NBER WORKING PAPER SERIES

ESTIMATION OF AFFINE TERM STRUCTURE MODELS WITH SPANNED OR UNSPANNED STOCHASTIC VOLATILITY

Drew D. Creal
Jing Cynthia Wu

Working Paper 20115
http://www.nber.org/papers/w20115

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2014

We thank Yacine Ait-Sahalia, Boragan Aruoba, Michael Bauer, Alan Bester, John Cochrane, Frank Diebold, Rob Engle, Jim Hamilton, Chris Hansen, Guido Kuersteiner, Ken Singleton, two anonymous referees, and seminar and conference participants at Chicago Booth, NYU Stern, NBER Summer Institute, Maryland, Bank of Canada, Kansas, UMass, and Chicago Booth Junior Finance Symposium for helpful comments. Drew Creal thanks the William Ladany Faculty Scholar Fund at the University of Chicago Booth School of Business for financial support. Cynthia Wu also gratefully acknowledges financial support from the IBM Faculty Research Fund at the University of Chicago Booth School of Business. This paper was formerly titled "Estimation of non-Gaussian affine term structure models." The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Estimation of Affine Term Structure Models with Spanned or Unspanned Stochastic Volatility
Drew D. Creal and Jing Cynthia Wu
NBER Working Paper No. 20115
May 2014, Revised October 2014
JEL No. C13,E43,G12

ABSTRACT

We develop new procedures for maximum likelihood estimation of affine term structure models with spanned or unspanned stochastic volatility. Our approach uses linear regression to reduce the dimension of the numerical optimization problem yet it produces the same estimator as maximizing the likelihood. It improves the numerical behavior of estimation by eliminating parameters from the objective function that cause problems for conventional methods. We find that spanned models capture the cross-section of yields well but not volatility while unspanned models fit volatility at the expense of fitting the cross-section.

Drew D. Creal
University of Chicago
Booth School of Business
5807 South Woodlawn Ave
Chicago, IL 60637
dcreal@chicagobooth.edu

Jing Cynthia Wu
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
cynthia.wu@chicagobooth.edu

1 Introduction

We propose new estimation procedures for affine term structure models (ATSMs) with spanned or unspanned stochastic volatility that use linear regression to simplify and stabilize estimation. For spanned models, our procedure recovers the maximum likelihood estimator but only requires numerically optimizing over a lower dimensional parameter space. The stability of our method makes it possible for us to study local maxima, explain why they exist, and their economic implications. We show how our insights from spanned models can be extended to estimate unspanned stochastic volatility (USV) models despite the fact that for USV models the likelihood function is not known in closed-form. Estimating a range of popular models, we find that models with spanned volatility fit the cross section of the yield curve better, while those with unspanned volatility fit the volatility better.

ATSMs are popular among policy makers, practitioners, and academic researchers for studying bond prices, monetary policy, and the macroeconomic determinants of discount rates; for overviews, see Piazzesi(2010), Duffee(2012), Gürkaynak and Wright(2012), and Diebold and Rudebusch(2013). As the literature on ATSMs has developed over the last decade, there is a consensus that estimation can be challenging; see, e.g. Duffee(2002), Ang and Piazzesi(2003), Kim and Orphanides(2005), and Hamilton and Wu(2012). New procedures for Gaussian ATSMs have made them easier to estimate, further increasing their popularity; see, Joslin, Singleton, and Zhu(2011), Christensen, Diebold, and Rudebusch(2011), Hamilton and Wu(2012), Adrian, Crump, and Moench(2012) and Diez de Los Rios(2013). However, these procedures do not address models with stochastic volatility. Moreover, in USV models as proposed by Collin-Dufresne and Goldstein(2002) and Collin-Dufresne, Goldstein, and Jones(2009), the likelihood function is not known in closed-form. Potential solutions to this problem are the closed-form expansions of the likelihood for continuous-time models developed by Aït-Sahalia(2008) and Aït-Sahalia and Kimmel(2010) and the expectation maximization (EM) algorithm. Combining our approach with these could potentially improve estimation; we demonstrate this for the EM algorithm explicitly.

Our main contribution are new procedures for estimating ATSMs with spanned or unspanned stochastic volatility. For models with spanned factors where volatility factors price bonds, we propose to maximize a concentrated likelihood that when optimized gives exactly the same estimator as maximizing the original likelihood function. However, it only requires numerically optimizing over a subset of the parameters. The concentrated likelihood function is simple to construct from linear regressions. Using this approach, estimation of spanned models only takes a fraction of a second to several minutes compared to hours when optimizing the original likelihood.

For USV models where the volatility factors do not price bonds, the log-likelihood function is not known in closed-form adding another layer of difficulty. Nevertheless, we show how the intuition behind the concentrated likelihood for spanned models can be extended to estimate USV models using the EM algorithm of Dempster, Laird, and Rubin(1977). The maximization step of the EM algorithm solves a similar problem as optimizing the likelihood function of a spanned Gaussian ATSM. Consequently, we can construct a concentrated objective function for the EM algorithm using linear regressions just as we did for spanned models.

