A Bayesian Analysis of Affine Term Structure Models.

Sébastien Blais [a,1]

[a] *Bank of Canada.*
First draft: June 12, 2006
This draft: 20 September 2009

Abstract

Dynamic term structure models are no-arbitrage structural economic factor models. In empirical applications, one specifies a statistical model for observational errors in order to accommodate the fact that the economic model imposes equality restrictions on observables that do not exactly hold in practice. Because term structure models involve unidentified structural parameters, they require normalization. This paper investigates the empirical importance of error modeling and normalization for inference for affine term structure models. At the methodological level, I propose and implement a new MCMC algorithm for Gaussian affine term structure models in which latent factors are drawn together with some parameters as a single block.

Comparing two popular approaches to modeling pricing errors, my analysis reveals that residuals from latent factor models have lower cross-correlations and autocorrelations than residuals from models where proxying factors are recovered by inverting the pricing equations. Because the latter models are special cases of the former in which some pricing errors are identically zero, this result implies that restrictions on error variances affect inference for factor dynamics. In order to investigate this issue, I compare latent factor models with homoscedastic and heteroscedastic errors: introducing heteroscedasticity further reduces residual cross-correlations and autocorrelations. While residuals from these independent-error models are correlated, modeling correlated errors increases residual autocorrelations. I use informative priors in order to obtain residuals that are compatible with the error correlation model but have low autocorrelations. I also propose an informative prior distribution for the dispersion of error standard deviations, which allows me to control the level of residual heteroscedasticity.

With respect to normalization, I provide evidence that factors are weakly identified from discount bond prices. This implies that a poor normalization can yield parameter point estimators with undesirable finite-sample properties. In particular, the maximum likelihood estimator can be severely biased and asymptotic confidence intervals unreliable. In contrast, Bayesian inference for pricing errors is valid. I demonstrate that Dai and Singleton's (2000, Journal of Finance) "canonical representation" makes inference particularly sensitive to these problems and I propose alternative normalizations.

1 Introduction

The topic of this paper is inference for *Dynamic Term Structure Models* (DTSM) (See Dai and Singleton, 2003, for a review). A DTSM is a factor model for the stochastic discount factor (SDF). Because the final nominal value of risk-free discount bonds is known with certainty, their prices are completely determined by the SDF. The joint specification of the factor physical dynamics and the functional forms of the short rate and the SDF defines a particular DTSM. The model I consider here is a discrete-time version of the Gaussian constant-diffusion essentially-affine DTSM of Duffee (2002), in which the short rate and the log-SDF are affine functions of factors that evolve as a Gaussian first-order vector autoregressive process under both the risk-neutral and physical measures. Under these assumptions, a N-dimensional vector y_t of discount rates at time t is affine in a K-dimensional vector X_t of factors,

$$y_t = \mathbf{A}(\psi) + \mathbf{B}(\psi)' X_t \tag{1.1}$$

$$X_t = X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \tilde{\epsilon}_t, \qquad \tilde{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{1.2}$$

where $\mathbf{A}(\psi)$ and $\mathbf{B}(\psi)$ are functions of the model's structural parameter vector $\psi \in \Psi$, which includes $\kappa^{\mathbb{P}}$, $\theta^{\mathbb{P}}$ and Σ. From now on, I refer to this model as the *affine term structure model* (ATSM).

DTSMs are increasingly popular in economics and their applications are varied. They are parsimonious and theoretically consistent descriptions of the term structure of interest rates. They are used to improve macroeconomic forecasts (Ang and Piazzesi, 2003), to estimate monetary policy rules (Ang, Dong, and Piazzesi, 2007), to enrich new Keynesian models (Hördahl, Tristani, and Vestin, 2006; Bekaert, Cho, and Moreno, 2006; Dewachter and Lyrio, 2006)), and to estimate structural parameters, such as preference parameters (Garcia and Luger, 2007).

One can see econometric inference for ATSM's as a sequence of five steps. First, economic theory gives a deterministic structural relationship, here given by equations (1.1-1.2), between the dynamics of the short rate and the term structure of interest rates. Second, to accommodate the fact that the model imposes equality restrictions on observables that do not exactly hold in practice, an observational error model is specified. This paper considers an additive Gaussian observational error vector $\tilde{e}_t$,

$$y_t = \mathbf{A}(\psi) + \mathbf{B}(\psi)' X_t + \tilde{e}_t, \qquad \tilde{e}_t \sim \mathcal{N}(\mathbf{0}, \Omega). \tag{1.3}$$

Email address: sblais@sebastienblais.com.

[1] The latest version of this working paper is available at http://www.sebastienblais.com/research/batsm.pdf. I would like to thank René Garcia , William McCausland and Monika Piazzesi and for helpful comments and suggestions. Financial support from CDP Capital is gratefully acknowledged.

Then, the model requires normalization because it involves unidentified structural parameters. For example, affine models are invariant with respect to linear transformations of the factors: for any invertible matrix $\mathbf{M}$, there exists a function $g_{\mathbf{M}} : \Psi \to \Psi$ such that equations (1.3) and (1.2) can be equivalently written as

$$\begin{aligned} y_t &= \mathbf{A}(\psi) + \mathbf{B}(\psi)'\mathbf{M}^{-1}\mathbf{M}X_t + \tilde{e}_t, \\ &= \mathbf{A}\left(g_{\mathbf{M}}(\psi)\right) + \mathbf{B}\left(g_{\mathbf{M}}(\psi)\right) Z_t + \tilde{e}_t, \\ \mathbf{M}X_t &= \mathbf{M}X_{t-1} + \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}(\mathbf{M}\theta^{\mathbb{P}} - \mathbf{M}X_{t-1}) + \mathbf{M}\tilde{\epsilon}_t, \\ &= Z_t = Z_{t-1} + \tilde{\kappa}^{\mathbb{P}}(\tilde{\theta}^{\mathbb{P}} - Z_{t-1}) + \tilde{\eta}_t, \qquad \tilde{\eta}_t \sim \mathcal{N}(\mathbf{0}, \tilde{\Sigma}), \end{aligned}$$

where $Z_t = \mathbf{M}X_t$, $\tilde{\kappa}^{\mathbb{P}} = \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}$, $\tilde{\Sigma} = \mathbf{M}\Sigma\mathbf{M}'$ and $\tilde{\theta}^{\mathbb{P}} = \mathbf{M}\theta^{\mathbb{P}}$. Parameter vectors ϕ and $g_{\mathbf{M}}(\psi)$ are thus observationally equivalent and the unnormalized model is globally unidentified.

Fourth, an inferential method is selected to retrieve the information about the economic model contained in the data. Finally, one chooses a loss function and makes decisions, in the form of estimates, tests or forecasts. This paper investigates the empirical importance of the error modeling and normalization steps for inference for ATSMs.

Error modeling

There are at least two distinct approaches to specifying errors in observations. The most popular approach in the macroeconomics and financial economics literatures[2] follows Chen and Scott (1993) and assumes that a number of discount rates equal to the number of factors is observed without error. This choice, where the error covariance matrix Ω has rank $N-K$, is justified on computational grounds, as the factors can then be recovered by inverting the pricing equation. I refer to this approach as the ***proxy*** modeling approach because it uses a deterministic function of observables to proxy latent factors. The second, ***latent-factor*** modeling approach follows Chen and Scott (1995) and assumes that all discount rates are observed with error. The errors covariance matrix Ω has rank N under this approach, which is popular in the empirical finance literature[3]. There are many theoretical reasons to prefer modeling errors on all yields: there is no need to make an arbitrary choice of which rates are observed without error; discount rates are often not observed but rather approximated from quoted coupon bond yields; and dynamics are more flexible.

To the best of my knowledge, these error modeling approaches have not been compared empirically. In this paper, I address the trade-off between model flexibility and

[2] Some examples are: Dai and Singleton (2002); Ang and Piazzesi (2003); Duffee (2002); Cheridito et al. (2003); Duarte (2003).

[3] Examples are: Jegadeesh and Pennacchi (1996); Ball and Torous (1996); Babbs and Nowman (1999); Geyer and Pichler (1999); Lamoureux and Witte (2002); Ang et al. (2007).

computational ease. I find that the latent-factor approach yields residuals that are significantly less correlated and autocorrelated than residuals from the proxy approach. This highlights the role of error modeling in the decomposition of observable dynamics into common and idiosyncratic components: restrictions on the latter affects inference for the former. In addition, a relatively higher pricing error on the short rate suggests that it might not be the best of proxying factors. Because DTSMs build on assumptions about short-rate dynamics, assuming that the short rate is observed without error is common practice; my results indicate that this particular modeling choice is not inferentially innocuous.

One often looks at residuals, as I do in this paper, for evidence of misspecification of the economic model. Finding such misspecification may lead the econometrician to more general error covariance specifications. On the other hand, an arbitrary covariance matrix may lead to over-parameterization. Because latent-factor models essentially decompose the dynamics of the observables into common and idiosyncratic components, error covariance modeling also allows the econometrician to specify which characteristics of the observables the common latent factors should capture. For example, if errors are modeled as i.i.d. random variables, factors are required to capture the heteroscedasticity, correlation and persistence of observables. In contrast, if errors are independent but heteroscedastic, factors might be better able to capture correlations and persistence. Factors might yet better capture persistence if errors are correlated.

The proxy approach is a special case of the latent-factor approach corresponding to a rather strong restriction on the covariance matrix Ω in which elements are equal to zero. Other restrictions are also likely to affect the factor-error decomposition. Imposing homoscedasticity and independence are examples of such restrictions. In order to consider these restrictions individually, I factorize the covariance matrix into a correlation matrix and a diagonal matrix of precisions (the inverse of errors variances). In this paper, I propose priors that operationalize *soft* restrictions on the correlation and precision matrices. My empirical results show that using these priors for modeling mildly heteroscedastic and cross-correlated errors yields residuals with lower autocorrelations than strict homoscedastic or independent error models.

Normalization

I consider likelihood-based inference methods, which rely on a parametric statistical model.

Definition 1 *A **parametric statistical model** is a triplet $(\mathcal{Y}, \mathcal{F}, \Psi)$, where $\mathcal{Y}$ is the sample space, $\mathcal{F} \equiv \left\{ f(y|\,\psi) \middle| \, y \in \mathcal{Y},\, \psi \in \Psi \right\}$ is a set of parametric probability density functions on $\mathcal{Y}$, and Ψ is the parameter set. The **likelihood function** of the model is the function $l(\psi|\, y) = f(y|\,\psi)$.*

The likelihood of ATSMs is invariant with respect to parameter transformations corresponding to affine transformations of the factors.

Definition 2 *A function $f : \Psi \to \mathbb{R}$ is **invariant with respect a bijective transformation** $T : \Psi \to \Psi$ if $f(T(\psi)) = f(\psi)$.*

If $l(\psi|y)$ is invariant with respect to T on Ψ for all $y \in \mathcal{Y}$ then we say that $T(\psi)$ and ψ are **observationally equivalent**. We will also say that f is invariant with respect a set of bijective transformations $\mathcal{T}(\Psi)$ if it is invariant with respect to all of its elements. The notation $\mathcal{T}(\Psi)$ makes dependence on the set Ψ explicit: $\mathcal{T}(\Psi)$ is a set of bijections on Ψ. For example, for $\Psi' \subseteq \Psi$, $\mathcal{T}(\Psi') = \{T : \Psi' \to \Psi' | \, T \in \mathcal{T}(\Psi)\}$. I will omit this dependence and write $\mathcal{T}$ when this causes no confusion. The following example illustrates this definition.

Example 1 *Consider*

$$y = bx + e, \quad x \sim \mathcal{N}(0, \sigma^2), \quad e \sim \mathcal{N}(0, 1),$$

for $(b, \sigma^2) \in \Psi = \mathbb{R} \times (0, \infty)$. In that case the likelihood function satisfies $l(b, \sigma^2 | \, y) = l\left(|Db|, \sigma^2/D^2 \middle| \, y\right)$ for any $D \neq 0$, and it is therefore invariant with respect to

$$\begin{aligned}
\mathcal{T}_D(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_D(b, \sigma^2) = \left(Db, \sigma^2/D^2\right), \, D > 0\right\} \\
\mathcal{T}_S(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_S(b, \sigma^2) = \left(Sb, \sigma^2\right), \, |S| = 1\right\} \\
\mathcal{T}_{SD}(\Psi) &= \left\{T : \Psi \to \Psi \middle| \, T_{SD}(b, \sigma^2) = \left(SDb, \sigma^2/D^2\right), \, D > 0, |S| = 1\right\} \\
&= \left\{T : \Psi \to \Psi \middle| \, T(\psi) = T_S\left(T_D(\psi)\right)\right\}
\end{aligned}$$

The parameters b and σ^2 enter the likelihood function as the product $b^2\sigma^2$. Transformations in $\mathcal{T}_D$ correspond to changing the scale of the unobserved factor x and reflect the fact that $(Db)^2 \frac{\sigma^2}{D^2} = b^2\sigma^2$ for $D > 0$. Transformations in $\mathcal{T}_S$ correspond to reflections of x across $x = 0$, which change its sign.

That the likelihood of ATSMs is invariant with respect to parameter transformations corresponding to affine transformations of the factors has the following meaning: If $\psi \in \Psi$ denotes the K-factor ATSM's parameter vector, y the observed panel of discount rates and $f(y | \, \psi, X)$ the probability density function of y conditional on a panel of factors X, then for any K-dimensional vector $\mathbf{t}$ and any invertible $K \times K$ matrix $\mathbf{M}$, there exists a bijection $T_{\mathbf{tM}}(\psi) : \Psi \to \Psi$ such that $f(y | \, T_{\mathbf{tM}}(\psi), \mathbf{M}(X - \mathbf{t})) = f(y | \, \psi, X)$.

In general, one addresses transformation invariance by normalizing the model.

Definition 3 *A normalization is a parameter subset $\Psi^N \subseteq \Psi$.*

A normalization $\Psi^N \subseteq \Psi$ breaks invariance with respect to a set of bijections $\mathcal{T}(\Psi)$ if $\mathcal{T}\left(\Psi^N\right) = \mathcal{T}_{\mathcal{I}}$, where

$$\mathcal{T}_{\mathcal{I}} = \left\{ T : \Psi \to \Psi \middle| \, T(\psi) = \psi \right\} \tag{1.4}$$

is a singleton: the identity transformation. Note that $\Psi^N \subseteq \Psi \Rightarrow \mathcal{T}\left(\Psi^N\right) \subseteq \mathcal{T}\left(\Psi\right)$. Breaking invariance with respect to a set of bijective transformations $\mathcal{T}(\Psi)$ is thus considering a parameter subset $\Psi^N \subseteq \Psi$ that is small enough to ensure that the only invariant bijection on that subset is the identity transformation, $\mathcal{T}(\Psi^N) = \left\{ T : \Psi^N \to \Psi^N \middle| \, T(\psi) = \psi \right\}$.

Example 2 (Example 1, continued) *Consider the normalizations*

$$\begin{aligned} \Psi^b &= \{ \psi \in \Psi | \, b > 0 \} \\ \Psi^{\sigma_1^2} &= \left\{ \psi \in \Psi \middle| \, \sigma^2 = 1 \right\}. \end{aligned}$$

Scaling and reflection transformations on these normalizations are:

$$\begin{aligned} \mathcal{T}_S\left(\Psi^b\right) &= \left\{ T : \Psi^b \to \Psi^b \middle| \, T_S(b, \sigma^2) = \left(Sb, \sigma^2\right), \, S = 1 \right\} = \mathcal{T}_{\mathcal{I}}, \\ \mathcal{T}_S\left(\Psi^{\sigma_1^2}\right) &= \mathcal{T}_S\left(\Psi\right), \\ \mathcal{T}_D\left(\Psi^b\right) &= \mathcal{T}_D\left(\Psi\right), \\ \mathcal{T}_D\left(\Psi^{\sigma_1^2}\right) &= \left\{ T : \Psi^{\sigma_1^2} \to \Psi^{\sigma_1^2} \middle| \, T_D(b, \sigma^2) = \left(Db, \sigma^2/D^2\right), \, D = 1 \right\} = \mathcal{T}_{\mathcal{I}}, \\ \mathcal{T}_{SD}\left(\Psi^b \cap \Psi^{\sigma_1^2}\right) &= \left\{ T : \Psi^b \cap \Psi^{\sigma_1^2} \to \Psi^b \cap \Psi^{\sigma_1^2} \middle| \right. \\ &\qquad \left. T_S(b, \sigma^2) = \left(SDb, \sigma^2/D^2\right), \, S = 1, D = 1 \right\} = \mathcal{T}_{\mathcal{I}}. \end{aligned}$$

Normalization Ψ^b breaks invariance with respect to reflections of factors because T_S is not a bijection on Ψ^b for $S \neq 1$. Similarly, $\Psi^{\sigma_1^2}$ breaks invariance with respect to $\mathcal{T}_D$ because T_D is not a bijection on $\Psi^{\sigma_1^2}$ for $D \neq 1$. Thus, $\Psi^b \cap \Psi^{\sigma_1^2}$ breaks invariance with respect to $\mathcal{T}_{SD}$.

One could consider normalization of arbitrary form, but it is natural to restrict attention to intersections of half hyper-spaces and hyper-planes,

$$\Psi^N = \bigcap_{i=1}^{I} \{ \psi \in \Psi | \, \mathbf{g}_i' \psi > 0 \} \cap \bigcap_{j=1}^{J} \left\{ \psi \in \Psi \middle| \, \mathbf{h}_j' \psi = 0 \right\},$$

for some conformable real vectors $\mathbf{g}_1, \ldots, \mathbf{g}_I, \mathbf{h}_1, \ldots, \mathbf{h}_J$. For example, one would break invariance with respect to a set of $(I+1)!$ invariant transformations with a normalization consisting in the intersection of I half hyper-spaces. In contrast, the intersection of J half hyper-planes would break invariance with respect to a set of invariant transformations that is equinumerous to $\mathbb{R}^J$ (*i.e* $\mathcal{T}$ has the same cardinality as $\mathbb{R}^J$).

Example 3 (Example 1, continued) *There are* 2! *transformations in $\mathcal{T}_S$ and the half-space $\{ \psi \in \Psi | \, b > 0 \}$ breaks invariance with respect to reflections. The set $\mathcal{T}_D$ is equinumerous to $\mathbb{R}$ (*e.g. *the natural logarithm is a bijection from $(0, \infty)$ to $\mathbb{R}$) and the line $\{ \psi \in \Psi | \, \sigma^2 = 1 \}$ breaks scale invariance.*

Because there are many ways to normalize a model, it is natural to ask how alternatives should be compared. A first natural criterion for choosing a normalization is that it should be observationally unrestrictive.

