrix},$$

the stationary distribution of the Markov chain.

2) **Step 1: Predictive Probabilities**
Given $\alpha_{t-1|t-1}$, we can compute

$$\alpha_{t|t-1} = P' \cdot \alpha_{t-1|t-1}.$$

3) **Step 2: Filtered Probabilities**
For any $1 \leq i \leq N$,

$$\begin{aligned} \mathbb{P}(s_t = i, s_{t-1} = j \mid \mathbf{X}_t) &\propto l(X_t \mid s_t = i, s_{t-1} = j, X_{t-1}) \cdot \mathbb{P}(s_t = i, s_{t-1} = j \mid \mathbf{X}_{t-1}) \\ &= l(X_t \mid s_t = i, s_{t-1} = j, X_{t-1}) \cdot \mathbb{P}(s_t = i \mid s_{t-1} = j) \cdot \mathbb{P}(s_{t-1} = j \mid \mathbf{X}_{t-1}) \\ &= l(X_t \mid s_t = i, s_{t-1} = j, X_{t-1}) \cdot P_{ji} \cdot \alpha_{t-1|t-1,j}. \end{aligned}$$

Since

$$\mathbb{P}(s_t = i \mid \mathbf{X}_t) = \sum_{j=1}^{N} \mathbb{P}(s_t = i, s_{t-1} = j \mid \mathbf{X}_t),$$

it follows that the kernels of the filtered probabilities in $\alpha_{t|t}$ are given as

$$k_{t|t} = \begin{pmatrix} \sum_{j=1}^{N} l(X_t \mid s_t = 1, s_{t-1} = j, X_{t-1}) \cdot P_{j1} \cdot \alpha_{t-1|t-1,j} \\ \vdots \\ \sum_{j=1}^{N} l(X_t \mid s_t = N, s_{t-1} = j, X_{t-1}) \cdot P_{jN} \cdot \alpha_{t-1|t-1,j} \end{pmatrix},$$

and the filtered probabilities as

$$\alpha_{t|t} = \frac{1}{\iota' k_{t|t}} k_{t|t}.$$

4) **Step 3: Sampling Regimes**

Sample s_T according to the filtered probabilities $\alpha_{T|T}$. Having sampled s_{t+1}, calculate the probabilities $\alpha_{t|T}$ as

$$\alpha_{t|T} = \frac{1}{\sum_{i=1}^{N} P_{i,s_{t+1}} \cdot \alpha_{t|t,i}} \begin{pmatrix} P_{1,s_{t+1}} \cdot \alpha_{t|t,1} \\ \vdots \\ P_{N,s_{t+1}} \cdot \alpha_{t|t,N}. \end{pmatrix}$$

s_t is then sampled according to the probabilities $\alpha_{t|T}$.