Our method outperforms conventional approaches both in terms of stability of convergence and speed. A study for a 3-factor model with one spanned volatility factor shows that our method guarantees convergence as long as it is locally identified, and it converges to a number of local maxima repeatedly. Aside from being able to find the global maximum, our method helps us to locate and understand the economic implications of different local maxima. Conversely, the conventional method of directly maximizing the original likelihood never converges fully to any of the local maxima, nor does it converge to the same point twice in repeated trials even when it is initialized under the same local mode. This makes it difficult for researchers to differentiate between points near a well-behaved local maximum having the same economic meaning and locations corresponding to local maxima that are economically different. The median time it takes for our new procedure is less than 2 minutes for this model, whereas the conventional approach takes over 2 hours.

Using our method, we shed light on how local maxima with different economic implications are created in non-Gaussian spanned models. In Gaussian models, different rotations of the factors (such as re-ordering of the factors) result in equivalent global maxima, with identical economic implications. In non-Gaussian models with spanned factors, rotations can have substantial economic impacts. The non-Gaussian state variables must be positive and enter the conditional variance. This creates an asymmetry between the Gaussian and non-Gaussian factors resulting in local maxima that are not economically equivalent.

Another contribution of this paper is to develop a family of discrete-time non-Gaussian ATSMs that encompasses continuous-time models, including both spanned models as in Duffie and Kan(1996), Duffee(2002), Cheridito, Filipovic, and Kimmel(2007), and Aït-Sahalia and Kimmel(2010) as well as USV models as proposed by Collin-Dufresne and Goldstein(2002). Gouriéroux, Monfort, and Polimenis(2002) proposed a one factor discrete-time non-Gaussian model and Le, Singleton, and Dai(2010) generalized it to have multiple factors. Our model encompasses any admissible rotation of a multivariate discrete-time Cox, Ingersoll, and Ross(1985) process, allowing the factors to be correlated. The model nests the risk-neutral dynamics of other discrete-time ATSMs.[1] In our model, the physical and risk neutral dynamics follow the same stochastic process but with different parameters. The market prices of risk have the extended affine form of Cheridito, Filipovic, and Kimmel(2007), which is different than Le, Singleton, and Dai(2010). Finally, we also provide the restrictions needed to generate USV in discrete-time versions of the continuous-time models studied by Collin-Dufresne, Goldstein, and Jones(2009) and Joslin(2010).

We apply our estimation method to a range of popular spanned and unspanned models with three and four factors. Judging by the estimated likelihood, a model with three spanned non-Gaussian factors has the highest likelihood followed by one of the USV models. Gaussian and non-Gaussian models with spanned factors fit the cross-section of yields equally well.

[1]In this paper, we do not consider the class of non-Gaussian ATSMs built from the non-central Wishart process of Gouriéroux, Jasiak, and Sufana(2009).

However, spanned models do not capture the volatility well at any maturity, even for the best fitting model. This is because the non-Gaussian state variables must simultaneously fit the conditional mean and variance. Maximum likelihood places more weight on the first moment. In order to guarantee unspanned volatility factors, USV models place restrictions on the bond loadings. This causes USV models to sacrifice some cross-sectional fit; their pricing errors are larger than spanned models. On the other hand, USV models fit the dynamics of yield curve volatility well. The USV restrictions are not unique and we show that the choice of which USV restrictions are imposed is not inconsequential.

This paper continues as follows. In Section 2, we specify a general class of discrete-time, non-Gaussian affine term structure models. In Section 3, we describe our new approach to estimation for both spanned and unspanned models. Section 4 describes the data and parameter restrictions of the models. Section 5 studies a three factor spanned model in depth. In Section 6, we study eight three and four factor spanned and unspanned models. In Section 7, we discuss directions for future research and conclude.

2 Model

In this section, we describe a class of discrete-time ATSMs with stochastic volatility that encompass both spanned models, as in Duffie and Kan(1996), Dai and Singleton(2000), Cheridito, Filipovic, and Kimmel(2007); and unspanned models, as proposed by Collin-Dufresne and Goldstein(2002).

2.1 Bond prices

The model has a $G \times 1$ vector of conditionally Gaussian state variables g_t, whose volatilities are captured by an $H \times 1$ vector of positive state variables h_t. Under the risk-neutral measure $\mathbb{Q}$, the Gaussian state variables follow a vector autoregression with conditional heteroskedasticity

$$
\begin{aligned}
g_{t+1} &= \mu_g^{\mathbb{Q}} + \Phi_g^{\mathbb{Q}} g_t + \Phi_{gh}^{\mathbb{Q}} h_t + \Sigma_{gh}\varepsilon_{h,t+1}^{\mathbb{Q}} + \varepsilon_{g,t+1}^{\mathbb{Q}}, \qquad \varepsilon_{g,t+1}^{\mathbb{Q}} \overset{\mathbb{Q}}{\sim} \mathrm{N}\left(0, \Sigma_{g,t}\Sigma_{g,t}'\right), \qquad (1) \\
\Sigma_{g,t}\Sigma_{g,t}' &= \Sigma_{0,g}\Sigma_{0,g}' + \sum_{i=1}^{H} \Sigma_{i,g}\Sigma_{i,g}' h_{it}, \\
\varepsilon_{h,t+1}^{\mathbb{Q}} &= h_{t+1} - \mathbb{E}^{\mathbb{Q}}\left(h_{t+1}|\mathcal{I}_t\right)
\end{aligned}
$$

where $\mathcal{I}_t$ captures agents' information set at time t.