Definition 4 *Suppose $l(\psi|\, y)$ is the likelihood function of a parametric statistical model $(\mathcal{Y}, \mathcal{F}, \Theta)$. A normalization $\Psi^N \subseteq \Psi$ is **observationally unrestrictive** if there exists a transformation $g : \Psi \to \Psi^N$ such that $l(g(\psi)|\, y) = l(\psi|\, y)$ for all $y \in \mathcal{Y}$. A normalization is observationally restrictive otherwise.*

Normalization does not merely ensure a parameter point estimator is well defined, it also affects its sampling distribution. For example, Hiller (1990) shows how normalization in structural equations models affects the finite-sample distribution of ordinary least squares and two-stage least squares estimators. Unfortunately, as Hamilton, Waggoner, and Zha (2007) note, "the fact that normalization can materially affect the conclusions one draws from likelihood-based methods is not widely recognized."

In Blais (2008), I show that it is in general not possible to ensure unimodal parameter point estimator sampling (or parameter posterior) distributions through an observationally unrestrictive normalization of the parameter space. Building on the work of Hamilton, Waggoner, and Zha (2007), I also argue that normalizations satisfying the following identification principle are more likely (over possible true parameter values) to produce unimodal distributions:

Definition 5 *A normalization $\Psi^N \subseteq \Psi$ satisfies the **identification principle** if it*

a) is observationally unrestrictive;
b) is connected;
c) provides global identification.

Note that intersections of connected spaces are connected. Global identification can be difficult to verify and one often considers the weaker concept of local identification. Local identification can be equivalently defined in terms of the Fisher information matrix. Rothenberg (1971) shows that the parametric model $\left(\mathcal{Y}, \mathcal{F}, \Psi^N\right)$ is locally identified at $\psi_1 \in \Psi^N$ if the Fisher information matrix

$$\mathfrak{I}(\psi_1) \equiv \int_{y \in \mathcal{Y}} \frac{\partial \log l(y|\, \psi)}{\partial \psi} \frac{\partial \log l(y|\, \psi)}{\partial \psi'} f(y|\, \psi)\, \mathrm{d}y \Big|_{\psi_1}$$

is non-singular in a neighborhood of ψ_1. For future reference, let Ψ^l be defined as follows:

Definition 6 *The parameter subspace $\Psi^l \subseteq \Psi$ where the Fisher information matrix is singular or $\log l(y|\, \psi) = -\infty$ is the **singularity parameter subspace**.*

A third identification concept is that of weak identification (or empirical underidentification in the psychometrics literature). Except in the context of instrumental variables (IV) and the generalized method of moments (GMM), weak identification has not been defined precisely. Dufour and Hsiao (2008) write: "More generally, any situation where a parameter may be difficult to determine because we are close to a case where a parameter ceases to be identifiable may be called *weak identification*." Many common situations fit this description. For example, multicollinearity issues arise in linear regression models when the sample covariance matrix of the regressors is "close" to being singular. If one restricts attention to ML inference, weak identification problems occur when the Fisher information matrix is close to being singular at the pseudo-true parameter values. In this paper, I say that a parametric model $\left(\mathcal{Y}, \mathcal{F}, \Psi^N\right)$ is **weakly identified** if the pseudo-true parameter value $\psi^0 \in \Psi^N$ is "close" to Ψ^l.

Weak identification has severe consequences for ML inference, which Dufour and Hsiao (2008) summarize thus:

> "...standard asymptotic distributional may remain valid, but they constitute very bad approximations to what happens in finite samples:
> (1) standard consistent estimators of structural parameters can be heavily biased and follow distributions whose form is far from the limiting Gaussian distribution, such as bimodal distributions, even with fairly large samples (Nelson and Startz, 1990; Hiller, 1990; Buse, 1992);
> (2) standard tests and confidence sets, such as Wald-type procedures based on estimated standard errors, become highly unreliable or completely invalid (Dufour, 1997)"

How close is too close? As weak identification is a finite-sample concern, one might be tempted to believe it is only a small-sample concern. However, Bound et al. (1995) present an IV situation in which weak identification difficulties persist even with 329000 observations. Obviously, if the instruments were uncorrelated with the regressors in population, increasing the sample size would be futile. In practice however, the statistician never knows the pseudo-true parameter values and he should favor inferential methods that are robust to weak identification.

If a normalization provides global identification, then weak identification difficulties only arise when the pseudo-true parameter value is close to the normalization's boundary. In contrast, if a normalization does not provide global identification, then weak identification problems can occurs if the pseudo-true parameter value is close to the singularity subspace. Therefore, an econometrician should use an observationally unrestrictive normalization providing global identification when one exists.

Example 4 (Location mixture) *Consider the following mixture of $K = 2$ normal distributions with common variance σ^2 and means μ_1 and μ_2*

$$f(y|\mu_1, \mu_2, \pi, \sigma) = \pi\mathcal{N}(y|\mu_1, \sigma) + (1 - \pi)\mathcal{N}(y|\mu_2, \sigma).$$

Label (or permutation) invariance refers to the likelihood function's invariance with respect to the re-labeling of the components. Here,

$$f(y|\,\mu_1,\mu_2,\pi,\sigma)=f(y|\,\mu_2,\mu_1,1-\pi,\sigma),$$

which establishes the invariance to the re-labeling (or permuting) of component indices 1 *and* 2*. In matrix notation, this set of invariant transformations is*

$$\mathcal{T}_{\mathbf{P}}\left(\Psi\right)=\left\{T:\Psi\rightarrow\Psi\middle|\,T_{\mathbf{P}}(\mu,\Pi,\sigma)=(\mathbf{P}\mu,\mathbf{P}\Pi,\sigma)\right\}$$

with $\mu=[\mu_1\,\mu_2]'$*,* $\Pi=[\pi\,1-\pi]'$ *and* $\mathbf{P}$ *is a permutation matrix,* i.e. *a matrix obtained by permuting the rows of an identity matrix. Permutation invariance implies that the likelihood function admits two equivalent global maxima, sitting at the summit* ***symmetric lobes****: if* $\left(\hat{\mu},\hat{\Pi},\hat{\sigma}\right)$ *is a global maximum, so is* $\left(\mathbf{P}\hat{\mu},\mathbf{P}\hat{\Pi},\hat{\sigma}\right)$*.*

In order to break permutation invariance, one might contemplate one of the two following normalizations:

$$\Psi^{\pi}=\{\psi\in\Psi|\,\pi>0.5\}$$
$$\Psi^{\mu}=\{\psi\in\Psi|\,\mu_1>\mu_2\}\,.$$

Either of these normalizations would break permutation invariance as

$$\mathcal{T}\left(\Psi^{\pi}\right)=\left\{T:\Psi^{\pi}\rightarrow\Psi^{\pi}\middle|\,T_{\mathbf{P}}(\mu,\Pi,\sigma)=(\mathbf{P}\mu,\mathbf{P}\Pi,\sigma)\,,\,\mathbf{P}=\mathcal{I}\right\}=\mathcal{T}_{\mathcal{I}}$$
$$\mathcal{T}\left(\Psi^{\mu}\right)=\left\{T:\Psi^{\mu}\rightarrow\Psi^{\mu}\middle|\,T_{\mathbf{P}}(\mu,\Pi,\sigma)=(\mathbf{P}\mu,\mathbf{P}\Pi,\sigma)\,,\,\mathbf{P}=\mathcal{I}\right\}=\mathcal{T}_{\mathcal{I}}.$$

However, these normalizations yield different finite-sample inference. Assume one obtains ML estimates $\hat{\psi}=[\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma},\hat{\pi}]'=[1,2,1,0.25]'$ *under* Ψ^{π}*.* $\mathcal{I}(\hat{\psi})$ *has full rank, which is can be verified numerically. However,* $\mathcal{I}(\psi)$ *is singular on* $\{\psi\in\Psi|\,\mu_1=\mu_2\}\subset\Psi^{\pi}$*. Therefore, the sampling distributions of the ML estimator of* μ_1 *and* μ_2 *can be multimodal under* Ψ^{π}*. Intuitively, this normalization would perform poorly if the data came from a mixture distribution with* $\pi=0.5$ *because component densities are equiprobable. The identification principle rules out* Ψ^{π} *because the information matrix is singular on* $\{\psi\in\Psi|\,\mu_1=\mu_2\}\subset\Psi^{\pi}$*. In contrast, the model is globally identified on* Ψ^{μ}*.*

In the latter example, the identification principle yields a unique normalization, under which the ML estimator has a unimodal sampling distribution for any $\psi\in\Psi^{\mu}$. In slightly more general models, the identification principle is less straightforward to apply, as it yields uncountably many normalizations. The practical guidance that the identification principle offers is thus incomplete. Moreover, there is no guarantee that any particular normalization ensures that the ML estimator has a unimodal sampling distribution.

Example 5 *Consider the location-and-scale mixture of normal distributions*

$$f(y_t|\mu_1,\mu_2,\pi,\sigma_1^2,\sigma_2^2) = \pi\phi(y_t|\mu_1,\sigma_1^2) + (1-\pi)\phi(y_t|\mu_2,\sigma_2^2).$$

The set where the information matrix is singular is not a line but a point,

$$\Psi^l = \{\psi \in \Psi |\, \mu_1 = \mu_2\} \cap \{\psi \in \Psi |\, \sigma_1 = \sigma_2\}\,.$$

The identification principle still rules out normalization Ψ^π, but the singularity set no longer separates the parameter space into two symmetric half-spaces. Normalizations Ψ^μ and

$$\Psi^\sigma = \{\psi \in \Psi |\, \sigma_1 > \sigma_2\}$$

both satisfy the identification principle, but neither ensures that all sampling distributions are unimodal. To illustrate, consider samples from a population with $\mu_1 = \mu_2$ and $\sigma_1 > \sigma_2$. Under Ψ^μ, the MLE of σ_1 and σ_2 will both have bimodal sampling distributions for sufficiently large samples (Geweke, 2007).

In this paper, I show that Dai and Singleton's (2000) normalization of ATSMs violates the identification principle and therefore makes inference particularly sensitive to weak identification problems. Normalization of affine transformations is best understood by considering simple affine transformations: translation, scaling, rotation, permutation (labeling) and reflection (signing) of the factors. Permutation and reflection invariance make the likelihood function symmetric and introduce weak identification problems. The consequences of permutation invariance for inference for finite mixture distributions is well documented (Redner and Walker, 1984; Stephens, 2000; Celeux et al., 2000; Frühwirth-Schnatter, 2001). I present evidence that these problems are empirically relevant for ATSMs and that normalization has important consequences for both ML and Bayesian inference. Furthermore, I propose uncountably many normalizations satisfying the identification principle.

Inference methodology

Because uncountably many normalizations of ATSMs satisfy the identification principle, the practical guidance that it offers is thus incomplete. Moreover, there is no guarantee that any particular normalization ensures that the ML estimator has a unimodal sampling distribution. For a given data set, some normalizations yield estimators with better finite-sample properties than others. Hamilton et al. (2007) suggest that one should " try several different normalizations" and "plot the small-sample distributions of parameters." Because one must resort to simulation methods (Stoffer and Wall, 1991), comparing normalizations can thus be computationally demanding or even intractable.

The Bayesian solution to this problem has computational and inferential advantages. Stephens (1997) shows that one can equivalently break permutation invariance within a posterior sampler or as a post-simulation step on the output from an unnormalized sampler. It turns out that his result also applies to reflection invariance.

Observationally unrestrictive normalizations contain no information and Frühwirth-Schnatter (2001) thus proposes choosing a normalization by inspection of the posterior distribution. An econometrician can therefore generate a single sample form the permutation- and reflection-invariant posterior and then normalize the parameter space using the information contained in the data. Moreover, permutation- and reflection-invariant posterior distributions are perfectly valid characterization of uncertainty for permutation- and reflection-invariant quantities (Frühwirth-Schnatter, 2001; Geweke, 2007). In particular, one can make valid and useful inference for observational errors using the permutation- and reflection-invariant posterior distributions. In contrast, inference for observational errors using ML parameter estimates can be completely invalid.

I thus proceed with a Bayesian analysis of the permutation- and reflection-invariant ATSM. I propose a Metropolis-within-Gibbs sampler in which latent factors are drawn together with some parameters as a single block. Because of the high correlations between latent factors and the parameters entering the pricing equations, the proposed sampler is numerically more efficient than one in which factors and parameters are drawn as separate blocks.

Few papers estimate DTSMs by Bayesian methods. Frühwirth-Schnatter and Geyer (1998),Lamoureux and Witte (2002), Müller et al. (2003), and Sanford and Martin (2005) consider CIR models. Ang et al. (2007) use an approximate Gibbs sampler where latent factors are de-meaned. Chib and Ergashev (2008) propose a numerically efficient, exact Gibbs sampler for ATSMs. To the best of my knowledge, none of the literature takes into account weak identification problems associated with permutation and reflection invariance. I use consider the reflection- and permutation-invariant posterior and show that some normalizations yield multimodal marginal posteriors.

This paper is organized as follows. In the first section, I briefly review some ATSM essentials, introduce notation and present the economic model. The second section pertains to error modeling and describes the proxy and latent-factor modeling approaches to error specification. I explain how normalization affects inference for ATSMs in the third section. Section 4 presents permutation- and reflection-invariant prior distributions. In the fifth section, I describe the posterior sampler and discuss its implementation. In the final section, I present empirical results.

2 Economic modeling

The dynamics of the SDF constrain the term structure. This section first presents sufficient conditions for obtaining affine discount rates. Conditions for analytical pricing are easier to express in terms of the risk-neutral measure . They impose a tight relationship between the SDF and the physical factor dynamics, while allowing for

considerable flexibility in other dimensions (see Dai, Le, and Singleton (2006) for an analysis of ATSMs in continuous time, and Bertholon, Monfort, and Pegoraro (2007) for more general pricing models). Given the risk-free dynamics, choosing physical factor dynamics fixes the risk premium, and *vice versa*. In this paper, I opt for a discrete-time version of Duffee's (2002) conditionally-Gaussian factor model, in which the SDF is exponential-affine in these factors.

2.1 Pricing discount bonds

In discrete time, given the nominal SDF at $t+1$, M_{t+1}, the price at time t of a discount bond maturing n periods from t, $P_{n,t}$, satisfies the difference equation

$$P_{n,t} = \mathrm{E}_t^{\mathbb{P}} \left[P_{n-1,t+1} M_{t+1} \right], \tag{2.1}$$

with boundary conditions

$$P_{0,t} = 1, \forall t, \tag{2.2}$$

and where the operator $\mathrm{E}_t^{\mathbb{P}}[.]$ refers to the conditional expectation at t under the physical measure $\mathbb{P}$. For future reference, I define the the log price of the n-period discount bond, $p_{n,t} \equiv \ln P_{n,t}$, and the continuously-compounded yield to maturity of the n-period discount bond as $y_{n,t} \equiv -\frac{p_{n,t}}{n}$. Equation (2.1) may or may not admit an analytical solution, depending on the joint physical dynamics of the prices and the SDF.

In discrete time, markets with are incomplete and the functional form of the SDF must be specified. In this model, the state of the economy at time t is completely specified by a K-dimensional vector of factors X_t. Following Gouriéroux, Monfort, and Polimenis (2002), the log SDF, m_{t+1}, is written in the simplest affine manner as

$$-m_{t+1} = \Lambda_t X_{t+1} + \gamma_t, \tag{2.3}$$

where Λ_t is the time-dependent price of risk. They show how to determine the time-dependent intercept γ_t from (2.1) for compound autoregressive processes, of which the Gaussian process I specify below is a special case.

Under the physical measure $\mathbb{P}$, the latent factors are given by

$$X_{t+1} = \mu_t^{\mathbb{P}} + \Sigma_t^{1/2} \epsilon_{t+1}. \tag{2.4}$$

where $\mu_t^{\mathbb{P}}$ and Σ_t are the conditional mean and covariance of the factor vector, $\Sigma_t^{1/2}$ is the upper Cholesky factor of Σ_t and ϵ_{t+1} is a vector of independent standard normal

random variables.

Writing (2.1) for the one-period bond,

$$\mathrm{e}^{-y_{1,t}} = \mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}}\right],$$

and solving for γ_t yields

$$\gamma_t = y_{1,t} - \Lambda_t \mu^{\mathbb{P}}_t + \frac{1}{2}\Lambda_t \Sigma_t \Lambda_t'. \tag{2.5}$$

Substituting γ_t in the expression for the log-SDF (2.3) gives

$$-m_{t+1} = y_{1,t} + \Lambda_t\left(X_{t+1} - \mu^{\mathbb{P}}_t\right) + \frac{1}{2}\Lambda_t \Sigma_t \Lambda_t'.$$

One still has to specify $\mu^{\mathbb{P}}_t$, Σ_t, $y_{1,t}$ and Λ_t in such a way that (2.1) has a simple solution. It turns out to be much easier to work under the risk-neutral measure $\mathbb{Q}$, for which (2.1) is written as

$$P_{n,t} = \mathrm{e}^{-y_{1,t}} \mathrm{E}^{\mathbb{Q}}_t\left[P_{n-1,t+1}\right], \tag{2.6}$$

and specify the short rate and the risk-neutral factor dynamics in a way that facilitates pricing. DTSMs are therefore often appropriately referred to as *short rate models*. Two such assumptions are the following:

Assumption 2.1 *A1. The short rate is affine in the factors. That is,*

$$Y_{1,t} \equiv Y_1(X_t) = \tilde{A}_1 + \tilde{\mathbf{B}}_1' X_t, \tag{2.7}$$

where $\tilde{A}_1$ *and* $\tilde{\mathbf{B}}_1$ *are constants.*

Assumption 2.2 *A2. Under the risk-neutral measure, the factor dynamics are given by an Gaussian VAR(1) process*

$$X_{t+1} = X_t + \kappa^{\mathbb{Q}}(\theta^{\mathbb{Q}} - X_t) + \Sigma_t^{1/2}\epsilon_{t+1} \tag{2.8}$$

Under equivalent assumptions, Ang and Piazzesi (2003) solve[4] (2.6) and show that prices of discount bonds maturing in $n > 0$ periods satisfy:

[4] Proof is provided in Appendix B for completeness.

$$
\begin{aligned}
P_{n,t} &= \exp\{-\tilde{A}_n - \tilde{\mathbf{B}}_n' X_t\} \\
\text{with} \quad \tilde{A}_{n+1} &= \tilde{A}_1 + \tilde{A}_n + \theta^{\mathbb{Q}\prime}\kappa^{\mathbb{Q}\prime}\tilde{\mathbf{B}}_n - \frac{1}{2}\tilde{\mathbf{B}}_n'\Sigma_t\tilde{\mathbf{B}}_n \\
\text{and} \quad \tilde{\mathbf{B}}_{n+1} &= \tilde{\mathbf{B}}_1 + (\mathcal{I} - \kappa^{\mathbb{Q}\prime})\tilde{\mathbf{B}}_n
\end{aligned} \tag{2.9}
$$

with the boundary conditions (2.2) $\tilde{A}_0 = 0$ and $\tilde{B}_0 = \mathbf{0}$.

Dai, Singleton, and Yang (2005) show[5] how to link physical and risk-neutral measures. Since the price $P_{n,t}$ of a *any* cash flow c_{t+1} can be calculated under (2.1) or (2.6) :

$$
\mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}} c_{t+1}\right] = \mathrm{e}^{-y_{1,t}}\mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[c_{t+1}\right],
$$

one can identify the risk-neutral measure

$$
\mathrm{d}\mathbb{Q} = \mathrm{e}^{y_{1,t}+m_{t+1}}\mathrm{d}\mathbb{P},
$$

and compute the risk-neutral unconditional mean,

$$
\mu_t^{\mathbb{Q}} \equiv \mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[X_{t+1}\right] = \mu_t^{\mathbb{P}} - \Sigma_t\Lambda_t'. \tag{2.10}
$$

Since the change of measure concerns only the conditional mean, Λ_t completely specifies the passage between physical and risk-neutral measure. Alternatively, it is completely specified by $\mu_t^{\mathbb{P}}$ and $\mu_t^{\mathbb{Q}}$ as

$$
\Lambda_t = \Sigma_t^{-1}\left(\mu_t^{\mathbb{P}} - \mu_t^{\mathbb{Q}}\right). \tag{2.11}
$$

2.2 Physical dynamics and risk premia

At this point, the conditionally Gaussian physical dynamics of factors (2.4) is still somewhat general and one needs to specify $\mu_t^{\mathbb{P}}$ and Σ_t to complete the model. Any function of X_t will do. This choice will determine the functional form of the risk premium. For example, regime switching processes can be considered in this framework, as in Dai, Singleton, and Yang (2005) or Monfort and Pegoraro (2007).

In this study, I consider a simple but popular VAR(1) process,

$$
X_{t+1} = X_t + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_t) + \Sigma^{1/2}\epsilon_{t+1}. \tag{2.12}
$$

[5] See Appendix D.

Using (2.11), the risk premium is

$$\begin{aligned}\Lambda_t &= \Sigma^{-1}\left(\left(\kappa^{\mathbb{P}} - \kappa^{\mathbb{Q}}\right) X_t + \kappa^{\mathbb{P}}\theta^{\mathbb{P}} - \kappa^{\mathbb{Q}}\theta^{\mathbb{Q}}\right) \\ &= \Sigma^{-1}\left(\lambda_0 + \lambda_1 X_t\right),\end{aligned}$$

where

$$\lambda_0 \equiv \kappa^{\mathbb{P}}\theta^{\mathbb{P}} - \kappa^{\mathbb{Q}}\theta^{\mathbb{Q}} \tag{2.13}$$

$$\lambda_1 \equiv \kappa^{\mathbb{P}} - \kappa^{\mathbb{Q}}. \tag{2.14}$$

These relations imply that there are only two K-dimensional vectors (from $\left\{\lambda_0, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}\right\}$) and two $K \times K$ matrices (from $\left\{\lambda_1, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}\right\}$) to specify.

3 Error modeling

The economic model presented in the previous section gives a deterministic relationship between the state variables and observed discount rates. The state variables consisting of K factors, the covariance matrix of $N > K$ discount rates has rank K. The econometrician must thus model observational errors in order to obtain a non-singular likelihood.

There is no standard terminology for the modeling of pricing errors in the literature. In this paper, I use *proxy* and *latent-factor* for the error modeling approaches. The *proxy* modeling approach, most often used in the macroeconomics or financial economics literature [6], follows Chen and Scott (1993) and assumes that only $N - K$ yields are observed with error. This is computationally convenient. But it is obviously awkward and theoretically unjustified to maintain, for example, that the model prices 5-year bonds exactly and 4-year bonds with error. Modeling errors on all yields is proposed by Chen and Scott (1995) and is consistent with the fact that the model is a mere simplification of reality and describes it imperfectly. This *latent-factor* modeling approach, is popular in the empirical finance literature [7]. There are two more reasons to preferring the latent-factor approach.

Even if the model were indeed *true*, the construction of discount rates from observable coupon bond yields introduces errors in the former. Coupon bond yields

[6] Examples are: Dai and Singleton (2000); Duffee (2002); Ang and Piazzesi (2003); Evans (2003) and Garcia and Luger (2007).

[7] See Jegadeesh and Pennacchi (1996); Geyer and Pichler (1999) and Babbs and Nowman (1999) for frequentist examples; and Frühwirth-Schnatter and Geyer (1998), Lamoureux and Witte (2002) and Ang et al. (2007)) for Bayesian studies.

are non-linear functions of discount rates. Using quoted yields directly in statistical inference thus presents a computational challenge and the standard approach is thus building discount rates in an *ad hoc* manner, before statistical inference and without reference to the model [8]. Adding pricing errors to the model is a way to account for such data pre-processing. As one alternative to pre-processing bond yields, one can use strip bonds (Lamoureux and Witte, 2002) which give discount rates directly. However, these are arguably less liquid bonds which leads to other problems.

From a statistical point a view, the vector autoregressive moving-average representation of the rate dynamics is quite different under the two error specifications. Simple algebra reveals [9] that the proxy approach implies a VAR(1) representation of the rate dynamics, while latent-factor approach implies a more general VARMA(1,1) representation.

Because one does inference for the economic and the statistical models jointly, a restrictive error model could lead one to wrongly infer that an ATSM provides an inappropriate description of the term structure. In this paper, I compare the proxy and latent-factor approaches by looking at the statistical properties of the residuals under these two specifications.

For notational convenience, let $A_n \equiv \tilde{A}_n/n$ and $\mathbf{B}_n \equiv \tilde{\mathbf{B}}_n/n$ denote the standardized pricing coefficients, which I stack in matrices $\mathbf{A}$ and $\mathbf{B}$. One observes N rates at time t, stacked in a vector y_t. Under the latent-factor modeling approach, pricing and measurement errors add up to a multivariate normal error of covariance Ω and one writes the system as

$$\begin{aligned} y_t &= \mathbf{A} + \mathbf{B}' X_t + \Omega^{1/2} u_t \\ X_t &= X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \Sigma^{1/2}\epsilon_t. \end{aligned} \tag{3.1}$$

Under the proxy modeling approach, the N yields are partitioned into sets of K perfectly observed yields, y_t^p, and $N-K$ imperfectly observed yields, y_t^i, resulting in the system

$$y_t^p = \mathbf{A}^p + \mathbf{B}^{p\prime} X_t$$

$$y_t^i = \mathbf{A}^i + \mathbf{B}^{i\prime} X_t + \Omega^{1/2} u_t \tag{3.2}$$

$$X_t = X_{t-1} + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_{t-1}) + \Sigma^{1/2}\epsilon_t. \tag{3.3}$$

[8] Bliss (1997) explains and compare several such methods. The problem is potentially more important for methods that impose some smoothness to the curve, as the cubic spline method of McCulloch (1975). Imposing a structure actually adds information not contained in the data but it also removes some information as bonds are not priced exactly.

[9] See Appendix C.

The likelihood is then computed by substituting $X_t = \mathbf{B}^{i'^{-1}}(y_t^p - \mathbf{A}^p)$ into (3.2-3.3), which highlights that proxying factors are affine transformations of the perfectly observed yields.

Because latent-factor models essentially decompose the dynamics of the observables into common and idiosyncratic components, error covariance modeling also allows the econometrician to specify which characteristics of the observables the common latent factors should capture. The proxy approach is a special case of the latent-factor approach corresponding to a rather strong restriction on the covariance matrix Ω in which elements are equal to zero. Other restrictions are also likely to affect the factor-error decomposition. Imposing homoscedasticity and independence are examples of such restrictions. In order to consider these restrictions individually, I factorize the covariance matrix into a correlation matrix $\mathbf{R}$ and a diagonal matrix ξ of precisions,

$$\Omega = \mathbf{DRD}' \tag{3.4}$$
$$\xi^{-1} = \mathbf{DD}', \tag{3.5}$$

where $\mathbf{D}$ is the diagonal matrix of standard deviations. In this paper, I propose priors on $\mathbf{R}$ and ξ that operationalize *soft* restrictions on the correlation and precision matrices.

4 Normalization

Let $\psi \in \Psi$ denote the K-factor ATSM's parameter vector,

$$\psi = \{\mathbf{A}_1, \mathbf{B}_1, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, \Sigma, \Omega\}.$$

For any K-dimensional vector $\mathbf{t}$ and any invertible $K \times K$ matrix $\mathbf{M}$,

$$f\left(y \middle| T_{\mathbf{tM}}(\psi), \mathbf{M}(X - \mathbf{t})\right) = f\left(y \middle| \psi, X\right)$$
$$f\left(y \middle| T_{\mathbf{tM}}(\psi)\right) = f\left(y \middle| \psi\right),$$

where

$$T_{\mathbf{tM}}\left(\mathbf{A}_1, \mathbf{B}_1, \theta^{\mathbb{P}}, \theta^{\mathbb{Q}}, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, \Sigma, \Omega\right) =$$

$$\left(\mathbf{A}_1 - \mathbf{B}_1'\mathbf{t}, \mathbf{M}'^{-1}\mathbf{B}_1, \mathbf{M}(\theta^{\mathbb{P}} - \mathbf{t}), \mathbf{M}(\theta^{\mathbb{Q}} - \mathbf{t}), \mathbf{M}\kappa^{\mathbb{P}}\mathbf{M}^{-1}, \mathbf{M}\kappa^{\mathbb{Q}}\mathbf{M}^{-1}, \mathbf{M}\Sigma\mathbf{M}', \Omega\right).$$

We thus say that the density of discount rates is invariant with respect to $\mathcal{T}_{\mathbf{tM}}(\Psi)$ and the parameter vectors ψ and $T_{\mathbf{tM}}(\psi)$ are observationally equivalent.

Decomposing affine transformations into simpler ones clarifies the identification problem. If an function is invariant with respect to some set of transformations $\mathcal{T}_f$ and $T_f(\psi) = T_g(T_h(\psi))$, then a normalization breaking invariance with respect to $\mathcal{T}_h$ and $\mathcal{T}_g$ breaks invariance with respect to $\mathcal{T}_f$. Several decompositions are possible, but a finer decomposition yields more insight than a coarser one. Here, I decompose affine transformations into translations, scaling, rotations, permutations and reflections:

$$
\begin{aligned}
\mathcal{T}_{\mathbf{t}} &= \left\{ T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{M} = \mathcal{I} \right\} \\
\mathcal{T}_{\mathbf{D}} &= \left\{ T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{D}, \mathbf{D}_{ii} > 0, \mathbf{D}_{ij} = 0, j \neq j \right\} \\
\mathcal{T}_{\mathbf{O}} &= \left\{ T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{O}, \mathbf{O}\mathbf{O}' = \mathcal{I}, |\mathbf{O}| = 1 \right\} \\
\mathcal{T}_{\mathbf{P}} &= \left\{ T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{P}, \mathbf{P}_{ij} \in \{0,1\}, \iota'\mathbf{P} = \iota', \mathbf{P}\iota = \iota \right\} \\
\mathcal{T}_{\mathbf{S}} &= \left\{ T_{\mathbf{tM}} \in \mathcal{T}_{\mathbf{tM}} \middle| \, \mathbf{t} = \mathbf{0}, \mathbf{M} = \mathbf{S}, |\mathbf{S}_{ii}| = 1, \mathbf{S}_{ij} = 0, j \neq j \right\}.
\end{aligned}
$$

These transformations have the following geometrical interpretations:

- $X + \mathbf{t}$ translates columns of X by $\mathbf{t}$;
- $\mathbf{D}$ is a diagonal scaling matrix with positive elements, $\mathbf{D}X$ changes the scale of columns of X;
- $\mathbf{O}$ is a rotation matrix, $\mathbf{O}X$ rotates the columns of X in Euclidean space;
- $\mathbf{P}$ is a permutation matrix, $\mathbf{P}X$ swaps the rows of X;
- $\mathbf{S}$ is diagonal reflection (or signing) matrix elements 1 or -1, $\mathbf{S}X$ changes the signs of columns of X;

4.1 Breaking invariance

A normalization Ψ^N breaks invariance with respect $\mathcal{T}_{\mathbf{tM}}$ if

$$\mathcal{T}_{\mathbf{t}}(\Psi^N) = \mathcal{T}_{\mathbf{P}}(\Psi^N) = \mathcal{T}_{\mathbf{D}}(\Psi^N) = \mathcal{T}_{\mathbf{S}}(\Psi^N) = \mathcal{T}_{\mathbf{O}}(\Psi^N) = \mathcal{T}_{\mathcal{I}}$$

where $\mathcal{T}_{\mathcal{I}}$ is defined by equation (1.4).

Dai and Singleton (2000) propose the following normalization of affine models:

$$\Psi^{DS} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{tri}} \cap \Psi^{\Sigma_{\mathcal{I}}} \cap \Psi^{B_1},$$

where

$$
\begin{aligned}
\Psi^{\theta^{\mathbb{P}}} &= \left\{\psi \in \Psi \middle|\, \theta^{\mathbb{P}} = \mathbf{0}\right\} \\
\Psi^{\kappa^{\mathbb{P}}_{tri}} &= \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{P}} \text{ is lower triangular}\right\} \\
\Psi^{\Sigma_{\mathcal{I}}} &= \{\psi \in \Psi |\, \Sigma = \mathcal{I}\} \\
\Psi^{B_1} &= \{\psi \in \Psi |\, B_1 > \mathbf{0}\}\,.
\end{aligned} \tag{4.1}
$$

It is straightforward to show that

$$
\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{\mathcal{I}}}) = \mathcal{T}_{\mathbf{S}}(\Psi^{B_1}) = \mathcal{T}_{\mathbf{t}}(\Psi^{\theta^{\mathbb{P}}}) = \mathcal{T}_{\mathcal{I}},
$$

which confirms that $\mathcal{T}_{\mathbf{tM}}(\Psi^{DS}) = T_{\mathcal{I}}$. For example, showing that $T_{\mathbf{P}}$ is not bijective on $\Psi^{\kappa^{\mathbb{P}}_{tri}}$ only requires a counterexample. Here, one looks for a lower triangular $\kappa^{\mathbb{P}}$ such that $\mathbf{P}\kappa^{\mathbb{P}}\mathbf{P}'$ is not lower triangular: the only permutation matrix $\mathbf{P}$ such that

$$
\mathbf{P}\begin{bmatrix}
1 & 0 & \dots & \dots & 0 \\
1 & 1 & \ddots & & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & & \ddots & 1 & 0 \\
1 & \dots & \dots & 1 & 1
\end{bmatrix}\mathbf{P}'
$$

is lower triangular is the identity matrix, *i.e.* $\mathbf{P} = \mathcal{I}$, which confirms that $\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{tri}}) = \mathcal{T}_{\mathcal{I}}$.

4.2 Weak identification

Dai and Singleton's normalization break invariance with respect to affine transformations but do not satisfy the identification principle, which makes inference sensitive to weak identification issues. In particular, difficulties arise in the region about $\Psi^{\kappa^{\mathbb{P}}_{diag}} = \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{P}} \text{ is diagonal}\right\} \subset \Psi^{\kappa^{\mathbb{P}}_{tri}}$, where the permutation normalization $\Psi^{\kappa^{\mathbb{P}}_{tri}}$ becomes ineffective. Indeed, the likelihood is invariant with respect to permutations on $\Psi^{\kappa^{\mathbb{P}}_{diag}}$,

$$
\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{diag}}) = \mathcal{T}_{\mathbf{P}}(\Psi) \neq \mathcal{T}_{\mathcal{I}}.
$$

Reflection invariance introduce weak identification inferential difficulties too. In both cases, difficulties arise because the Fisher information matrix is singular on the following parameter subspace, where some parameters are locally unidentified:

$$
\tilde{\Psi} = \left(\bigcup_{(i,j)\in\{1,\dots,K\}, j\neq j} \Psi^{\kappa^{\mathbb{P}}_{ij}} \cap \Psi^{\kappa^{\mathbb{Q}}_{ij}} \cap \Psi^{B_{1,ij}} \cap \Psi^{\Sigma_{ij}}\right) \cup \left(\bigcup_{k=1}^{K} \Psi^{B_{k,0}}\right), \tag{4.2}
$$

where

$$
\begin{aligned}
\Psi^{\kappa^{\mathbb{Q}}_{ij}} &= \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{Q}}_{ii} = \kappa^{\mathbb{Q}}_{jj}\right\} \\
\Psi^{\kappa^{\mathbb{P}}_{ij}} &= \left\{\psi \in \Psi \middle|\, \kappa^{\mathbb{P}}_{ii} = \kappa^{\mathbb{P}}_{jj}\right\} \\
\Psi^{\mathbf{B}_{1,ij}} &= \left\{\psi \in \Psi \middle|\, \mathbf{B}_{1,i} = \mathbf{B}_{1,j}\right\} \\
\Psi^{\Sigma_{ij}} &= \left\{\psi \in \Psi \middle|\, \Sigma_{ii} = \Sigma_{jj}\right\} \\
\Psi^{\mathbf{B}_{k,0}} &= \left\{\psi \in \Psi \middle|\, \text{the } k^{th} \text{ row of } \mathbf{B} \text{ is a vector of zeros}\right\}.
\end{aligned} \tag{4.3}
$$

Intuitively, if any factor k contains too little information about the discount rates or if any two factors (i, j) are too similar then identification problems arise. The latter situation is known as the *label switching* problem in the finite mixture literature (Redner and Walker, 1984; Celeux et al., 2000; Frühwirth-Schnatter, 2001) because it is then difficult to break permutation invariance. The former could then be referred to as *sign switching*. In this case, it is difficult to break reflection invariance, which corresponds to changes in factor signs. In that sense, permutation and reflection invariance introduce weak identification issues.