The volatility factors h_t are an affine transformation of the exact discrete-time equivalent of a multivariate Cox, Ingersoll, and Ross(1985) process

$$
\begin{aligned}
h_{t+1} &= \mu_h + \Sigma_h w_{t+1} & (2) \\
w_{i,t+1} &\sim \text{Gamma}\left(\nu_{h,i}^{\mathbb{Q}} + z_{i,t+1}^{\mathbb{Q}}, 1\right), \qquad i = 1, \ldots, H & (3) \\
z_{i,t+1}^{\mathbb{Q}} &\sim \text{Poisson}\left(\mathrm{e}_i' \Sigma_h^{-1} \Phi_h^{\mathbb{Q}} \Sigma_h w_t\right), \qquad i = 1, \ldots, H & (4)
\end{aligned}
$$

where e_i denotes the i-th column of the identity matrix I_H. We discuss the admissibility restrictions and interpretation of the parameters of the model in Section 2.2.

The price of a zero-coupon bond with maturity n at time t is the expected price of the same asset at time $t+1$ discounted by the short rate r_t under the risk neutral measure

$$
P_t^n = \mathbb{E}_t^{\mathbb{Q}}\left[\exp\left(-r_t\right) P_{t+1}^{n-1}\right].
$$

The short rate is a linear function of the state vector

$$
r_t = \delta_0 + \delta_{1,h}' h_t + \delta_{1,g}' g_t.
$$

Given the dynamics of g_t and h_t under $\mathbb{Q}$, bond prices are an exponentially affine function of the state variables

$$P_t^n = \exp\left(\bar{a}_n + \bar{b}_{n,h}' h_t + \bar{b}_{n,g}' g_t\right).$$

The loadings $\bar{a}_n, \bar{b}_{n,h}$ and $\bar{b}_{n,g}$ can be expressed recursively in matrix notation as

$$\begin{aligned}
\bar{a}_n &= -\delta_0 + \bar{a}_{n-1} + \mu_g^{\mathbb{Q}\prime}\bar{b}_{n-1,g} + \left[\mu_h - \Phi_h^{\mathbb{Q}}\mu_h + \Sigma_h\nu_h^{\mathbb{Q}}\right]'\bar{b}_{n-1,h} + \frac{1}{2}\bar{b}_{n-1,g}'\Sigma_{0,g}\Sigma_{0,g}'\bar{b}_{n-1,g} \\
&\quad + \mu_h'\Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\left(I_H - \left[\operatorname{diag}\left(\iota_H - \Sigma_h'\bar{b}_{n-1,gh}\right)\right]^{-1}\right)\Sigma_h'\bar{b}_{n-1,gh} \\
&\quad - \nu_h^{\mathbb{Q}\prime}\left[\log\left(\iota_H - \Sigma_h'\bar{b}_{n-1,gh}\right) + \Sigma_h'\bar{b}_{n-1,gh}\right] && (5) \\
\bar{b}_{n,h} &= -\delta_{1,h} + \Phi_{gh}^{\mathbb{Q}\prime}\bar{b}_{n-1,g} + \Phi_h^{\mathbb{Q}\prime}\bar{b}_{n-1,h} + \frac{1}{2}\left(I_H \otimes \bar{b}_{n-1,g}'\right)\Sigma_g\Sigma_g'\left(\iota_H \otimes \bar{b}_{n-1,g}\right) \\
&\quad - \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\left(I_H - \left[\operatorname{diag}\left(\iota_H - \Sigma_h'\bar{b}_{n-1,gh}\right)\right]^{-1}\right)\Sigma_h'\bar{b}_{n-1,gh} && (6) \\
\bar{b}_{n,g} &= -\delta_{1,g} + \Phi_g^{\mathbb{Q}\prime}\bar{b}_{n-1,g} && (7)
\end{aligned}$$

with initial values $\bar{a}_1 = -\delta_0, \bar{b}_{1,g} = -\delta_{1,g}$ and $\bar{b}_{1,h} = -\delta_{1,h}$, see Appendix B for a derivation. The matrix $\Sigma_g\Sigma_g'$ is a $GH \times GH$ block diagonal matrix with elements $\Sigma_{i,g}\Sigma_{i,g}'$ for $i = 1, \ldots, H$ and $\bar{b}_{n-1,gh} = \Sigma_{gh}'\bar{b}_{n-1,g} + \bar{b}_{n-1,h}$. The loadings must satisfy the restriction that the i-th component of $\Sigma_h'\bar{b}_{n-1,gh} < 1$ for $i = 1, \ldots, H$.

Bond yields $y_t^n \equiv -\frac{1}{n}\log\left(P_t^n\right)$ are linear in the factors

$$y_t^n = a_n + b_{n,h}'h_t + b_{n,g}'g_t \tag{8}$$

with $a_n = -\frac{1}{n}\bar{a}_n$, $b_{n,h} = -\frac{1}{n}\bar{b}_{n,h}$ and $b_{n,g} = -\frac{1}{n}\bar{b}_{n,g}$.

Gouriéroux and Jasiak(2006) built the univariate version of the non-Gaussian process (2)-(4) and Gouriéroux, Monfort, and Polimenis(2002) used it to construct a one factor ATSM. Le, Singleton, and Dai(2010) extended the process to allow for multiple factors; their specification under $\mathbb{Q}$ is (2)-(4) but with $\mu_h = 0$ and Σ_h diagonal.