With respect to permutation invariance, the subset (4.2) suggests that any of the following normalizations satisfy the identification principle and would thus yield estimators with better finite-sample properties than $\Psi^{\kappa^{\mathbb{P}}_{tri}}$:

$$
\begin{aligned}
\Psi^{B_{1,ord}} &= \left\{\psi \in \Psi \middle|\, (\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) > 0, \ldots, (\mathbf{B}_{1,K-1} - \mathbf{B}_{1,K}) > 0\right\} \\
\Psi^{\kappa^{\mathbb{P}}_{ord}} &= \left\{\psi \in \Psi \middle|\, (\kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2}) > 0, \ldots, (\kappa^{\mathbb{P}}_{K-1,K-1} - \kappa^{\mathbb{P}}_{K,K}) > 0\right\} \\
\Psi^{\kappa^{\mathbb{Q}}_{ord}} &= \left\{\psi \in \Psi \middle|\, (\kappa^{\mathbb{Q}}_{1,1} - \kappa^{\mathbb{Q}}_{2,2}) > 0, \ldots, (\kappa^{\mathbb{Q}}_{K-1,K-1} - \kappa^{\mathbb{Q}}_{K,K}) > 0\right\}.
\end{aligned}
$$

For example, $\Psi^{\kappa^{\mathbb{P}}_{ord}} \cap \Psi^{\kappa^{\mathbb{P}}_{diag}}$ satisfies the identification principle and breaks invariance with respect to rotation and permutation as

$$
\begin{aligned}
\mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{P}}_{diag}}) &= \mathcal{T}_{\mathcal{I}} \\
\mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{P}}_{ord}}) &= \mathcal{T}_{\mathcal{I}}.
\end{aligned} \tag{4.4}
$$

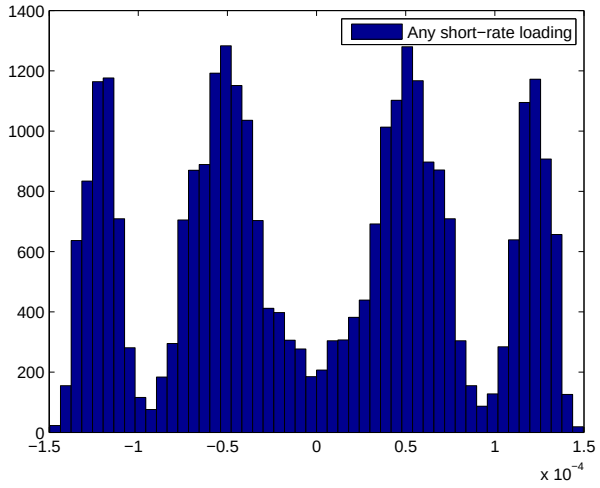
From the the factor loading equation (2.9), reflection normalizations satisfying the identification principle $\Psi^{\mathbf{B}_{k,0}}$ involve elements of $\mathbf{B}_1$ and $\kappa^{\mathbb{Q}}$. For example, Ψ^{B_1} and $\{\psi \in \Psi |\, \mathbf{B}_{1k} > 0,\, \mathbf{B}_{2j} > 0,\, j \neq k\}$ satisfy the identification principle. In terms of the structural parameters $\mathbf{B}_1$ and $\kappa^{\mathbb{Q}}$, this normalization is

$$
\{\psi \in \Psi |\, \mathbf{B}_{1i} > 0,\, \mathbf{B}_{2j} > 0,\, j \neq i\} = \left\{\psi \in \Psi \middle|\, \mathbf{B}_{1i} > 0,\, \kappa^{\mathbb{Q}} | \mathbf{B}_{1,i} > 0,\, j \neq i\right\}.
$$

For the data set I consider in this paper, Figure 1 shows the histogram of a sample from the permutation- and reflection-invariant posterior of $\mathbf{B}_1$. While one factor can perhaps be identified, the other factors cannot be identified from the posterior of $\mathbf{B}_1$

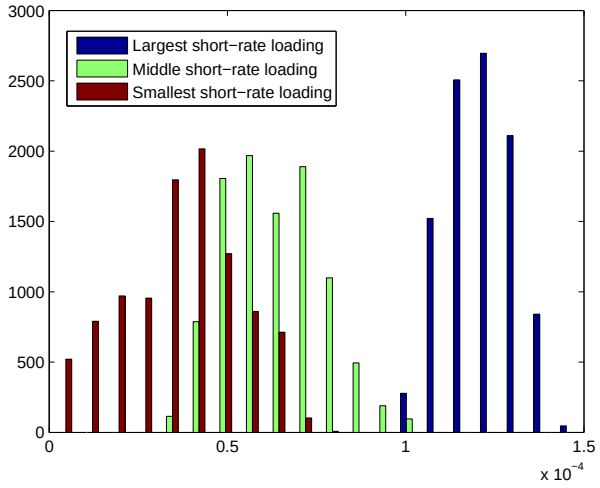
so normalization $\Psi^{B_{1,ord}}$ would not break permutation invariance effectively. Moreover, there is some significant posterior probability in the region about zero and $\Psi^{\mathbf{B}_1}$ would thus not break reflection invariance effectively.

Fig. 1. Sample form the permutation- and reflection-invariant posterior distribution of B_1.



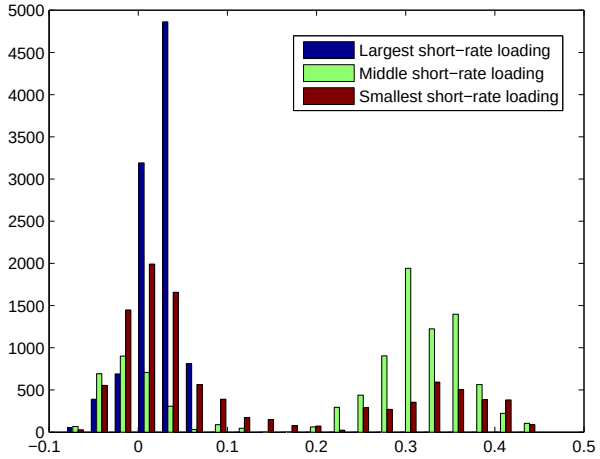
Normalization $\Psi^{\mathbf{B}_1} \cap \Psi^{B_{1,ord}}$ would of course yield unimodal marginal posteriors for $\mathbf{B}_1$ (Figure 2). However, it does not yield unimodal marginal posteriors for every parameters. For example, Figure 3 shows that normalization $\Psi^{\mathbf{B}_1} \cap \Psi^{B_{1,ord}}$ yields bimodal marginal posteriors for two elements of the diagonal of $\kappa^{\mathbb{Q}}$.

Fig. 2. Sample form the normalized posterior distribution of B_1.



There are situations where none of the few normalizations described above yields unimodal posterior distributions (Geweke, 2007). Fortunately, uncountably many normalizations satisfy the identification principle. While there is no guarantee that there

Fig. 3. Sample form the normalized posterior distribution of $\kappa^{\mathbb{Q}}_{kk}$, $k = 1, \ldots, 3$.



exists a normalization ensuring that posteriors are unimodal, an uncountably large set is more likely to contain one than a small finite set.

In order to obtain one family of normalizations satisfying the identification, consider, for example, the following hyperplanes:

$$\left\{\psi \in \Psi |\, \kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2} = 0\right\}$$
$$\left\{\psi \in \Psi |\, \mathbf{B}_{1,1} - \mathbf{B}_{1,2} = 0\right\}.$$

Each of these hyperplanes defines two half-spaces, and normalizations satisfying the identification principle consist in one of these half-spaces, *e.g.*

$$\left\{\psi \in \Psi |\, \kappa^{\mathbb{P}}_{1,1} > \kappa^{\mathbb{P}}_{2,2}\right\}$$
$$\left\{\psi \in \Psi |\, \mathbf{B}_{1,1} > \mathbf{B}_{1,2}\right\}.$$

These normalizations satisfy the identification because their frontier includes the singularity set and their interiors do not intersect with the singularity set (Definition 6). Note that, for any odd bijections $g_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, the half-spaces defined by the following hyperplanes also satisfy the identification principle:

$$\left\{\psi \in \Psi |\, g_1(\kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2}) = 0\right\}$$
$$\left\{\psi \in \Psi |\, g_2(\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) = 0\right\}.$$

Examples of useful bijections are those defining changes of coordinate system. Furthermore, the half-spaces defined by any convex combination of the above hyperplanes,

$$\left\{\psi \in \Psi \middle| \alpha g_1(\kappa^{\mathbb{P}}_{1,1} - \kappa^{\mathbb{P}}_{2,2}) + (1-\alpha) g_2(\mathbf{B}_{1,1} - \mathbf{B}_{1,2}) = 0, \alpha \in [0,1]\right\}$$

also satisfy the identification principle.

4.3 Observational restrictions

One might remark that a triangular matrix has more non-zero elements than a diagonal matrix of the same dimension and conclude that $\Psi^{\kappa^{\mathbb{P}}_{diag}}$ imposes observational restrictions. This is indeed the case. Note, however, that Dai and Singleton's (2000) normalization is already observationally restrictive as breaking scale invariance does not require restrictions on the off-diagonal elements of Σ,

$$\mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{\mathcal{I}}}) = \mathcal{T}_{\mathbf{D}}(\Psi^{\Sigma_{corr}}) = \mathcal{T}_{\mathcal{I}},$$

where

$$\Psi^{\Sigma_{corr}} = \{\psi \in \Psi | \, \Sigma \text{ is a correlation matrix}\}. \tag{4.5}$$

The empirical investigation I present in this paper does not address whether these restrictions reduce model flexibility in a significant manner. I estimate a model that has the same number of parameters as Dai and Singleton's (2000) canonical representation, and I do not break permutation- and reflection-invariance. My normalization is thus

$$\Psi^{B} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{diag}} \cap \Psi^{\Sigma_{corr}},$$

for which

$$\mathcal{T}_{\mathbf{S}}(\Psi^{B}) = \mathcal{T}_{\mathbf{S}}(\Psi)$$
$$\mathcal{T}_{\mathbf{P}}(\Psi^{B}) = \mathcal{T}_{\mathbf{P}}(\Psi).$$

Other normalizations impose observational restrictions. For example, Ang, Dong, and Piazzesi (2007) use

$$\Psi^{ADP} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{P}}_{tri}} \cap \Psi^{\Sigma_{diag}} \cap \Psi^{B_{1,\iota}}$$

where

$$\Psi^{B_{1,\iota}} = \{\psi \in \Psi | \, B_1 = \iota\}$$

The normalization $\Psi^{B_{1,\iota}}$ is observationally restrictive because the short rate is then affine in all K factors, *i.e.* models where an element of $\mathbf{B}_1$ is zero and the corresponding factor drives only risk premium are ruled out.

Since the work of Litterman and Scheinkman (1991), many econometricians look for factor interpretations in terms of level, slope and curvature (LSC) of the term structure. For a simpler term structure model, Gouriéroux et al. (2002) argue that rotation invariance implies that such factors must be looked for in an uncountable set, which is impractical. Indeed, by continuously rotating factors, one might find ones with interpretations close to level, slope and curvature of the term structure. Christensen, Diebold, and Rudebusch (2009) show that Nelson-Siegel term structure models are observationally restrictive affine models with LSC factors where

$$\kappa^{\mathbb{Q}} = \begin{bmatrix} 0 & \kappa & -\kappa \\ 0 & 0 & \kappa \\ 0 & 0 & 0 \end{bmatrix}.$$

From equation (2.9) however, LSC factors can be obtained more generally through the following parameter restriction:

$$\Psi^{\kappa^{\mathbb{Q}}_{LSC}} = \left\{ \psi \in \Psi \,\middle|\, \kappa^{\mathbb{Q}} = \begin{bmatrix} 0 & \kappa^{\mathbb{Q}}_{1,2} & \kappa^{\mathbb{Q}}_{1,3} \\ 0 & 0 & \kappa^{\mathbb{Q}}_{2,3} \\ 0 & 0 & \kappa^{\mathbb{Q}}_{3,3} \end{bmatrix} \right\}. \tag{4.6}$$

Because $\Psi^{\kappa^{\mathbb{Q}}_{LSC}}$ breaks invariance with respect to rotations and permutations,

$$\mathcal{T}_{\mathbf{O}}(\Psi^{\kappa^{\mathbb{Q}}_{LSC}}) = \mathcal{T}_{\mathbf{P}}(\Psi^{\kappa^{\mathbb{Q}}_{LSC}}) = \mathcal{T}_{\mathcal{I}},$$

the following normalization breaks invariance with respect to affine transformations:

$$\Psi^{LSC} = \Psi^{\theta^{\mathbb{P}}} \cap \Psi^{\kappa^{\mathbb{Q}}_{LSC}} \cap \Psi^{\Sigma_{corr}} \cap \Psi^{B_1}.$$

5 Parameterization and prior specification

5.1 Parameterization

Reparameterization consists in defining a one-to-one mapping from a parameter space to another one, which often takes the form of a change of coordinate system. Some

parameterizations yield parameters that are easier to interpret than others, which facilitates prior specification: I would prefer correlations to covariances on that basis. Other parameterizations may affect the numerical efficiency or stability of some algorithms. For example, one often considers the logarithm of a standard deviation in maximization routines because it maps the real line onto the positive half-line. Parameterization also affects the performance of posterior simulator (See Frühwirth-Schnatter, 2004, for an application to state space models.) Note that while I could use one parameterization for prior specification and another one for numerical efficiency reasons, this does not seem to be necessary here.

5.1.1 Long term discount rate factor loadings

My reparameterization of $\kappa^{\mathbb{Q}}$ is based on a novel analytic solution of the factor loadings recurrence equation (2.9). Assuming that $\kappa^{\mathbb{Q}}$ is eigendecomposable,

$$\begin{aligned}
\mathbf{B}_n &= \sum_{i=0}^{n-1} (\mathcal{I} - \kappa^{\mathbb{Q}'})^i \mathbf{B}_1 \\
&= [\mathcal{I} - (\mathcal{I} - \kappa^{\mathbb{Q}'})^n][\mathcal{I} - (\mathcal{I} - \kappa^{\mathbb{Q}'})]^{-1}\mathbf{B}_1 \\
&= [\mathcal{I} - (\delta\gamma\delta^{-1})^n][\mathcal{I} - (\delta\gamma\delta^{-1})]^{-1}\mathbf{B}_1 \\
&= [\delta\delta^{-1} - (\delta\gamma\delta^{-1})^n][\delta\delta^{-1} - (\delta\gamma\delta^{-1})]^{-1}\mathbf{B}_1 \\
&= \delta[\mathcal{I} - \gamma^n][\mathcal{I} - \gamma]^{-1}\delta^{-1}\mathbf{B}_1, \qquad (5.1)
\end{aligned}$$

where the third line uses the eigendecomposition of $\mathcal{I} - \kappa^{\mathbb{Q}'}$, i.e.

$$\mathcal{I} - \kappa^{\mathbb{Q}'} = \delta\gamma\delta^{-1}. \qquad (5.2)$$

These eigenvalues, γ, thus play a central role in long-term factor loadings via γ^n, and one should expect the data to be informative about these quantities. Note that this reparameterization is not a bijection: it assumes that $\kappa^{\mathbb{Q}}$ is eigendecomposable, *i.e.* that it has K distinct eigenvalues. Recall that invertibility does not ensure eigendecomposability. Furthermore, I restrict the parameter space to matrices with real-valued eigenvalues. Complex eigenvalues would generate a sinusoidal pattern in factor loadings, in which, for example, odd-month maturities could be more sensitive to some factor than even-month maturities.

Eigenvector are defined up to a scalar multiplication so I consider normalized unit-length eigenvectors with positive first-element, which I parameterize in polar coordinates, omitting the radial coordinate. Define the matrix of angles, $\phi \equiv [\phi_1 \dots \phi_K]$, where the vector $\phi_j \in (-\frac{\pi}{2}, \frac{\pi}{2}]^{K-1}$, $j = 1, \dots, K$, contains the angles associated with the eigenvector δ_j:

$$\phi_{k,j} \equiv \arctan\left(\frac{\delta_{k+1,j}}{\sqrt{\sum_{i=1}^{k}\delta_{i,j}^2}}\right) \quad \text{for} \quad k = 1, \ldots, K-1. \tag{5.3}$$

For the benchmark model I estimate in this paper, the average posterior correlations of the elements of $[\gamma, \phi]$ and those of $\kappa^{\mathbb{Q}}$ are respectively 0.26 and 0.38. Although I do not investigate the effect of this parameterization on the numerical efficiency of the posterior sampler, lower posterior correlation may result in better mixing.

5.1.2 Short rate factor loadings

For $K > 1$, I use a parameterization (ζ, σ) of the short rate factor loadings $\mathbf{B}_1$ in polar coordinates. I define ζ to be the $K-1$-vector of angles $[\zeta_1, \ldots, \zeta_{K-1}] \in (0, 2\pi] \times (-\frac{\pi}{2}, \frac{\pi}{2}]^{K-2}$, where

$$\zeta_k \equiv \arctan\left(\frac{\mathbf{B}_{1,k+1}}{\sqrt{\sum_{i=1}^{k}\mathbf{B}_{1,i}^2}}\right) \quad \text{for} \quad k = 1, \ldots, K-1, \tag{5.4}$$

and σ to be the logarithm of the Euclidean norm of $\mathbf{B}_1$,

$$\sigma \equiv \log\left(\sqrt{\mathbf{B}_1'\mathbf{B}_1}\right). \tag{5.5}$$

This parameterization in natural logarithm results from a computational consideration: the posterior distribution of the norm of $\sqrt{\mathbf{B}_1'\mathbf{B}_1}$ has thick tails which the posterior sampler has difficulties exploring. Considering σ improves mixing considerably.

From equation (5.1), note that e^{σ} can be interpreted as the common standard deviation of factor innovations: rates can be written as

$$\begin{aligned} Y_{n,t} &= A_n + \mathbf{B}_1^{*\prime}\delta^{-1\prime}[\mathcal{I} - \gamma]^{-1}[\mathcal{I} - \gamma^n]\delta' X_t^* \\ X_t^* &= (\mathrm{I} - \kappa^{\mathbb{P}})X_{t-1}^* + e_t^*, \end{aligned} \tag{5.6}$$

with $\mathbf{B}_1^* = \mathrm{e}^{-\sigma}\mathbf{B}_1$ a unit-length factor-loading vector, and $e_t^* \sim \mathcal{N}(0, \mathrm{e}^{\sigma}\Sigma)$, where Σ is a correlation matrix.

5.1.3 Error covariance matrix

I parameterize the error covariance matrix as a diagonal matrix of precisions ξ and a correlation matrix $\mathbf{R}$ as in (3.4), which I repeat here:

$$\begin{aligned}\Omega &= \mathbf{DRD}' \\ \xi^{-1} &= \mathbf{DD}',\end{aligned}$$

where $\mathbf{D}$ is the diagonal matrix of standard deviations.

5.1.4 Stationarity

Over short horizons, the dynamics of interest rates might not be well described by stationary processes. In order increase flexibility, I do not impose factor stationarity and consider the level of factors at $t = 1$ as an extra parameter, X_1. I therefore allow for co-integration, as yields could share a common unit-root factor.

5.1.5 Parameterization summary

To summarize the parameterizations and normalizations I use in this paper, descriptions of the parameters are given in Table 1. Because I make inference for permutation- and reflection-invariant ATMSs, these normalizations do not break permutation or reflection invariance.

5.1.6 Mapping parameters between parameter subspaces

Because permutation- and reflection-invariance implies that the likelihood function has $K!2^K$ symmetric modes, an observationally unrestrictive normalization consists in an element of a partition of the parameter space into $K!2^K$ observationally equivalent subspaces. The next section describes an extension of Frühwirth-Schnatter's (2001) permutation sampler that maps a parameter vector $\psi \in \Psi$ to $\psi^N \in \Psi^N$, where Ψ^N has one of the two following interpretations. It can be a normalization, in which case the algorithm is used in order to normalize a sample from an un-normalized posterior sampler. It could also be a randomly chosen element of the partition associated with a normalization, in which case the algorithm is used in order to efficiently explore all $K!2^K$ observationally equivalent subspaces. Because permutation and reflection matrices are orthogonal matrices, mapping one parameter subspace to another is achieved by the following transformation

$$T_{\mathbf{SP}}\left(\left\{\mathbf{B}_1, \Lambda_0, \Sigma, \kappa^{\mathbb{P}}, \kappa^{\mathbb{Q}}, X\right\}\right)$$

Table 1
Summary of parameterization and restrictions.

Estimated parameters	
A_1	Mean short-rate; positive.
(ζ, σ)	Spherical parameterization (log-radius) of $\mathbf{B}_1$, (5.4-5.5).
$\kappa^{\mathbb{P}}$	Physical mean-reversion diagonal matrix (4.4).
Σ	Factor correlation matrix (4.5).
(γ, ϕ)	Eigendecomposition of $\kappa^{\mathbb{Q}}$; spherical parameterization (unit radius) of eigenvectors (5.2-5.3).
λ_0	Mean risk premium.
ξ	Pricing error precisions (3.4).
$\mathbf{R}$	Pricing error correlation matrix (3.4).
X_1	Factor vector at $t = 1$.
Derived parameters	
$\mathbf{B}_1$	Short-rate factor loading (5.4-5.5).
$\kappa^{\mathbb{Q}}$	Risk-neutral mean-reversion matrix (5.2-5.3).
λ_1	Factor coefficient in risk premium (2.14).
$\theta^{\mathbb{Q}}$	Risk-neutral factor mean (2.13).
Ω	Pricing error covariance (3.4).
Fixed parameters	
$\theta^{\mathbb{P}}$	Physical factor mean (4.1).

Estimated parameters are used directly in the inference and have prior distributions associated to them. *Derived parameters* are functions of the *estimated parameters*. *Fixed parameters* are constrained by normalizations.

$$= \left\{ \mathbf{SPB}_1, \mathbf{SP}\Lambda_0, \mathbf{SP}\kappa^{\mathbb{P}}\mathbf{P}'\mathbf{S}', \mathbf{SP}\kappa^{\mathbb{Q}}\mathbf{P}'\mathbf{S}', \mathbf{SP}\Sigma\mathbf{P}'\mathbf{S}', \mathbf{SP}X \right\} \tag{5.7}$$

where $\mathbf{S}$ is a reflection matrix and $\mathbf{P}$ is a permutation matrix.

A few properties of this mapping should be noted, as I use them in order to propose permutation- and reflection-invariant prior distributions. First, pre-multiplying a vector by a reflection matrix $\mathbf{S}$ changes its direction and pre-multiplying it by $\mathbf{P}$ changes its orientation. In both cases, the Euclidean norm is preserved. In particular,

$$\sigma = \log\left(\sqrt{\mathbf{B}_1'\mathbf{B}_1}\right) = \log\left(\sqrt{(\mathbf{SPB}_1)'(\mathbf{SPB}_1)}\right). \tag{5.8}$$

Second, if the eigendecomposition of a matrix $\mathbf{A}$ is $\mathbf{A} = \delta\gamma\delta^{-1}$, then

$$\mathbf{SPAP'S'} = (\mathbf{SP}\delta)\gamma(\mathbf{SP}\delta)^{-1}. \tag{5.9}$$

So the mapping changes the direction and orientation of the eigenvectors of $\mathbf{A}$ and leave its eigenvalues unchanged.

5.2 Prior distributions

I propose permutation- and reflection-invariant priors. For finite mixture distributions, Geweke (2007, p. 3537) argues that "If the state labels have no substantive interpretation, then the prior density must also be permutation invariant." His argument applies to reflection invariance as well. Prior distribution hyper-parameters are given in Appendix A.

There are many ways to designing permutation- and reflection-invariant priors, as all one needs to do is ensure that no information is provided with respect either permutations or reflections. The conceptually simplest approach is to specify arbitrary prior distributions and consider the equiprobable mixture of these priors over all $K!2^K$ permutation and reflection combinations. Alternative approaches require some analysis to see how each element of each parameter is affected by permutation and reflection. Reparameterization sometimes helps in this analysis. Some parameters are naturally reflection-invariant, *e.g.* γ or the diagonal of $\kappa^{\mathbb{P}}$. Exchangeable prior distributions are permutation-invariant for some parameters, *e.g.* the diagonal elements of $\kappa^{\mathbb{P}}$ [10]. As a special case, i.i.d. univariate priors are permutation-invariant. Priors that are symmetric with respect to 0 are reflection-invariant. They are equivalently specified as priors on the absolute values of the parameters.

In this section, I propose conditionally conjugate priors when they are available. An exchangeable normal distribution has the form $\mathcal{N}(\mu\iota, \sigma^2((1-\rho)\mathcal{I} + \rho\iota\iota'))$.

5.2.1 Prior distribution of ξ

I use a hierarchical prior for $\xi \equiv \text{diag}(\Omega^{-1})$ that is a Inverse-Gamma-scale-mixture of an N-dimensional vector of conditionally independent Gamma distributions. Specifically,

[10] This might perhaps sound tautological, as an exchangeable distribution defined as a permutation invariant distribution. However, permutations of the parameters need not correspond to permutations of the factors.

$$p(\xi|\gamma_\Omega^0,\nu_\Omega^0,\beta_\Omega^0) = \int_0^\infty \prod_{n=1}^N \mathcal{G}\left(\xi_n|\gamma_\Omega^0,\frac{\eta}{\gamma_\Omega^0}\right)\mathcal{IG}\left(\eta|\nu_\Omega^0,\beta_\Omega^0\right)\,\mathrm{d}\eta.$$

One can integrate η out and write this mixture in closed form as (See appendix E.)

$$p(\xi|\gamma_\Omega^0,\nu_\Omega^0,\beta_\Omega^0) = \frac{{\gamma_\Omega^0}^{N\gamma_\Omega^0}{\beta_\Omega^0}^{\nu_\Omega^0}\Gamma(N\gamma_\Omega^0+\nu_\Omega^0)}{\Gamma(\nu_\Omega^0)\Gamma(\gamma_\Omega^0)^N}\frac{\prod_{n=1}^N \xi_n^{\gamma_\Omega^0-1}}{\left(\beta_\Omega^0+\gamma_\Omega^0\sum_{n=1}^N\xi_n\right)^{N\gamma_\Omega^0+\nu_\Omega^0}}. \tag{5.10}$$

This prior allows one to express separately prior knowledge about the global scale of the precisions and their dispersion. A large value of γ_Ω^0 corresponds to strong belief that errors are nearly identically distributed. A small value of ν_Ω^0 expresses little knowledge about the scale of precisions. β_Ω^0 is a level parameter that centers precisions around $\beta_\Omega^0/(\nu_\Omega^0-1)$.

Note that this prior is conditionally conjugate in any Gaussian model.

5.2.2 Prior distribution of $\mathbf{R}$ *and* Σ

$\mathbf{R}$ and Σ are correlation matrices and I use a prior distribution proposed by Barnard, McCulloch, and Meng (2000). They obtain this distribution by integrating the standard deviations out of an inverse-Wishart-distributed covariance matrix with identity matrix scale parameter. Defining the one-to-one mapping $g(\mathbf{R},\mathbf{D}) = \mathbf{DRD}' = \Omega$, which decomposes a covariance matrix Ω into a diagonal matrix of standard deviations $\mathbf{D}$ and a correlation matrix $\mathbf{R}$, the distribution is

$$\begin{aligned} p(\mathbf{R}|\tau) &= \int \mathcal{IW}\left(\mathbf{DRD}'|\mathcal{I},\tau\right)\,\left|\mathcal{J}(\mathbf{D},\mathbf{R})\right|\,\mathrm{d}\mathbf{D} \\ &= |\mathbf{R}|^{\frac{1}{2}(\tau-1)(N-1)-1}\left(\prod_{i=1}^N|\mathbf{R}_{(ii)}|\right)^{-\frac{\tau}{2}}, \end{aligned}$$

where $\mathcal{IW}\left(\mathbf{DRD}'|\mathbf{W},\tau\right)$ is the Inverse Wishart distribution with shape τ and scale $\mathbf{W}$, $\mathcal{J}(\mathbf{D},\mathbf{R})$ is the Jacobian of the mapping $g(\cdot)$ and $\mathbf{R}_{(ii)}$ is the ith principal submatrix of $\mathbf{R}$. It has the property that individual correlations have Beta marginal distributions $\mathcal{B}eta(\frac{\tau-N+1}{2},\frac{\tau-N+1}{2})$ extended to $[-1,1]$ (*i.e.* $(\mathbf{R}_{ij}+1)/2$ has a Beta marginal distribution), which is uniform over $[-1,1]$ for $\tau = N+1$. My priors are thus $p(\mathbf{\Sigma}|\tau_\Sigma^0)$ and $p(\mathbf{R}|\tau_\mathbf{R}^0)$. Note that $p(\mathbf{\Sigma}|\tau_\Sigma^0)$ is permutation- and reflection-invariant because all correlations have identical marginal priors that are symmetric with respect to 0.

5.2.3 Joint prior distribution of A_1 and λ_0

The system of difference equations (2.9) that the pricing coefficients satisfy introduces non-linearities that are generally viewed as preventing an analytical expression of the risk premium's conditional posterior distribution. I propose the following novel solution of (2.9) in order to obtain the conditional posterior of $\mathbf{a} \equiv [A_1 \quad \lambda_0']'$.

Write the pricing equations as

$$\begin{aligned}
\tilde{A}_n &= \mathbf{a}' \begin{bmatrix} n \\ \sum_{i=1}^{n-1} \tilde{\mathbf{B}}_i \end{bmatrix} - \frac{1}{2} \left(\sum_{i=1}^{n-1} \tilde{\mathbf{B}}_i' \Sigma \tilde{\mathbf{B}}_i \right) \\
\tilde{\mathbf{B}}_n &= \mathbf{B}_1 + (\mathrm{I} - \kappa^{\mathbb{Q}'}) \tilde{\mathbf{B}}_{n-1},
\end{aligned}$$

and define

$$\begin{aligned}
\Delta_{1(K+1\times N)} &= \begin{bmatrix} 1 & \dots & 1 \\ \frac{\sum_{i=1}^{n_1-1} \tilde{\mathbf{B}}_i}{n_1} & \dots & \frac{\sum_{i=1}^{n_N-1} \tilde{\mathbf{B}}_i}{n_N} \end{bmatrix} \\
\Delta_{2(N\times 1)} &= \frac{1}{2} \begin{bmatrix} \frac{\sum_{i=1}^{n_1-1} \tilde{\mathbf{B}}_i' \Sigma \tilde{\mathbf{B}}_i}{n_1} & \dots & \frac{\sum_{i=1}^{n_N-1} \tilde{\mathbf{B}}_i' \Sigma \tilde{\mathbf{B}}_i}{n_N} \end{bmatrix}'.
\end{aligned}$$

The conditional posterior is then

$$p(\mathbf{a}|y, X, \Omega, \Sigma, \kappa^{\mathbb{Q}}, \mu_{\lambda_0}, \Sigma_{\lambda_0}) = N\left(\mathbf{a}|\hat{\mu}_{\mathbf{a}}, \hat{\Sigma}_{\mathbf{a}}\right) p(\mathbf{a}|\mu_{\mathbf{a}}, \Sigma_{\mathbf{a}}),$$

where

$$\hat{\Sigma}_{\mathbf{a}}^{-1} = T \Delta_1 \Omega^{-1} \Delta_1' \tag{5.11}$$

$$\hat{\mu}_{\mathbf{a}} = \hat{\Sigma}_{\mathbf{a}} \Delta_1 \Omega^{-1} \left[T\Delta_2 - T\iota_N + \sum_{t=1}^{T} y_t - \mathbf{B}' X_t \right], \tag{5.12}$$

and $\{n_1, n_2, \dots, n_N\}$ is the set of N maturities, which shows that the conditional posterior admits conjugate Gaussian priors.

The rank of $\hat{\Sigma}_{\mathbf{a}}$ is at most equal to that of Δ_1, which is $K+1$ unless loadings are constant over maturities for one factor ($\mathbf{B}_{n,k} = b_k$ for $n = 1, \dots, N$), or the entire term structure is identically sensitive to two factors ($\mathbf{B}_{n,k} = \mathbf{B}_{n,j}$, for $n = 1, \dots, N$ and $K \neq j$). The latter case corresponds to a parameter subspace $\Psi^{\mathbf{B}_{k,0}}$ defined by equation (4.3).

The former case corresponds to a situation where one factor has a pure level interpretation: a change in that factor shifts the entire term structure. Asymptotically,

one would expect the sample mean of this factor to be equal to its population mean, which is zero by my normalization $\Psi^{\theta^{\mathbb{P}}}$ (see equation 4.1), and expect the factor to describe time variations in the short rate mean around A_1. In finite sample however, A_1 is imprecisely estimated: because factors are highly correlated, the factor's sample mean $\bar{X}$ can be significantly different from zero, its population value by normalization. This provides an interesting explanation of the poor performance of Gibbs samplers for ATSMs. The sample mean of discount rates is

$$\bar{y} = \mathbf{a}'\Delta_1 - \frac{1}{2}\Delta_1 + \mathbf{B}'\bar{X}.$$

Because $\mathbf{a}$ and $\bar{X}$ play similar roles in the description of average discount rates, simulations not reported in this paper reveal that Gibbs sampling schemes where $\mathbf{a}$ and X are drawn as separate blocks result in poor mixing. One can overcome this inferential difficulty by fixing the value of A_1 to some reasonable value (which is observationally restrictive), or by constraining the factors' sample mean to being zero (Ang et al., 2007) (so that factors are not longer drawn from their full conditional posterior). The posterior sampler I describe in the next section, which draws $\mathbf{a}$ and X as a single block, is exact and mixes much better than the Gibbs schemes described above.

My priors are

$$\begin{aligned} p(A_1) &= \mathcal{N}(A_1|\mu^0_{A_1}, \Sigma^0_{A_1})\mathbf{1}_{A_1>0} \\ p(\lambda_{0,k}) &= \mathcal{N}\left(\lambda_0|\mu^0_{\lambda_0}, \Sigma^0_{\lambda_0}\right), \end{aligned}$$

for $k = 1, \ldots, K$, where the truncation $\mathbf{1}_{A_1>0}$ reflects my personal belief that the mean nominal short rate considered in this paper is positive.

5.2.4 Prior distribution of σ

The logarithm of the Euclidean norm of $\mathbf{B}_1$ (the global scale of factor innovations, see equations 5.5 and 5.6) is normally distributed

$$p(\sigma|\mu^0_\sigma, \Sigma^0_\sigma) \equiv \mathcal{N}(\sigma|\mu^0_\sigma, \Sigma^0_\sigma).$$

From equation (5.8), any prior on σ is permutation- and reflection-invariant.

5.2.5 Prior distribution of ζ

The short-rate vector of factor loadings is a priori uniformly distributed on a K-dimensional hyper-sphere with radius e^σ, which implies the following prior distribution on the angles:

$$p(\zeta) \equiv \frac{1}{2\pi} \prod_{k=2}^{K} \frac{1}{4\pi} \cos(\zeta_k).$$

Since permutations and reflections only change the direction and orientation of the factor loadings, this prior is permutation- and reflection-invariant.

5.2.6 Prior distribution of $\kappa^{\mathbb{P}}$

I do not impose stationarity and use a i.i.d. normal distribution

$$p(\kappa^{\mathbb{P}}_{k,k}) = \mathcal{N}(\kappa^{\mathbb{P}}_{k,k} | \mu^0_{\kappa^{\mathbb{P}}}, \Sigma^0_{\kappa^{\mathbb{P}}}),$$

for $k = 1, \ldots, K$.

5.2.7 Prior distribution of γ

The eigenvalues of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori i.i.d. normally distributed

$$p(\gamma_k | \mu^0_\gamma, \Sigma^0_\gamma) \equiv \mathcal{N}(\gamma_k | \mu^0_\gamma, \Sigma^0_\gamma).$$

From (5.9), any prior is permutation- and refection-invariant.

5.2.8 Prior distribution of ϕ

Eigenvectors are defined up to a scalar multiplication so I consider the normalized unit-length eigenvectors with positive first-element. The K eigenvectors of $\mathcal{I} - \kappa^{\mathbb{Q}'}$ are a priori uniformly distributed on the unit half-sphere, which implies the following prior distribution on the angles:

$$p(\phi_k) \equiv \frac{1}{\pi} \prod_{j=2}^{K} \frac{1}{4\pi} \cos(\phi_{k,j}),$$

for $k = 1, \ldots, K$. Again, permutations and reflections only affects the eigenvector directions and orientations, and this prior is thus permutation- and reflection-invariant.

6 Posterior simulator

This section describes a Metropolis-within-Gibbs sampler combined with a extension of Früwirth-Schnatter's (2001) permutation sampler.

6.1 MCMC algorithm

Defining the parameter vector

$$\vartheta \equiv \{A_1, \lambda_0, \Omega, \sigma, \zeta, \gamma, \phi, \Sigma\},$$

my Metropolis-Hastings update of the chain consists of the following cycle of parameter and state updates:

Given the state of the Markov chain at iteration $(m-1)$,

(1) Generate $\kappa^{\mathbb{P}*} \sim p\left(\kappa^{\mathbb{P}} \middle| y, X_1^{(m-1)}, \vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right)$.

(2) Generate $X_1^* \sim p\left(X_1 \middle| y, \kappa^{\mathbb{P}*}, \vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right)$.

(3) Generate $(\vartheta', X'_{t=2:T}) \sim q\left(\vartheta, X_{t=2:T} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)$.