2.1.1 Unspanned stochastic volatility models

USV models (see Collin-Dufresne and Goldstein(2002)) impose restrictions on the parameters under $\mathbb{Q}$ guaranteeing that the bond loadings for the volatility factors $b_{n,h}$ in (8) are zero for all maturities. Yields consequently only depend on the Gaussian factors $y_t^n = a_n + b_{n,g}' g_t$. The bond loadings (5)-(7) simplify to

$$\begin{aligned} \bar{a}_n &= -\delta_0 + \bar{a}_{n-1} + \mu_g^{\mathbb{Q}\prime} \bar{b}_{n-1,g} + \frac{1}{2} \bar{b}_{n-1,g}' \Sigma_{0,g} \Sigma_{0,g}' \bar{b}_{n-1,g} && (9) \\ \bar{b}_{n,g} &= -\delta_{1,g} + \Phi_g^{\mathbb{Q}\prime} \bar{b}_{n-1,g} && (10) \end{aligned}$$

which are the same as Gaussian ATSMs. Unlike Gaussian ATSMs, however, USV models constrain some of the $\mathbb{Q}$ parameters (i.e. elements of $\Phi_g^{\mathbb{Q}}$) in order to set the non-Gaussian loadings to zero. In Section 4.2.2, we provide conditions under which discrete-time ATSMs exhibit USV as in Collin-Dufresne, Goldstein, and Jones(2009).

2.2 Physical dynamics

Analogous to the popular class of Gaussian ATSMs, we specify the dynamics of g_t and h_t under $\mathbb{P}$ to have the same dynamics as under $\mathbb{Q}$. The Gaussian state variables follow a vector autoregression with conditional heteroskedasticity

$$\begin{aligned} g_{t+1} &= \mu_g + \Phi_g g_t + \Phi_{gh} h_t + \Sigma_{gh} \varepsilon_{h,t+1} + \varepsilon_{g,t+1}, \quad \varepsilon_{g,t+1} \sim \mathrm{N}\left(0, \Sigma_{g,t} \Sigma_{g,t}'\right), && (11) \\ \Sigma_{g,t} \Sigma_{g,t}' &= \Sigma_{0,g} \Sigma_{0,g}' + \sum_{i=1}^{H} \Sigma_{i,g} \Sigma_{i,g}' h_{it}, \\ \varepsilon_{h,t+1} &= h_{t+1} - \mathbb{E}\left(h_{t+1} | \mathcal{I}_t\right). \end{aligned}$$

The Gaussian state variables are a function of the non-Gaussian state variables through both the autoregressive term $\Phi_{gh} h_t$ and the covariance term $\Sigma_{gh} \varepsilon_{h,t+1}$. The parameters controlling the conditional mean are different under $\mathbb{P}$ and $\mathbb{Q}$ measures, while the scale parameters Σ_{gh} and $\Sigma_{i,g}$ for $i = 0, \ldots, H$ are the same.

The model for h_{t+1} under the physical measure $\mathbb{P}$ is

$$h_{t+1} = \mu_h + \Sigma_h w_{t+1} \tag{12}$$

$$w_{i,t+1} \sim \text{Gamma}\left(\nu_{h,i} + z_{i,t+1}, 1\right), \qquad i = 1, \ldots, H \tag{13}$$

$$z_{i,t+1} \sim \text{Poisson}\left(\mathrm{e}_i' \Sigma_h^{-1} \Phi_h \Sigma_h w_t\right), \qquad i = 1, \ldots, H \tag{14}$$

where $\nu_h = (\nu_{h,1}, \ldots, \nu_{h,H})$ are shape parameters, Φ_h is a matrix controlling the autocorrelation of h_{t+1}, Σ_h is a scale matrix, and μ_h is a vector determining the lower bound of h_{t+1}. Sufficient conditions for non-negativity of h_t are that elements of μ_h, Σ_h, and $\Sigma_h^{-1}\Phi_h\Sigma_h$ are non-negative. The discrete-time equivalent of the Feller condition $\nu_{h,i} > 1$ ensures that the process does not attain its lower bound. A similar set of restrictions must be satisfied under $\mathbb{Q}$. The scale parameters Σ_h are the same under both probability measures and so is the parameter μ_h. The latter restriction is required for no-arbitrage.

The conditional mean of the volatility factors h_{t+1} can be written in matrix form as

$$\mathbb{E}\left(h_{t+1}|\mathcal{I}_t\right) = \left(I_H - \Phi_h\right)\mu_h + \Sigma_h \nu_h + \Phi_h h_t.$$

It is a linear function of its own lag h_t, similar to a vector autoregression. The conditional variance is also an affine function of h_t

$$\mathbb{V}\left(h_{t+1}|\mathcal{I}_t\right) = \Sigma_h \text{diag}(\nu_h - 2\Sigma_h^{-1}\Phi_h\mu_h)\Sigma_h' + \Sigma_h \text{diag}\left(2\Sigma_h^{-1}\Phi_h h_t\right)\Sigma_h'.$$

In Appendix A.2, we provide the transition density of h_{t+1} for any admissible rotation.

A nice property of the model (11)-(14) for the vector $(h_t', g_t')'$ is that any admissible affine transformation remains within the same family of distributions.

Proposition 1 *Let g_t and h_t follow the process of (11)-(14) with parameters θ. Consider* *an admissible affine transformation of the form*

$$\begin{pmatrix} \tilde{h}_t \\ \tilde{g}_t \end{pmatrix} = \begin{pmatrix} c_h \\ c_g \end{pmatrix} + \begin{pmatrix} C_{hh} & C_{hg} \\ C_{gh} & C_{gg} \end{pmatrix} \begin{pmatrix} h_t \\ g_t \end{pmatrix}.$$

The new process $\tilde{g}_t$ and $\tilde{h}_t$ remains in the same family of distributions under updated parameters $\tilde{\theta}$. The parameters ν_h and $\Sigma_h^{-1}\Phi_h\Sigma_h$ are invariant to rotation.

Proof: See Appendix C.1.

The admissibility restrictions and the relationship between the new and old parameterizations can be found in Appendix C.1. This proposition helps to understand identification in Section 4.2.1. The admissibility constraints ensure that the non-Gaussian state variables always remain positive after applying a transformation from $(h_t', g_t')'$ to $\left(\tilde{h}_t', \tilde{h}_t'\right)'$ and that there exists another admissible rotation to get back to the original factors.

2.3 Stochastic discount factor

In this section, we demonstrate how an agent gets compensated for risk exposure when holding a zero-coupon bond under stochastic volatility. Given the dynamics of the state vector under $\mathbb{P}$ and $\mathbb{Q}$ measures, the market prices of risk have the extended affine form of Cheridito, Filipovic, and Kimmel(2007). The full expression is in Appendix D. To provide intuition, the log of the stochastic discount factor (SDF) can be decomposed up to a first order approximation into the risk free rate plus three components describing risk compensation

$$m_{t+1} = -r_t - \frac{1}{2}\lambda_{gt}'\lambda_{gt} - \lambda_{gt}'\epsilon_{g,t+1} - \lambda_{wt}'\epsilon_{w,t+1} - \lambda_{zt}'\epsilon_{z,t+1} \qquad (15)$$

where $\epsilon_{i,t+1}$ are standardized shocks with mean zero and identity covariance matrix, and λ_{it} is the price of risk i for each of the three types of shocks in the model.[2] In addition to the risk-free r_t, the agent gets compensated for being exposed to the Gaussian shock $\epsilon_{g,t+1}$ in

[2] This approximation is not used during estimation.

(11), the gamma shock $\epsilon_{w,t+1}$ in (13), and the Poisson shock $\epsilon_{z,t+1}$ in (14). The prices of these risks are defined as

$$\begin{aligned}
\lambda_{gt} &= \mathbb{V}(g_{t+1}|\mathcal{I}_t, h_{t+1}, z_{t+1})^{-1/2}\left[\mathbb{E}\left(g_{t+1}|\mathcal{I}_t, h_{t+1}, z_{t+1}\right) - \mathbb{E}^{\mathbb{Q}}\left(g_{t+1}|\mathcal{I}_t, h_{t+1}, z_{t+1}\right)\right] \\
\lambda_{wt} &= \mathbb{V}(w_{t+1}|\mathcal{I}_t, z_{t+1})^{-1/2}\left[\mathbb{E}(w_{t+1}|\mathcal{I}_t, z_{t+1}) - \mathbb{E}^{\mathbb{Q}}(w_{t+1}|\mathcal{I}_t, z_{t+1})\right], \\
\lambda_{zt} &= \mathbb{V}(z_{t+1}|\mathcal{I}_t)^{-1/2}\left[\mathbb{E}(z_{t+1}|\mathcal{I}_t) - \mathbb{E}^{\mathbb{Q}}(z_t|\mathcal{I}_t)\right].
\end{aligned}$$

The market prices of risk have an intuitive form as the Sharpe ratio measuring per unit risk compensation. Specifically, they are the difference in the conditional means of each shock under $\mathbb{P}$ and $\mathbb{Q}$ standardized by a conditional standard deviation. The time-varying quantities of risk are a feature of non-Gaussian models that are not available in Gaussian models.

USV models In USV models, the components of the SDF associated with the Gaussian factors (the first three terms in (15)) are the only parts that are directly observable from bond yields. The risk premium for non-Gaussian factors can only be estimated jointly by also observing derivatives because the $\mathbb{Q}$ parameters of the volatility process do not enter the likelihood and are unidentified by observing only yields.

2.4 State space representation

Define x_t as the vector of spanned factors; i.e., $x_t = (g_t', h_t')'$ in spanned models and $x_t = g_t$ in USV models. Stacking y_t^n in order for N different maturities $n_1, n_2, ..., n_N$ gives $Y_t = A + Bx_t$ where $A = (a_{n_1}, \ldots, a_{n_N})'$, $B = (b_{n_1}', ..., b_{n_N}')'$. If more yields are observed than the number of spanned factors ($N > N_1$), not all yields can be priced exactly. Following the literature, we assume that N_1 linear combinations of the yields $Y_t^{(1)} = S_{Y_1}Y_t$ are priced without error and the remaining $N_2 = N - N_1$ linear combinations $Y_t^{(2)} = S_{Y_2}Y_t$ are observed with Gaussian measurement errors. Given this assumption, the observation equations are

$$Y_t^{(1)} = A_1 + B_1 x_t \tag{16}$$

$$Y_t^{(2)} = A_2 + B_2 x_t + \eta_t \qquad \eta_t \sim \mathrm{N}(0, \Omega) \tag{17}$$

where $A_1 \equiv S_{Y_1} A, A_2 \equiv S_{Y_2} A, B_1 \equiv S_{Y_1} B$, and $B_2 \equiv S_{Y_2} B$. The state space representation of the model is completed using the dynamics of the state variables (11)-(14).