(4) Take

$$(\vartheta^*, X^*_{t=2:T}) = \begin{cases} (\vartheta', X'_{t=2:T}) & \text{with probability } \rho \\ \left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)}\right) & \text{with probability } 1-\rho \end{cases},$$

where

$$\rho = \min\left\{ \frac{p\left(\vartheta', X'_{t=2:T} \middle| y, \kappa^{\mathbb{P}*} X_1^*\right)}{p\left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)} \frac{q\left(\vartheta^{(m-1)}, X_{t=2:T}^{(m-1)} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)}{q\left(\vartheta', X'_{t=2:T} \middle| y, \kappa^{\mathbb{P}*}, X_1^*\right)} \right\}.$$

(5) Generate $\mathbf{S}$ uniformly over the $K!$ signing matrices.

(6) Generate $\mathbf{P}$ uniformly over the 2^K permutation matrices.

(7) Take (see equation 5.7)

$$\left\{\mathbf{B}_1^{(m)}, \Lambda_0^{(m)}, \Sigma^{(m)}, \kappa^{\mathbb{P}(m)}, \kappa^{\mathbb{Q}(m)}, X^{(m)}\right\} = T_{\mathbf{SP}}\left(\left\{\mathbf{B}_1^*, \Lambda_0^*, \Sigma^*, \kappa^{\mathbb{P}*}, \kappa^{\mathbb{Q}*}, X^*\right\}\right).$$

The proposal in the Metropolis-Hastings defined by steps (3-4) is

$$q\left(\vartheta', X'_{t=2:T} \middle| y, \vartheta, \kappa^{\mathbb{P}}, X_1\right) = p\left(X'_{t=2,\ldots,T} \middle| y, \vartheta', \kappa^{\mathbb{P}}, X_1\right) \mathcal{N}\left(\vartheta' \middle| \vartheta, \Sigma_\vartheta\right), \tag{6.1}$$

where the density $p\left(X'_{t=2,\ldots,T} \middle| y, \vartheta', \kappa^{\mathbb{P}}, X_1\right)$ can be computed exactly using an algorithm independently suggested by Carter and Kohn (1994) and Frühwirth-Schnatter (1994), and used extensively, among others, by Kim and Nelson (1998). The parameter Σ_ϑ is chosen by the econometrician (See Robert and Casella (2004) for a discussion).

6.2 Mixture sampler

Steps (5-7) define a mixture sampler that generalizes Frühwirth-Schnatter's (2001) permutation sampler to permutation- and reflection-invariant linear state space models. Like the permutation sampler, the mixture sampler comes in two flavors. As used above, it allows to efficiently explore all symmetric modes of the posterior distribution. Alternatively, it can operationalize a normalization in the following manner.

As already mentioned, an observationally unrestrictive normalization consists in an element of a partition of the parameter space into $K!2^K$ observationally equivalent subspaces,

$$\Psi = \bigcup_{i=1}^{K!2^K} \Psi_i^N.$$

Assume one considers normalization $\Psi^N = \Psi_1^N$. In order to map Ψ onto Ψ_1^N, for $i = 1, \ldots, K!2^K - 1$, take $\mathbf{P}_i$ and $\mathbf{S}_i$ such that $T_{\mathbf{P}_i\mathbf{S}_i}(\psi) \in \Psi_1^N$ for $\psi \in \Psi_i^N$. Steps (5-6) are thus replaced by

(5a) Take $\mathbf{S} = \mathbf{S}_i$ such that $T_{\mathbf{S}_i}(\psi^*) \in \Psi^N$.
(6a) Take $\mathbf{P} = \mathbf{P}_i$ such that $T_{\mathbf{P}_i}(\psi^*) \in \Psi^N$.

In order to facilitate the interpretation of the parameters, one would like to find a normalization which yields unimodal parameter posterior distributions. Hamilton et al. (2007) show that normalizations satisfying the identification principle are more likely to yield such posteriors. The search can thus be restricted to normalizations satisfying the identification principle. However, there are uncountably many such normalizations. Because permutation and reflection normalizations can be implemented as a post-simulation step (Stephens, 1997; Geweke, 2007), a Bayesian analysis makes comparing a large number of normalizations computationally feasible. In contrast, one must obtain the sampling distribution of the ML estimator by simulations methods (See Stoffer and Wall (1991) for an application to linear state space models) in order to see whether a particular normalization produces unimodal sampling distributions.

7 Empirical results

In this section, I investigate the empirical role of error modeling. I use a panel of monthly sampled continuously-compounded discount rates from the Fama CRSP data files. Maturities are 1, 3, 12, 36 and 60 months, and the 204 observations of the curve run from January 1988 to December 2004. The 1- and 3-month rates are from the CRSP Risk Free Rates File and the longer maturities are from the Fama-Bliss Discount Bonds File. Discount bond rates were originally built from bootstrapping a

filtered set of observed coupon Treasuries and are used by Ang and Bekaert (2002), Dai et al. (2005) and Ang and Piazzesi (2003), among many others.

My benchmark model is the 3-latent-factor affine model with homoscedastic errors, which I label $A^L_{\Omega=\omega\mathcal{I}}$. I compare it to four alternatives (which I summarize in Table 2 for clarity): the 3-latent-factor affine model with heteroscedastic errors, $A^L_{\Omega=\text{diag}(\xi^{-1})}$; the 3-latent-factor affine model with heteroscedastic and correlated errors, $A^L_{\Omega=\mathbf{DRD}'}$; the 3-proxying-factor affine model with homoscedastic errors on the 3-month and 3-year rates ($A^P_{\Omega=\omega\mathcal{I}}$); and the 3-principal-component model (PC) [11]. I compare these models through the posterior distribution of several statistics of interest.

Because residuals are functions of the parameter vector, they are random vectors too. It is therefore possible to consider the posterior distribution of residual statistics, which I approximate using a sample from my posterior simulator. For each model, the posterior sampler runs for 500 000 iterations, of which I keep every 100th iteration to lighten some computations. For example, I obtain a posterior sample for the mean short-rate residual by computing

$$\bar{e}_1^{(m)} = \frac{1}{T}\sum_{t=1}^{T} y_t - \mathbf{A}(\psi^{(m)}) - \mathbf{B}(\psi^{(m)})' X_t^{(m)}$$

for $m = 1, \ldots, 50\,000$, while I compute the posterior median of $\bar{e}_1$ as

$$\text{median}(\bar{e}_1) = \arg\max_e \left\{ \frac{1}{50\,000} \sum_{m=1}^{50\,000} \mathbf{1}(\bar{e}_1^{(m)} < e) \leq \frac{1}{2} \right\}.$$

Tables in this section report the posterior median and 95%-inter-quantile credibility intervals for such statistics. For expositional briefness, I will say that a parameter is significant if its 95%-inter-quantile credibility interval does not include zero.

Table 2
Model notation

Model	Description
$A^L_{\Omega=\omega\mathcal{I}}$	3 latent factors; homoscedastic errors (Benchmark).
$A^L_{\Omega=\text{diag}(\xi^{-1})}$	3 latent factors; heteroscedastic errors.
$A^L_{\Omega=\mathbf{DRD}'}$	3 latent factors; heteroscedastic and correlated errors.
$A^P_{\Omega=\omega\mathcal{I}}$	3 proxying factors; homoscedastic errors.
PC	3 principal components.

[11] This model is presented in Appendix F for completeness.

7.1 Observational errors

Table 3 reports the 95-%-inter-quantile credibility intervals for pricing error and absolute error statistics for affine models, and sample statistics for the principal components model. In order to compare the benchmark affine model with homoscedastic errors $A^L_{\Omega=\omega\mathcal{I}}$ (Panel a) to model $A^L_{\Omega=\text{diag}(\xi^{-1})}$ (Panel b), I use a relatively uninformative prior on the dispersion of precisions ($\gamma^0_\Omega = 1.01$). Allowing for high heteroscedasticity reveals that the short rate is relatively mispriced by the economic model, with errors in the order of 10 basis points (bp) on average that exhibit a standard deviation of almost 40 bp, while the errors on other maturities are not significantly different from zero. Absolute errors confirm this pattern.

Because DTSMs are derived from an hypothesis on the short rate (see equation 2.7), this is of central concern. This hypothesis justifies the general use of the short rate as a proxying factor. It therefore seems that the short rate is badly "measured" in some way, compared to other maturities. Note that one obtains even larger pricing errors on the short rate from the Fama Treasury Bill Term Structure Files derived from 6-month Treasury Bills. One possible explanation is that the bootstrapping method used to extract the 1-month rate is bound to result in higher measurement errors than for other maturities. In the 6-month Fama Treasury Bill Term Structure Files, the maximum maturity mismatch is 4 days, which is more significant on the 1-month rate than on the 3-month rate. Short-rate residuals would also be larger than other maturities if there were an omitted short-rate-specific factor. For example, if short term instruments are held for liquidity reasons, then there would be some priced liquidity factor: investors would accept a return lower than the pure time-value return for the liquidity services provided by these instruments.

Comparing the proxy (Panel d) and latent-factor modeling approaches (Panel a), errors on $N-K$ yields are now distributed on N rates, and residuals are accordingly smaller in absolute terms (approximately 2 bp *versus* 3 bp) and are less variable (approximately 14 bp *versus* 17 bp). It is perhaps surprising that three rates from the $A^L_{\Omega=\text{diag}(\xi^{-1})}$ model (Panel b) have relatively small residuals. This might lead one to conclude that the proxy approach might not be too restrictive if one happens to pick the right rates to proxy latent factors. However, residual size is not necessarily the best metric to evaluate a model if the objective is extracting *common* factors from a panel of interest rates.

Introducing correlation in addition to a limited dispersion of precisions (Panel c) does not change the overall picture: the short rate still has larger and more variable residuals than longer rates. Note that a slightly more informative prior on precision dispersion, from $\gamma^0_\Omega = 1.01$ (Panel b) to $\gamma^0_\Omega = 5$ (Panel c), is sufficient to keep precisions within some common range. For example, the standard deviation of the 60-month residuals is 0.1 bp for $\gamma^0_\Omega = 1.01$ while all five standard deviations are between 12 and 26 bp for $\gamma^0_\Omega = 5$. Table 4 investigates this issue in more detail.

Table 3
Pricing errors (in basis points) statistics - covariance modeling.

	Maturity	1	3	12	36	60
Errors	Median	0.97 (−1.38, 3.31)	−2.37* (−4.67, −0.13)	2.22* (0.05, 4.38)	−1.19 (−3.44, 1.03)	0.46 (−1.82, 2.74)
	Mean	1.25 (−0.62, 3.10)	−2.60* (−4.34, −0.86)	2.19* (0.43, 3.93)	−1.25 (−3.07, 0.57)	0.43 (−1.45, 2.31)
	Std dev	14.64* (13.14, 16.28)	14.66* (13.29, 16.08)	12.78* (11.46, 14.17)	12.80* (11.49, 14.17)	12.79* (11.47, 14.17)
Abs errors	Median	9.62* (8.10, 11.29)	9.75* (8.26, 11.37)	8.75* (7.37, 10.24)	8.68* (7.32, 10.16)	8.64* (7.28, 10.15)
	Mean	11.57* (10.34, 12.93)	11.73* (10.58, 12.94)	10.36* (9.24, 11.52)	10.27* (9.17, 11.44)	10.22* (9.11, 11.39)
	Max	47.78* (36.49, 67.12)	48.33* (37.46, 65.56)	38.13* (30.69, 50.25)	37.85* (30.35, 49.99)	37.59* (30.20, 49.46)
	Std dev	9.05* (7.94, 10.29)	9.16* (8.14, 10.25)	7.81* (6.89, 8.82)	7.75* (6.84, 8.74)	7.71* (6.79, 8.70)
Panel a: $A^L_{\Omega=\omega\mathcal{I}}$						
Errors	Median	11.24* (5.81, 16.75)	−0.00 (−0.12, 0.10)	1.24 (−1.07, 3.68)	−0.07 (−1.37, 1.25)	0.00 (−0.02, 0.02)
	Mean	13.65* (9.57, 17.75)	−0.00 (−0.10, 0.08)	1.20 (−0.71, 3.20)	−0.15 (−1.21, 0.90)	−0.00 (−0.01, 0.02)
	Std dev	38.58* (34.94, 42.45)	0.57* (0.28, 1.08)	13.70* (11.99, 15.52)	7.71* (6.63, 8.70)	0.10* (0.08, 0.17)
Abs errors	Median	24.65* (20.82, 28.81)	0.38* (0.18, 0.74)	9.12* (7.56, 10.87)	5.08* (4.18, 6.06)	0.07* (0.05, 0.12)
	Mean	30.96* (27.89, 34.22)	0.45* (0.22, 0.86)	10.90* (9.48, 12.45)	6.11* (5.24, 6.93)	0.08* (0.06, 0.14)
	Max	162.64* (126.48, 208.30)	1.66* (0.78, 3.34)	42.64* (33.38, 57.17)	23.58* (18.43, 31.55)	0.29* (0.22, 0.53)
	Std dev	26.73* (23.78, 29.82)	0.34* (0.17, 0.65)	8.40* (7.24, 9.65)	4.71* (3.99, 5.41)	0.06* (0.05, 0.10)
Panel b: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$						
Errors	Median	−6.16* (−10.19, −1.77)	2.10 (−0.38, 4.51)	−1.52 (−4.18, 1.03)	−1.44 (−5.32, 2.11)	−0.87 (−4.61, 2.47)
	Mean	−9.10* (−12.69, −4.57)	2.19* (0.06, 4.41)	−1.45 (−3.64, 0.66)	−1.39 (−4.79, 1.40)	−1.52 (−4.79, 1.29)
	Std dev	25.94* (24.01, 28.25)	12.20* (10.30, 14.79)	13.49* (11.06, 16.79)	20.49* (15.98, 25.39)	20.09* (15.28, 24.84)
Abs errors	Median	14.30* (12.00, 17.03)	8.32* (6.72, 10.42)	9.07* (7.11, 11.47)	13.85* (10.33, 18.06)	13.71* (9.93, 17.85)
	Mean	19.22* (17.43, 21.43)	9.89* (8.33, 12.08)	10.77* (8.86, 13.37)	16.36* (12.72, 20.54)	16.12* (12.13, 20.15)
	Max	134.34* (113.95, 156.42)	37.51* (28.71, 51.48)	42.91* (30.95, 61.69)	63.01* (45.10, 86.83)	60.52* (42.58, 82.37)
	Std dev	19.64* (17.47, 21.85)	7.54* (6.26, 9.30)	8.29* (6.70, 10.54)	12.41* (9.60, 15.46)	12.13* (9.15, 15.04)
Panel c: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 50, \gamma^0_\Omega = 5$						

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

Principal components, introduced in the term structure literature by Litterman and Scheinkman (1991), are often presented as the standard benchmark and are indeed hard to beat in terms of various measures of error size, but the latent-factor modeling approach is definitely a serious competitor. Once one sets the short rate apart as a special mispriced rate, the models fare equally well with respect to most metrics. For example, the principal components model and the heteroscedastic model $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}$

Table 3
Pricing errors (in basis points) statistics - covariance modeling (Continued).

	Maturity	1	3	12	36	60
Errors	Median		-4.33^{*} $(-6.99, -1.61)$		-1.97 $(-4.83, 0.90)$	
	Mean		-4.44^{*} $(-6.47, -2.45)$		-2.13 $(-4.50, 0.36)$	
	Std dev		19.58^{*} $(18.02, 21.22)$		15.20^{*} $(13.73, 16.74)$	
Abs errors	Median		13.12^{*} $(11.31, 15.18)$		10.29^{*} $(8.72, 12.03)$	
	Mean		15.75^{*} $(14.40, 17.21)$		12.23^{*} $(10.99, 13.56)$	
	Max		71.58^{*} $(54.73, 95.55)$		46.20^{*} $(36.98, 61.44)$	
	Std dev		12.43^{*} $(11.16, 13.75)$		9.29^{*} $(8.25, 10.47)$	
Panel d: $A^{P}_{\Omega=\omega\mathcal{I}}$						
Errors	Median	-0.1	0.2	-0.0	-0.5	-0.3
	Std dev	4.9	9.2	4.9	5.6	4.5
Abs errors	Median	2.9	5.3	3.0	3.9	3.4
	Mean	3.7	7.0	3.7	4.6	3.7
	Max	21.4	41.1	22.7	18.0	11.9
	Std dev	3.1	5.9	3.2	3.3	2.6
Panel e: PC						

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

(Panel b) yield residuals and absolute residuals with medians and standard deviations of the same order.

Table 4 presents a sensibility analysis with respect to the prior precision dispersion parameter, γ^{0}_{Ω}, when errors are uncorrelated. As γ^{0}_{Ω} goes from 1.01 (Table 3, Panel b) to 2 (Panel a), the prior allows the standard deviation of the short rate to be singled out, as it gets more than twice as high as any other maturity. Also, the mean residual on the 3-month rate is now significantly different from zero. Increasing γ^{0}_{Ω} to 5 (Panel b) yield similar results: significant residual means for the short and 3-month rates, and a high short-rate residual standard deviation. But a further increase, to 50, somewhat changes the pattern: while the short-rate residual standard deviation is still higher than that of the other maturities, residual means are similar to those of the heteroscedastic model $A^{L}_{\Omega=\omega\mathcal{I}}$ (Panel a).

7.2 Cross-section properties

I next examine the correlations of pricing residuals. I consider posterior sample covariance matrices for models in which the covariance matrix is at most diagonal. Table 5 shows low but significant cross-correlations between some adjacent maturities for the benchmark homoscedastic model (Panel a). None of the correlations from the het-

Table 4
Pricing errors (in basis points) statistics - precision modeling.