3 Estimation methodology

In this section, we introduce new estimation procedures for spanned and unspanned models, both of which use least square regressions to simplify and stabilize estimation. Our approach is based on the following observations: (i) The parameter vector θ can be separated into those parameters that enter the bond loadings and those that do not (e.g. μ_g, Φ_g, Φ_{gh}); (ii) Given the parameters that enter the loadings, we can calculate the bond loadings and solve for the spanned factors $x_t = B_1^{-1}\left(Y_t^{(1)} - A_1\right)$ using (16); (iii) The $\mathbb{P}$ parameters (e.g. μ_g, Φ_g, Φ_{gh}) of the Gaussian VAR for the factors g_t plus Ω enter the objective function as a quadratic form. The first order conditions for these parameters $(\mu_g, \Phi_g, \Phi_{gh}, \Omega)$ are linear and can be solved by running (generalized) least squares regressions of (11). Using this basic insight, we show how to eliminate these parameters from the numerical optimization problem.

3.1 Spanned models

Given the parameters of the model θ, the likelihood function is

$$\begin{aligned} p\left(Y_{1:T};\theta\right) &= p\left(Y_{1:T}^{(2)}|Y_{1:T}^{(1)};\theta\right) p\left(Y_{1:T}^{(1)};\theta\right) \\ &= \prod_{t=1}^{T} p\left(Y_t^{(2)}|Y_t^{(1)};\theta\right) \left|J\left(\theta\right)\right|^{-T} \prod_{t=1}^{T} p\left(g_t|h_t,\mathcal{I}_{t-1};\theta\right) \prod_{t=1}^{T}\prod_{i=1}^{H} p\left(h_{it}|\mathcal{I}_{t-1};\theta\right) \end{aligned} \tag{18}$$

where $J(\theta)$ is the Jacobian of the transformation from $x_t = (g_t', h_t')'$ to $Y_t^{(1)}$, see Appendix E for the log-likelihood $\ell(\theta) = \log p(Y_{1:T}; \theta)$.[3] Direct maximization of the log-likelihood is however extremely challenging as interest rates are close to non-stationary, the bond loadings are non-linear functions of the models' parameters, and the maximization must impose the condition that $h_t > 0$.

Our approach to spanned models is a result of the following proposition:

Proposition 2 *If the model is given by equations (11)-(14) and (16)-(17) with all spanned factors, then the maximum likelihood estimator $\hat{\theta} = \underset{\theta}{argmax}\ \ell(\theta)$ can be solved by maximizing the concentrated likelihood $\max_{\theta_m} \ell\left(\hat{\theta}_c(\theta_m), \theta_m\right)$, where $\theta_c = (\mu_g, \Phi_g, \Phi_{gh}, \Omega)$ and θ_m are the remaining parameters of the model. The function $\hat{\theta}_c(\theta_m)$ is obtained by solving $\max_{\theta_c} \ell(\theta_m, \theta_c)$ using the GLS estimates for the $\mathbb{P}$ dynamics (11) and the OLS estimates for the variance-covariance matrix Ω in (17).*

Proof: See Appendix F.1.

The proposition raises two points. First, optimizing the concentrated likelihood gives exactly the same solution as maximizing the original likelihood function in (18). However, it only requires numerically optimizing over θ_m instead of both θ_m and θ_c. Second, the concentrated likelihood function can be constructed from linear regressions. The method we propose is an immediate result of Proposition 2.

Procedure 1 *Maximize the concentrated log-likelihood function $\max_{\theta_m} \ell\left(\hat{\theta}_c(\theta_m), \theta_m\right)$. For a given value of θ_m, the concentrated likelihood can be constructed as follows:*

(i.) Given θ_m, calculate the bond loadings A and B and the state variables g_t and h_t from $x_t = B_1^{-1}\left(Y_t^{(1)} - A_1\right)$.

[3] The stationary distribution is only known for special sub-classes of the affine family of models. In this paper, we assume a diffuse initial condition and start from $t = 2$. This provides an analytical solution for the first order conditions of the likelihood. If a researcher wants to include the stationary distribution as the initial condition, we recommend using our procedure first, and then using these estimates as starting values to optimize the likelihood with the initial condition. While including the initial conditions enforces stationarity, it can also introduce a downward bias in estimates of the autoregressive parameters; see, e.g. Bauer, Rudebusch, and Wu(2012).

(ii.) Given g_t and h_t, calculate $\varepsilon_{h,t+1}$ and $\Sigma_{g,t}$. Run a GLS regression

$$g_{t+1} - \Sigma_{gh}\varepsilon_{h,t+1} = \mu_g + \Phi_g g_t + \Phi_{gh} h_t + \Sigma_{g,t}\varepsilon_{g,t+1} \tag{19}$$

to calculate $\hat{\mu}_g(\theta_m), \hat{\Phi}_g(\theta_m), \hat{\Phi}_{gh}(\theta_m)$.

(iii.) Calculate the covariance matrix

$$\hat{\Omega}(\theta_m) = \frac{1}{T-1}\sum_{t=2}^{T}\left(Y_t^{(2)} - A_2 - B_2 x_t\right)\left(Y_t^{(2)} - A_2 - B_2 x_t\right)' \tag{20}$$

(iv.) Substitute $\hat{\theta}_c(\theta_m) = \left(\hat{\mu}_g(\theta_m), \hat{\Phi}_g(\theta_m), \hat{\Phi}_{gh}(\theta_m), \hat{\Omega}(\theta_m)\right)$ back into the original likelihood to form the concentrated likelihood.

In Appendix F, we also derive the analytical gradients of the concentrated log-likelihood. Our derivation is based on the following proposition. It decomposes the gradient into pieces according to whether a parameter enters the bond loadings, the $\mathbb{P}$ dynamics, or both.

Proposition 3 *The gradient of the concentrated log-likelihood $\ell\left(\hat{\theta}_c(\theta_m), \theta_m\right)$ can be decomposed into three terms:*

$$\begin{aligned} \frac{d\ell\left(\hat{\theta}_c(\theta_m), \theta_m, A(\theta_m), B(\theta_m)\right)}{d\theta_m'} &= \frac{\partial\ell\left(\hat{\theta}_c, \theta_m, A, B\right)}{\partial\theta_m'} + \frac{\partial\ell\left(\hat{\theta}_c, \theta_m, A, B\right)}{\partial A'}\frac{\partial A(\theta_m)}{\partial\theta_m'} \\ &+ \frac{\partial\ell\left(\hat{\theta}_c, \theta_m, A, B\right)}{\partial vec(B')'}\frac{\partial vec\left(B(\theta_m)'\right)}{\partial\theta_m'}. \end{aligned}$$

The first term is the partial derivative of the $\mathbb{P}$ dynamics and Jacobian with respect to θ_m. This measures the effect parameters have on the log-likelihood through the time series of the factors. The second and third terms measure the effect parameters have on the log-likelihood through the bond loadings A and B.

Proof: See Appendix F.2

The expressions for the gradient can be used for other affine models such as models for defaultable bonds and credit default swaps.

3.2 Unspanned models

In a USV model, the pricing equation (16) can be inverted to calculate the Gaussian factors g_t conditional on the parameters that enter the bond loadings. However, the volatility factors h_t cannot be observed by inverting the pricing formula. The likelihood of the model $p\left(Y_{1:T};\theta\right) = p\left(Y_{1:T}^{(2)}|Y_{1:T}^{(1)};\theta\right)p\left(Y_{1:T}^{(1)};\theta\right)$ is no longer known in closed-form. The first term $p\left(Y_{1:T}^{(2)}|Y_{1:T}^{(1)};\theta\right)$ remains the same as in (18). The second term $p\left(Y_{1:T}^{(1)};\theta\right)$ is associated with the $\mathbb{P}$ dynamics of the factors and is an integral over the path of the latent volatility

$$p\left(Y_{1:T}^{(1)};\theta\right) = \left|J\left(\theta\right)\right|^{-T}\int\ldots\int p\left(g_{1:T}|h_{0:T-1};\theta\right)p\left(h_{0:T-1};\theta\right)dh_0\ldots dh_{T-1}$$

where $J\left(\theta\right)$ is the Jacobian from g_t to $Y_t^{(1)}$. This integral does not have a closed-form solution. We use the Monte Carlo Expectation Maximization (MCEM) algorithm to estimate the model, see Wei and Tanner(1990).

The EM algorithm consists of two steps: the expectation and maximization steps, which are iterated back and forth until convergence of the algorithm to a stationary point of the likelihood. The first step calculates the expected value of the complete data log-likelihood

$$\begin{aligned} Q\left(\theta|\theta^{(i)}\right) = \mathbb{E}\Bigg[&\sum_{t=1}^{T}\log p\left(Y_t^{(2)}|g_t;\theta\right) - T\log\left|J\left(\theta\right)\right| + \sum_{t=1}^{T}\log p\left(g_t|g_{t-1},h_{t-1};\theta\right) \\ &+\sum_{t=1}^{T-1}\log p\left(h_t|h_{t-1};\theta\right) + p\left(h_0;\theta\right)\Bigg]. \end{aligned} \tag{21}$$

This expectation is taken with respect to the posterior distribution $p\left(h_{0:T-1}|Y_{1:T};\theta^{(i)}\right)$, which depends on the parameters $\theta^{(i)}$ from the previous iteration. The function $Q\left(\theta|\theta^{(i)}\right)$ is known as the intermediate quantity of the EM algorithm and it is a function of θ. In the second step of the EM algorithm, the intermediate quantity is maximized $\theta^{(i+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta|\theta^{(i)}\right)$

to determine the parameters for the next iteration.

For USV models, the intermediate quantity has the same form as the log-likelihood for spanned Gaussian models. This means that maximization of the intermediate quantity at each iteration of the EM algorithm is (essentially) equivalent to estimating a Gaussian ATSM. This is why we can construct a concentrated version of the intermediate quantity from the output of linear regressions. For USV models, we separate the parameters into three groups: $\theta_c = (\mu_g, \Phi_g, \Omega)$ are the parameters that can be concentrated out, $\theta_{m,h} = (\nu_h, \Phi_h, \Sigma_h)$ are the parameters that govern the dynamics of the volatility, and $\theta_{m,b}$ are the parameters that enter the bond loadings. We only need to optimize numerically over $\theta_{m,h}$ and $\theta_{m,b}$, as the parameters in θ_c can be determined analytically as a function of $\theta_{m,b}$. Moreover, the intermediate quantity can be additively separated into two pieces $Q\left(\theta|\theta^{(i)}\right) = Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right) + Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$. The first component corresponds to the first three terms in (21), and depends only on the parameters $\theta_{m,b}$ and θ_c. The remaining terms in (21) are the second component $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$, which depend only on the volatility parameters $\theta_{m,h}$.