	Maturity	1	3	12	36	60
Errors	Median	-4.46^* $(-7.80, -1.11)$	2.66^* $(0.86, 4.52)$	-1.47 $(-3.26, 0.32)$	0.57 $(-1.18, 2.38)$	-0.28 $(-2.13, 1.58)$
	Mean	-7.26^* $(-10.30, -4.13)$	2.75^* $(1.21, 4.33)$	-1.36 $(-2.87, 0.16)$	0.65 $(-0.85, 2.22)$	-0.23 $(-1.84, 1.36)$
	Std dev	24.12^* $(22.62, 25.64)$	10.25^* $(9.13, 11.43)$	9.32^* $(8.30, 10.42)$	9.31^* $(8.15, 10.74)$	9.31^* $(8.12, 10.81)$
Abs errors	Median	12.84^* $(10.93, 14.93)$	7.09^* $(5.96, 8.34)$	6.27^* $(5.23, 7.40)$	6.19^* $(5.11, 7.46)$	6.22^* $(5.09, 7.54)$
	Mean	17.45^* $(16.09, 18.87)$	8.45^* $(7.48, 9.49)$	7.47^* $(6.60, 8.44)$	7.40^* $(6.44, 8.59)$	7.41^* $(6.42, 8.64)$
	Max	124.52^* $(108.29, 142.26)$	31.27^* $(25.09, 40.46)$	29.71^* $(23.10, 40.07)$	28.89^* $(22.50, 38.72)$	28.40^* $(22.10, 38.58)$
	Std dev	18.15^* $(16.57, 19.91)$	6.43^* $(5.62, 7.35)$	5.75^* $(5.02, 6.53)$	5.71^* $(4.91, 6.69)$	5.68^* $(4.88, 6.69)$
Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 2$						
Errors	Median	-4.20^* $(-7.70, -0.59)$	2.73^* $(0.93, 4.57)$	-1.57 $(-3.44, 0.20)$	0.73 $(-1.12, 2.67)$	-0.32 $(-2.32, 1.63)$
	Mean	-6.87^* $(-10.21, -3.25)$	2.84^* $(1.34, 4.35)$	-1.45 $(-3.06, 0.06)$	0.75 $(-0.81, 2.44)$	-0.27 $(-1.99, 1.43)$
	Std dev	23.82^* $(22.22, 25.40)$	10.34^* $(9.12, 11.80)$	9.46^* $(8.30, 10.73)$	9.79^* $(8.34, 11.33)$	9.87^* $(8.20, 11.50)$
Abs errors	Median	12.76^* $(10.89, 14.88)$	7.17^* $(5.97, 8.56)$	6.37^* $(5.29, 7.63)$	6.49^* $(5.27, 7.90)$	6.57^* $(5.23, 8.02)$
	Mean	17.26^* $(15.89, 18.69)$	8.53^* $(7.47, 9.79)$	7.60^* $(6.61, 8.70)$	7.78^* $(6.60, 9.08)$	7.85^* $(6.49, 9.20)$
	Max	121.75^* $(103.44, 140.29)$	31.61^* $(25.23, 41.15)$	30.13^* $(23.24, 41.01)$	30.08^* $(23.26, 40.29)$	29.92^* $(22.59, 40.97)$
	Std dev	17.78^* $(15.90, 19.76)$	6.49^* $(5.62, 7.54)$	5.83^* $(5.03, 6.73)$	6.00^* $(5.04, 7.04)$	6.01^* $(4.92, 7.12)$
Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$						
Errors	Median	-2.41 $(-5.63, 0.74)$	3.07^* $(1.18, 4.91)$	-1.98^* $(-4.02, -0.01)$	1.08 $(-1.12, 3.39)$	-0.35 $(-2.71, 1.97)$
	Mean	-4.61^* $(-7.27, -1.77)$	3.26^* $(1.61, 4.80)$	-1.86^* $(-3.59, -0.16)$	1.08 $(-0.76, 2.92)$	-0.35 $(-2.41, 1.64)$
	Std dev	22.43^* $(20.95, 24.11)$	10.95^* $(9.78, 12.20)$	10.22^* $(9.17, 11.38)$	10.97^* $(9.79, 12.32)$	11.10^* $(9.76, 12.65)$
Abs errors	Median	12.23^* $(10.43, 14.45)$	7.62^* $(6.39, 8.87)$	6.98^* $(5.84, 8.27)$	7.38^* $(6.14, 8.76)$	7.43^* $(6.20, 8.78)$
	Mean	16.22^* $(14.90, 17.66)$	9.08^* $(8.09, 10.12)$	8.27^* $(7.38, 9.30)$	8.77^* $(7.76, 9.93)$	8.87^* $(7.77, 10.11)$
	Max	110.72^* $(94.81, 129.22)$	34.06^* $(27.29, 43.49)$	32.99^* $(25.94, 45.27)$	33.24^* $(26.76, 43.16)$	33.20^* $(26.57, 44.08)$
	Std dev	16.14^* $(14.84, 17.55)$	6.93^* $(6.13, 7.92)$	6.32^* $(5.58, 7.18)$	6.69^* $(5.87, 7.63)$	6.75^* $(5.86, 7.74)$
Panel c: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 50$						

Posterior medians of the median, mean, maximum and standard deviation of errors for the 1-, 3-, 12-, 36- and 60-month posterior discount rates. 95% credibility intervals are presented in parentheses and a * indicates that credibility interval does not include 0.

eroscedastic model are significant when precision dispersion is a priori high (Panel b), but lower dispersion yields significant correlations between adjacent maturities (Panel c). One could argue that *all* the correlations from the proxying-factor model (Panel d) are significantly different from zero, but as there is only one such correlation, that would arguably be abusive. A larger number of rates would be necessary to

reach any meaningful conclusion. Note that cross-correlations for all affine models are especially small compared with those from the principal components model (Panel e), which confirms that looking exclusively at measures of residual size does not tell the whole story.

That correlations decrease as one allows heteroscedasticity highlights the role of error modeling in the statistical decomposition of observables into common and idiosyncratic components. The observables' heteroscedasticity is thus partially captured by the idiosyncratic component, which allows factors to better capture other dimensions. Here, factors better capture correlations. In terms of prior specification, the main message from Table 5 is the following. While a very relatively uninformative prior precision dispersion ($\gamma_\Omega^0 = 1.01$) is compatible with uncorrelated errors, a slightly more informative prior ($\gamma_\Omega^0 = 5$) is not. Therefore, if an informative prior is used in order to keep precisions within some common range, then the error model should allow for some degree of correlation.

7.3 Time-series properties

Table 6 shows the posterior error autocorrelations and partial autocorrelations for all maturities (rows) and the first 3 orders (columns) for models presented in Table 3. Principal components (Panel d) do not model dynamics and PC consequently presents the worst performance. In spite of the fact that model $A^P_{\Omega=\text{diag}(\xi^{-1})}$ is a "dynamic term structure model", it produces residuals that are as autocorrelated (Panel c). In both cases, patterns in the coefficients suggest high-order ARMA structures. Latent-factor modeling of pricing errors (Panels a and b) seems more in line with the error model's assumption of time serial independence, although there still exists some residual dynamics. These results are consistent with the implied richer VARMA(1,1) representation of latent-factor models. Comparing the i.i.d error model (Panel a) to the more general model with heteroscedastic and correlated errors (Panel b), prior precision dispersion and correlation modeling is likely to affect the residual dynamics. I investigate these issues in Tables 7 and 8.

Table 5
Sample covariance of pricing errors.

14.66* (13.14 , 16.28)				
−0.15* (-0.27 , -0.02)	14.66* (13.29 , 16.08)			
0.04 (-0.10 , 0.17)	−0.01 (-0.14 , 0.13)	12.79* (11.46 , 14.17)		
0.08 (-0.06 , 0.21)	−0.14* (-0.27 , -0.01)	0.13 (-0.01 , 0.26)	12.82* (11.49 , 14.17)	
−0.04 (-0.18 , 0.10)	0.07 (-0.07 , 0.20)	−0.02 (-0.15 , 0.12)	0.11 (-0.03 , 0.24)	12.80* (11.47 , 14.17)

Panel a: $A^L_{\Omega=\omega\mathcal{I}}$

38.60* (34.94 , 42.45)				
0.00 (-0.13 , 0.14)	0.58* (0.28 , 1.08)			
−0.11 (-0.23 , 0.02)	0.00 (-0.14 , 0.14)	13.72* (11.99 , 15.52)		
0.07 (-0.05 , 0.20)	−0.00 (-0.14 , 0.14)	−0.02 (-0.14 , 0.11)	7.69* (6.63 , 8.70)	
−0.00 (-0.14 , 0.14)	−0.00 (-0.14 , 0.14)	0.00 (-0.14 , 0.14)	−0.00 (-0.14 , 0.14)	0.11* (0.08 , 0.17)

Panel b: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$

23.82* (22.22 , 25.40)				
0.18* (0.07 , 0.29)	10.37* (9.12 , 11.80)			
−0.12* (-0.23 , -0.00)	0.26* (0.13 , 0.39)	9.47* (8.30 , 10.73)		
0.08 (-0.03 , 0.19)	−0.13 (-0.26 , 0.01)	0.22* (0.07 , 0.37)	9.79* (8.34 , 11.33)	
−0.02 (-0.14 , 0.10)	−0.08 (-0.21 , 0.06)	0.10 (-0.04 , 0.24)	0.60* (0.48 , 0.71)	9.84* (8.20 , 11.50)

Panel c: $A^L_{\Omega=\mathrm{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

19.59* (18.02 , 21.22)	
−0.15* (-0.27 , -0.03)	15.21* (13.73 , 16.74)

Panel d: $A^P_{\Omega=\omega\mathcal{I}}$

4.85				
-1.00	9.19			
0.90	-0.91	4.89		
0.57	-0.53	0.14	5.63	
-0.75	0.73	-0.38	-0.97	4.52

Panel e: PC

Posterior medians and 95%-inter-quantile credibility intervals for the standard-deviation (diagonal, in basis points) and correlations of pricing errors. A * indicates that credibility interval does not include 0

Table 6
Sample autocorrelations and partial autocorrelations of pricing errors.

	Autocorrelation		
Order	1	2	3
e_1	0.05 $(-0.07, 0.18)$	0.07 $(-0.06, 0.19)$	-0.01 $(-0.14, 0.11)$
e_3	0.39^* $(0.30, 0.47)$	0.24^* $(0.15, 0.34)$	0.20^* $(0.10, 0.30)$
e_{12}	0.23^* $(0.10, 0.36)$	0.08 $(-0.05, 0.22)$	0.09 $(-0.06, 0.21)$
e_{36}	0.38^* $(0.26, 0.48)$	0.27^* $(0.15, 0.38)$	0.22^* $(0.10, 0.34)$
e_{60}	0.24^* $(0.10, 0.37)$	0.15 $(-0.00, 0.29)$	0.16^* $(0.01, 0.31)$

Panel a: $A^L_{\Omega=\omega\mathcal{I}}$

	Autocorrelation		
Order	1	2	3
e_1	0.27^* $(0.18, 0.35)$	0.11^* $(0.02, 0.20)$	-0.06 $(-0.15, 0.03)$
e_3	0.27^* $(0.13, 0.42)$	0.10 $(-0.04, 0.24)$	0.07 $(-0.06, 0.21)$
e_{12}	0.23^* $(0.09, 0.36)$	0.00 $(-0.13, 0.13)$	0.01 $(-0.13, 0.15)$
e_{36}	0.42^* $(0.28, 0.54)$	0.13 $(-0.01, 0.26)$	0.05 $(-0.06, 0.18)$
e_{60}	0.36^* $(0.17, 0.50)$	0.05 $(-0.11, 0.20)$	0.02 $(-0.09, 0.15)$

Panel b: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 50, \gamma^0_\Omega = 5$

	Autocorrelation		
Order	1	2	3
e_3	0.46^* $(0.42, 0.51)$	0.33^* $(0.28, 0.39)$	0.25^* $(0.19, 0.31)$
e_{36}	0.72^* $(0.71, 0.75)$	0.55^* $(0.52, 0.60)$	0.44^* $(0.40, 0.51)$

Panel c: $A^O_{\Omega=\omega\mathcal{I}}$

	Autocorrelation		
Order	1	2	3
e_1	0.46	0.33	0.24
e_3	0.44	0.32	0.23
e_{12}	0.34	0.23	0.17
e_{36}	0.74	0.57	0.48
e_{60}	0.71	0.55	0.44

Panel d: PC

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

Table 7 presents a sensibility analysis with respect to correlation prior specification. Comparing all models, it seems that correlations do not affect residual dynamics much when errors are heteroscedastic: there is little residual dynamics when errors are uncorrelated (Panel a). However, first-order partial autocorrelations do increase as prior correlations are less informative.

Table 7
Sample autocorrelations and partial autocorrelations of pricing errors - Correlation modeling.

	Autocorrelation		
Order	1	2	3
e_1	0.24^* $(0.15, 0.32)$	0.09^* $(0.01, 0.17)$	-0.09 $(-0.16, 0.00)$
e_3	0.25^* $(0.11, 0.38)$	0.10 $(-0.04, 0.23)$	0.10 $(-0.03, 0.23)$
e_{12}	0.27^* $(0.14, 0.40)$	0.10 $(-0.04, 0.23)$	0.07 $(-0.07, 0.20)$
e_{36}	0.25^* $(0.13, 0.36)$	0.11 $(-0.02, 0.23)$	0.10 $(-0.02, 0.23)$
e_{60}	0.15^* $(0.01, 0.28)$	0.01 $(-0.13, 0.16)$	0.09 $(-0.04, 0.23)$

Panel a: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

	Autocorrelation		
Order	1	2	3
e_1	0.25^* $(0.17, 0.34)$	0.10^* $(0.01, 0.19)$	-0.08 $(-0.16, 0.01)$
e_3	0.25^* $(0.11, 0.38)$	0.09 $(-0.05, 0.23)$	0.08 $(-0.05, 0.21)$
e_{12}	0.20^* $(0.07, 0.33)$	0.02 $(-0.11, 0.16)$	0.03 $(-0.10, 0.17)$
e_{36}	0.24^* $(0.12, 0.36)$	0.03 $(-0.09, 0.15)$	0.03 $(-0.09, 0.16)$
e_{60}	0.14^* $(0.01, 0.26)$	-0.07 $(-0.19, 0.05)$	0.02 $(-0.10, 0.14)$

Panel b: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 250, \gamma^0_\Omega = 5$

	Autocorrelation		
Order	1	2	3
e_1	0.25^* $(0.16, 0.33)$	0.08 $(-0.01, 0.17)$	-0.10^* $(-0.19, -0.01)$
e_3	0.34^* $(0.20, 0.46)$	0.14 $(-0.01, 0.28)$	0.09 $(-0.06, 0.23)$
e_{12}	0.31^* $(0.19, 0.42)$	0.02 $(-0.10, 0.15)$	-0.03 $(-0.15, 0.10)$
e_{36}	0.56^* $(0.47, 0.63)$	0.26^* $(0.14, 0.37)$	0.13^* $(0.02, 0.25)$
e_{60}	0.53^* $(0.43, 0.62)$	0.22^* $(0.09, 0.35)$	0.12 $(-0.00, 0.25)$

Panel c: $A^L_{\Omega=\mathbf{DRD}'}, \tau^0_\Omega = 6, \gamma^0_\Omega = 5$

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

Table 8 considers the impact of precision dispersion prior specification on residual dynamics. An uninformative prior (Panel a) singles out two rates: the residuals on the 3-month and 60-month rates look serially independent. This suggests that two factors are closely associated with these maturities. This is confirmed by small mean residuals (Table 3, Panel b). In contrast, more informative priors (Panels b and c) yield residuals that have more similar dynamics across maturities. In the extreme case of homoscedastic errors (Table 6, Panel a), the short rate is singled out as a factor, which leaves no residual dynamics. However, other maturities have significant high-order residual dynamics, which further suggests that this factor contains infor-

mation that is specific to the short rate.

Table 8
Sample autocorrelations and partial autocorrelations of pricing errors - Precision modeling.

	Autocorrelation		
Order	1	2	3
e_1	0.17^* $(0.04, 0.29)$	0.10 $(-0.02, 0.22)$	0.03 $(-0.09, 0.16)$
e_3	-0.01 $(-0.14, 0.13)$	-0.00 $(-0.14, 0.13)$	-0.00 $(-0.14, 0.13)$
e_{12}	0.20^* $(0.07, 0.33)$	0.11 $(-0.03, 0.24)$	0.10 $(-0.03, 0.24)$
e_{36}	0.19^* $(0.05, 0.32)$	0.10 $(-0.03, 0.24)$	0.12 $(-0.01, 0.26)$
e_{60}	-0.00 $(-0.14, 0.13)$	-0.01 $(-0.14, 0.13)$	-0.00 $(-0.14, 0.13)$

Panel a: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 1.01$

	Autocorrelation		
Order	1	2	3
e_1	0.24^* $(0.15, 0.32)$	0.09^* $(0.01, 0.17)$	-0.09 $(-0.16, 0.00)$
e_3	0.25^* $(0.11, 0.38)$	0.10 $(-0.04, 0.23)$	0.10 $(-0.03, 0.23)$
e_{12}	0.27^* $(0.14, 0.40)$	0.10 $(-0.04, 0.23)$	0.07 $(-0.07, 0.20)$
e_{36}	0.25^* $(0.13, 0.36)$	0.11 $(-0.02, 0.23)$	0.10 $(-0.02, 0.23)$
e_{60}	0.15^* $(0.01, 0.28)$	0.01 $(-0.13, 0.16)$	0.09 $(-0.04, 0.23)$

Panel b: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 5$

	Autocorrelation		
Order	1	2	3
e_1	0.18^* $(0.11, 0.27)$	0.06 $(-0.02, 0.13)$	-0.11^* $(-0.18, -0.03)$
e_3	0.30^* $(0.19, 0.43)$	0.15^* $(0.01, 0.27)$	0.13 $(-0.00, 0.26)$
e_{12}	0.28^* $(0.16, 0.40)$	0.10 $(-0.03, 0.22)$	0.07 $(-0.06, 0.20)$
e_{36}	0.24^* $(0.13, 0.35)$	0.08 $(-0.04, 0.21)$	0.09 $(-0.03, 0.21)$
e_{60}	0.17^* $(0.03, 0.29)$	0.01 $(-0.14, 0.15)$	0.09 $(-0.04, 0.24)$

Panel c: $A^L_{\Omega=\text{diag}(\xi^{-1})}, \gamma^0_\Omega = 50$

Posterior medians and 95%-inter-quantile credibility intervals of pricing errors sample autocorrelations (right panels) and partial autocorrelations (left panels) for the first three orders (columns) in the 3-factor models, for each maturity (rows). A * indicates that credibility interval does not include 0.

While this clearly merits further investigation, with perhaps a larger number of rates, it seems that a general error model with relatively informative priors can help the econometrician affect the decomposition of observables into common components and idiosyncratic errors. In particular, for the data set considered here, hyperparameter values $\tau_\Omega^0 = 50$ and $\gamma_\Omega^0 = 5$ yield residuals that are roughly consistent with the error model, although they still leave some residual dynamics. More generally, the following facts emerge:

(1) Low prior correlations are inconsistent with residual correlations when prior heteroscedasticity is low (Table 5, Panel c);
(2) Less informative correlation priors increase first-order residual partial autocorrelations (Table 7, Panel c);
(3) Low prior heteroscedasticity increases residual dynamics for all maturities except the short rate (Table 8, Panels a and c).
(4) Less informative precision dispersion priors yield rate-specific factors (Table 3, Panel b);

8 Concluding remarks

Modeling observational errors on all discount rates is desirable on both theoretical and empirical grounds, and is computationally feasible. Assuming that some rates are observed without error is observationally restrictive and yields residuals with high variances, cross-correlations and autocorrelations. Because factor models decompose observable dynamics into common and idiosyncratic components, error modeling is not inferentially innocuous. Extreme error modeling choices illustrate the relevant issues: the likelihood function is singular if all rates are observed without error, while the model is globally unidentified if the error dynamics are as rich as those of the factors.

Between the extremes, the econometrician has considerable room for modeling common and idiosyncratic components. For example, modeling heteroscedastic errors highlights that some rates are better proxying factor candidates than others. Because these errors capture part of observable heteroscedasticity, factors can better capture cross-correlations and autocorrelation. However, modeling heteroscedastic errors shares some drawbacks with the proxying-factor approach: factors better describe some discount rates at the expense of others. I show how an informative heteroscedasticity prior specification mitigates this problem and yield factors describing features that are common to the entire panel of discount rate rather than a small subset thereof. In addition, modeling low cross-correlations through an informative prior helps further reduce residuals autocorrelations.

Inference for affine models is complicated by weak identification problems, which make the Bayesian methodology particularly appealing for at least two reasons. Be-

cause one may have to evaluate a large number of normalizations before one that yields estimators with good finite-sample properties is found, the fact that ML estimator sampling distributions must be obtained by simulations methods makes this approach computationally prohibitive. In contrast, normalizations can be compared at little computational cost using a sample from the un-normalized posterior distribution. Moreover, Bayesian inference for observational errors does not rely on biased parameter point estimators and therefore provides valid diagnostics of model adequacy. These computational and inferential considerations make the proposed methodology an ideal candidate for empirical macroeconomic work, especially with relatively small data sets, of the order of a few hundred months or quarters.

I leave a number of important empirical questions unanswered. How binding are parameter restrictions that yield factors with level, slope and curvature interpretations? Does the role of error modeling changes as one observes a larger number of maturities? As for the modeling of factors, the scale normalization and short rate factor loadings parameterization proposed in this paper yields a parameter, σ, which can be interpreted as the factors's common variance. This parameterization lends itself to the specification of a simple stochastic volatility model with a single common volatility factor.

References

Ang, A., Bekaert, G., 2002. Regime switches in interest rates. Journal of Business and Economic Statistics 20, 163–182.

Ang, A., Dong, S., Piazzesi, M., 2007. No-arbitrage Taylor rules, Working paper, Columbia University and University of Chicago.

Ang, A., Piazzesi, M., 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. Journal of Monetary Economics 50(4), 745–787.

Babbs, S., Nowman, K., 1999. Kalman filtering of generalized Vasicek term structure models. Journal of Financial and Quantitative Analysis 34 (1), 115–130.

Ball, C., Torous, W., 1996. Unit roots and the estimation of interest rate dynamics. Journal of Empirical Finance 3, 215–238.

Barnard, J., McCulloch, R., Meng, X.-L., 2000. Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. Statistica Sinica 10, 1281–1311.

Bekaert, G., Cho, S., Moreno, A., June 2006. New-Keynesian macroeconomics and the term structure, Working paper, Graduate School of Business, Columbia University.

Bertholon, H., Monfort, A., Pegoraro, F., 2007. Econometric asset pricing modelling, CREST DP.

Blais, S., 2008. Forecasting with wealkly identified linear state space models, Working paper, Université de Montréal.

Bliss, R. R., 1997. Testing term structure estimation methods. Advances in Futures

and Options Research 9, 197–231.
Bound, J., Jaeger, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlations betweeen the instruments and the endogenous explanatory variables is weak. Journal of the American Statistical Association 90, 443–450.
Box, G., Jenkins, G., Reinsel, G., 1994. Time Series Analysis, 3rd Edition. Prentice Hall.
Buse, A., 1992. The bias of instrumental variables estimators. Econometrica 60, 173–180.
Carter, C., Kohn, P., 1994. On the Gibbs sampling for state space models. Biometrika 81, 541–553.
Celeux, G., Hurn, M., Robert, C., 2000. Computational and inferential difficulties with mixture prosterior distribution. Journal of the American Statistical Association 95 (451), 957–970.
Chen, R., Scott, L., 1993. Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. Journal of Fixed Income 3, 14–31.
Chen, R., Scott, L., 1995. Multi-factor Cox-Ingersoll-Ross models of the term-structure: estimates and tests from a Kalman filter model., Working paper, University of Georgia.
Cheridito, P., Filipović, D., Kimmel, R., December 2003. Market price of risk specifications for affine models: Theory and evidence, Princeton University.
Chib, S., Ergashev, B., September 2008. Analysis of multi-factor affine yield curve models, wroking paper, Washington University in St. Louis and the Federal Reserve Bank of Richmond.
Christensen, J. H. E., Diebold, F. X., Rudebusch, G. D., 2009. The affine arbitrage-free class of Nelson-Siegel term structure models, working paper, University of Pennsylvania and Federal Reserve Bank of San Francisco.
Dai, Q., Le, A., Singleton, K., March 2006. Discrete-time dynamic term structure models with generalized market prices of risk, Working paper, Graduate School of Business, Stanford University.
Dai, Q., Singleton, K., 2000. Specification analysis of affine term structure models. 55 Journal of Finance, 1943–1978.
Dai, Q., Singleton, K., 2002. Expectation puzzles, time-varying risk premia, and affine models of the term structure. Journal of Financial Economics 63, 415–441.
Dai, Q., Singleton, K., 2003. Term structure dynamics in theory and reality. Review of Financial Studies 16, 361–678.
Dai, Q., Singleton, K., Yang, W., 2005. Are regime shifts priced in U.S. Treasury markets?, Working paper, New York University.
Dewachter, H., Lyrio, M., 2006. Learning, macroeconomic dynamics and the term structure of interest rates, Working paper, Catholic University of Leuven.
Duarte, J., 2003. Evaluating an alternative risk preference in affine term structure models, financeLab, Working paper FLWP-2003-2.
Duffee, G., 2002. Term premia and interest rate forecasts in affine models. Journal of Finance 57, 405–443.

Dufour, J.-M., 1997. Some impossibility theorems in econometrics, with applications to structural and dynamic models. Econometrica 65, 1365–1389.

Dufour, J.-M., Hsiao, C., 2008. "Identification", The New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan.

Evans, M., 2003. Real risk, inflation risk, and the term structure. Economic Journal 113(487), 345–389.

Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. Journal of Time Series Analysis 15, 183–202.

Frühwirth-Schnatter, S., 2001. Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194–205.

Frühwirth-Schnatter, S., 2004. Efficient Bayesian parameter estimation. In: Harvey, A., Koopman, S. J., Shephard, N. (Eds.), State Space and Unobserved Component Models: Theory and Applications. Cambridge University Press, pp. 123–151.

Frühwirth-Schnatter, S., Geyer, A., 1998. Bayesian estimation of econometric multifactor Cox-Ingersoll-Ross models of the term structure of interest rates via MCMC methods, Working paper, Vienna University of Economics and Business Administration.

Garcia, R., Luger, R., February 2007. Risk aversion, intertemporal substitution, and the term structure of interest rates., Working paper, Université de Montréal and Emory University.

Geweke, J., 2007. Interpretation and inference in mixture models: Simple MCMC works. Computational Statistics & Data Analysis 51, 3529–3550.

Geyer, A., Pichler, S., 1999. A state-space approach to estimate and test multifactor Cox-Ingersoll-Ross models of the term structure. Journal of Financial Research 22 (1), 107–130.

Gouriéroux, C., Monfort, A., Polimenis, V., 2002. Affine term structure models, Working paper, CREST.

Hamilton, J., Waggoner, D., Zha, T., 2007. Normalization in econometrics. Econometric Reviews 26, 221 – 252.

Hiller, G. H., 1990. On the normalization of structural equations: properties of direction estimators. Econometrica 58 (5), 1181–1194.

Hördahl, P., Tristani, O., Vestin, D., 2006. A joint econometric model of macroeconomic and term-structure dynamics. Journal of Econometrics 131, 405–444.

Jegadeesh, N., Pennacchi, G., 1996. The behavior of interest rates implied by the term structure of eurodollar futures. Journal of Money, Credit and Banking 28 (3), 426–446.

Kim, C., Nelson, C., 1998. Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. The Review of Economics and Statistics 80, 188–201.

Lamoureux, C., Witte, H., 2002. Empirical analysis of the yield curve: The information in the data viewed through the window of Cox, Ingersol and Ross. The Journal of Finance 57, 1479–1520.

Litterman, R., Scheinkman, J., 1991. Common factors affecting bond returns. Journal of Fixed Income, 54–61.

McCulloch, J., 1975. The tax-adjusted yield curve. Journal of Finance 30, 811–830.
Monfort, A., Pegoraro, F., 2007. Switching VARMA term structure models. Journal of Financial Econometrics 51 (1), 105–153.
Müller, P., Polson, N., Stroud, J., 2003. Nonlinear state-space models with state-dependent variances. Journal of the American Statistical Association 98, 377–386.
Nelson, C. R., Startz, R., 1990. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. Journal of Business 63 (S125-S140).
Redner, R. A., Walker, H. F., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26 (2), 195–239.
Robert, C. P., Casella, G., 2004. Monte Carlo Statistical Methods, 2nd Edition. Springer.
Rothenberg, T. J., May 1971. Identification in parametric models. Econometrica 39 (3), 577–591.
Sanford, A., Martin, G., 2005. Simulation-based Bayesian estimation of affine term structure models. Computational Statistics and Data Analysis, Special Issue on Computational Econometrics 2 49, 527–554.
Stephens, M., 1997. Bayesian methods for mixtures of normal distributions. Ph.D. thesis, University of Oxford.
Stephens, M., 2000. Dealing with label switching in mixture models. Journal of the Royal Statistical Society B 62, 795–809, part 4.
Stoffer, D. S., Wall, K. D., 1991. Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. Journal of the American Statistical Association 86 (416), 1024–1033.

A Prior distribution hyper-parameters

Table A.1
Prior distribution parameters for the 3-factor models.

Parameter	Value
$\mu^0_{A_1}$	4e-003
$\Sigma^0_{A_1}$	1e-005
$\mu^0_{\sigma_0}$	-5
$\Sigma^0_{\sigma_0}$	1e+003
$\mu^0_{\lambda_0}$	0
$\Sigma^0_{\lambda_0}$	100
μ^0_{γ}	0
Σ^0_{γ}	5
$\mu^0_{\kappa^{\mathbb{P}}}$	0
$\Sigma^0_{\kappa^{\mathbb{P}}}$	5
τ^0_{Σ}	4
γ^0_{Ω}	1.01, 2, 5, 10
τ^0_{Ω}	6, 250 ,1000
ν^0_{Ω}	2
β^0_{Ω}	1e+009

B Solution to the pricing difference equation

The solution to the pricing difference equation (2.6) is due to Ang and Piazzesi (2003) and is provided here for completeness. Assume the solution is of the form $P_{n,t} = \exp\{\tilde{A}_n + \tilde{B}_n' X_t\}$,

$$\begin{aligned} P_{n,t} &= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}_1' X_t\} \mathbf{E}_t^{\mathbb{Q}}[P_{n-1,t+1}] \\ \exp\{\tilde{A}_n + \tilde{\mathbf{B}}_n' X_t\} &= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}_1' X_t\} \mathbf{E}_t^{\mathbb{Q}}[\exp\{\tilde{A}_{n-1} + \tilde{\mathbf{B}}_{n-1}' X_{t+1}\}] \\ &= \exp\{\tilde{A}_1 + \tilde{\mathbf{B}}_1' X_t\} \\ &\quad \times \exp\{\tilde{A}_{n-1} + \tilde{\mathbf{B}}_{n-1}' \left(X_t + \kappa^{\mathbb{Q}}(\theta^{\mathbb{Q}} - X_t)\right) + \frac{1}{2} \tilde{\mathbf{B}}_{n-1}' \Sigma \tilde{\mathbf{B}}_{n-1}, \end{aligned}$$

and match the coefficients to get the recursions (2.9).

C VARMA-representation of yields

To simplify exposition, consider an N-factor model, where there are just as many latent factors as there are observed yields. The K-factor model, with $K < N$ can then viewed as a constrained N-factor model. I first rewrite (2.12) with $\nu_t \equiv \Sigma^{1/2} \epsilon_t$

$$X_{t+1} = X_t + \kappa^{\mathbb{P}}(\theta^{\mathbb{P}} - X_t) + \nu_{t+1}$$

and (3.1) with $\zeta_t \equiv \Omega^{1/2} u_t$

$$y_t = \mathbf{A} + \mathbf{B}' X_t + \zeta_t.$$

When yields are observed without any measurement error and the model is assumed to be perfect, one can inverse the pricing equations to solve for the yields and obtain

$$y_{t+1} = (I - \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1})\mathbf{A} + \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1} y_t + \mathbf{B}' \nu_{t+1}$$

When all yields are subject to measurement with errors or when the model describes reality imperfectly, one obtains the same VAR(1) process but with measurement errors in the variables

$$(y_{t+1} - \zeta_{t+1}) = (I - \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1})\mathbf{A} + \mathbf{B}'(I - \kappa^{\mathbb{P}})\mathbf{B}'^{-1}(y_t - \zeta_t) + \mathbf{B}' \nu_{t+1}$$

Such a process is equivalent to a VARMA(1,1) (Box, Jenkins, and Reinsel, 1994).

D From physical drift to risk-neutral drift in a conditionally Gaussian model with log-linear SDF

This proof is based on Dai, Singleton, and Yang (2005). Since the price of a *any* cash flow c_{t+1} can be calculated under both measure, i.e.

$$\mathbf{E}^{\mathbb{P}}_{X_{t+1}|X_t}\left[\mathrm{e}^{m_{t+1}}c_{t+1}\right] = \mathrm{e}^{-y_{1,t}}\mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[c_{t+1}\right],$$

we can identify the risk-neutral measure,

$$\mathrm{d}\mathbb{Q} = \mathrm{e}^{y_{1,t}+m_{t+1}}\mathrm{d}\mathbb{P}.$$

We can then compute the risk-neutral trend,

$$\begin{aligned}
\mu_t^{\mathbb{Q}} \equiv & \mathbf{E}^{\mathbb{Q}}_{X_{t+1}|X_t}\left[X_{t+1}\right] \\
= & \int_{\mathbb{R}^K} X_{t+1}\exp\left\{-\Lambda_t\left(X_{t+1}-\mu_t^{\mathbb{P}}\right)-\frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\}\mathrm{d}\mathbb{P} \\
= & \exp\left\{\Lambda_t\mu_t^{\mathbb{P}}-\frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
& \times\int_{\mathbb{R}^K} X_{t+1}\mathrm{e}^{-\Lambda_t X_{t+1}}\,\mathrm{d}\mathbb{P} \\
= & \exp\left\{\Lambda_t\mu_t^{\mathbb{P}}-\frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
& \times\frac{\partial}{\partial\Lambda_t}-\int_{\mathbb{R}^K}\mathrm{e}^{-\Lambda_t X_{t+1}}\,\mathrm{d}\mathbb{P} \\
= & -\exp\left\{\Lambda_t\mu_t^{\mathbb{P}}-\frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
& \times\frac{\partial}{\partial\Lambda_t}\exp\left\{-\Lambda_t\mu_t^{\mathbb{P}}+\frac{1}{2}\Lambda_t\Sigma_t\Lambda_t'\right\} \\
= & \mu_t^{\mathbb{P}}-\Sigma_t\Lambda_t'. \qquad \text{(D.1)}
\end{aligned}$$

E Inverse-Gamma-mixture of Gamma densities

Our hierarchal prior for $\xi \equiv \text{diag}\,(\Omega^{-1})$ is a Inverse-Gamma-mixture of Gamma densities. Specifically,

$$p(\xi|\gamma,\nu,\beta) = \int_0^\infty \prod_{n=1}^N G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right) IG\left(\eta|\nu,\beta\right)\,\mathrm{d}\eta$$

with

$$G\left(\xi_n|\gamma,\frac{\eta}{\gamma}\right) = \frac{(\frac{\gamma}{\eta})^\gamma}{\Gamma(\gamma)}\Xi_n^{\gamma-1}\exp\left(-\frac{\gamma}{\eta}\Xi_n\right)$$

and

$$IG\left(\eta|\nu,\beta\right) = \frac{\beta^\nu}{\Gamma(\nu)}\frac{\exp\left(-\beta/\eta\right)}{\eta^{\nu+1}}.$$

One can write this mixture in closed form as

$$p(\xi|\gamma,\nu,\beta) = \frac{p(\Xi,\eta|\gamma,\nu,\beta)}{p(\eta|\Xi,\nu,\beta)}$$

since

$$\begin{aligned} p(\eta|\Xi,\gamma,\nu,\beta) &\propto p(\Xi|\gamma,\eta)p(\eta|\nu,\beta) \\ &\propto \left(\frac{\gamma}{\eta}\right)^{N\gamma}\exp\left(-\frac{\gamma}{\eta}\sum_{n=1}^N\Xi_n\right)\frac{\exp\left(-\beta/\eta\right)}{\eta^{\nu+1}} \\ &\propto IG\Big(\eta\Big|N\gamma+\nu,\beta+\gamma\sum_{n=1}^N\Xi_n\Big). \end{aligned}$$

Explicitly,

$$p(\xi|\gamma,\nu,\beta) = \frac{\gamma^{N\gamma}\beta^\nu\Gamma(N\gamma+\nu)}{\Gamma(\nu)\Gamma(\gamma)^N}\frac{\prod_{n=1}^N\xi_n^{\gamma-1}}{\left(\beta+\gamma\sum_{n=1}^N\xi_n\right)^{N\gamma+\nu}}$$

The mean of ξ_n is

$$
\begin{aligned}
\mathbf{E}[\xi_n] &= \int_0^\infty \xi_n \left[\int_0^\infty G\left(\xi_n|\gamma, \frac{\eta}{\gamma}\right) IG\left(\eta|\nu,\beta\right) \, \mathrm{d}\eta \right] \, \mathrm{d}\xi_n \\
&= \int_0^\infty \left[\int_0^\infty \xi_n G\left(\xi_n|\gamma, \frac{\eta}{\gamma}\right) \, \mathrm{d}\xi_n \right] IG\left(\eta|\nu,\beta\right) \, \mathrm{d}\eta \\
&= \int_0^\infty \eta IG\left(\eta|\nu,\beta\right) \, \mathrm{d}\eta \\
&= \frac{\beta}{\nu - 1}
\end{aligned}
$$

Note that this prior is conditionally conjugate in a Gaussian model.

F Principal components

To build orthogonal factors, one takes the singular-value decomposition of the yield sample covariance matrix

$$\hat{\Sigma}_y^2 = \delta\gamma\delta'.$$

For $K < N$ principal components, consider the K first columns of δ, call it δ_K and take

$$\hat{y}_{PC} = \bar{y} + (y - \bar{y})\delta_K\delta_K'$$

where $\bar{y}$ is the sample mean.