Our procedure can be implemented as follows:

Procedure 2 *The maximum likelihood estimator for USV models can be obtained by iterating over the following two steps:*

(a.) E-step: compute the expectations in the intermediate quantity $Q\left(\theta|\theta^{(i)}\right)$ *from (21).*

(b.) M-step: maximize $Q\left(\theta|\theta^{(i)}\right)$ *over* θ *to determine* $\theta^{(i+1)}$*. This can be separated into two sub-steps.*

(b1.) Maximize $Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right)$ *with respect to* $\theta_{m,b}$ *and* θ_c *to determine* $\theta_{m,b}^{(i+1)}$ *and* $\theta_c^{(i+1)}$*. This can be solved equivalently by maximizing* $Q_1\left(\theta_{m,b}, \hat{\theta}_c(\theta_{m,b})|\theta^{(i)}\right)$ *with respect to* $\theta_{m,b}$*, where the concentrated objective function can be constructed as follows:*

(i.) Given $\theta_{m,b}$*, calculate the bond loadings* A *and* B *and the state variables* g_t *from* $x_t = B_1^{-1}\left(Y_t^{(1)} - A_1\right)$*.*

(ii.) *Given* g_t*, run a GLS regression*

$$g_{t+1} \;=\; \mu_g + \Phi_g g_t + \bar{S}_t^{\frac{1}{2}} \varepsilon_{g,t+1} \tag{22}$$

to calculate $\hat{\mu}_g(\theta_{m,b}), \hat{\Phi}_g(\theta_{m,b})$*.*

(iii.) *Calculate the covariance matrix* $\hat{\Omega}(\theta_{m,b})$ *as in (20).*

(iv.) *Substitute* $\hat{\theta}_c(\theta_{m,b}) = \left(\hat{\mu}_g(\theta_{m,b}), \hat{\Phi}_g(\theta_{m,b}), \hat{\Omega}(\theta_{m,b})\right)$ *back into the intermediate quantity.*

(b2.) *Maximize* $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$ *with respect to* $\theta_{m,h}$ *to determine* $\theta_{m,h}^{(i+1)}$*.*

Appendix H contains details of our implementation. This procedure for USV models has many similarities with our earlier procedure for spanned models.[4] The difference between the GLS regression in (19) of Procedure 1 and the regression in (22) of Procedure 2 is their covariance matrices. The EM algorithm imputes the latent values of h_t by taking their expectations. Finally, we note that with only a few minor modifications, the analytical gradients for the likelihood of the spanned model from Proposition 3 can be used to calculate the gradients of the intermediate quantity in (21).

In our experience, it takes only a few iterations of the EM algorithm to approach the maximum when estimating the USV models of Section 4.2.2. The rate of convergence of the EM algorithm near the maximum is known to be slow. Once the EM algorithm approaches the maximum, a researcher can switch to alternative estimation procedures for non-Gaussian state space models.[5]

For USV models, the expectation in Step 1 cannot be calculated in closed-form, requiring a Monte Carlo version of the EM algorithm. We calculate the expectations using sequential

[4] There are several versions of the EM algorithm all of which lead to the MLE; see Meng and Rubin(1993). These authors discuss issues such as concentrating the intermediate quantity as well as sequentially maximizing the intermediate quantity over subsets of θ.

[5] USV models are an example of a non-linear, non-Gaussian state space model. General approaches for estimating non-Gaussian state space models include importance sampling (Durbin and Koopman(2012) and Richard and Zhang(2007)), particle filters (see Malik and Pitt(2011)), and MCMC (see Jacquier, Johannes, and Polson(2007)).

Monte Carlo methods or particle filters; see Creal(2012) for a survey. In particular, we use the particle filtering algorithm of Godsill, Doucet, and West(2004) to draw paths of $h_{0:T-1}$ from the joint posterior distribution $p\left(h_{0:T-1}|Y_{1:T};\theta\right)$; see Appendix H.2 for details. The particle filter also allows us to calculate filtered (one-sided) estimates of the volatility as well as an estimate of the likelihood function $p\left(Y_{1:T};\theta\right)$.

3.3 Discussion

We discuss how our approach can be applied to a wide range of ATSMs.

Example #1: observable macroeconomic variables

Our approach can estimate models with observable macroeconomic variables and with homoskedastic or heteroskedastic shocks; examples with homoskedasticity are Ang and Piazzesi(2003) and Hamilton and Wu(2012). Our procedure works the same as before except the state vector x_t now contains the yield factors as well as the observed macroeconomic factors. For step $(i.)$ of Procedure 1 or 2, we back out the latent component of x_t conditional on a subset of the parameters, the yields $Y_t^{(1)}$, and the macro variables. Given x_t, we can concentrate a large number of parameters out of the objective function including many of the parameters introduced by adding the macroeconomic variables.

Example #2: Hidden factors

Recently, Duffee(2011) argued that more than three factors are needed to explain the time-series dynamics of yields and risk premia, where these additional factors are "hidden" from the cross-section of yields because the factors are not priced. For simplicity, we illustrate the basic ideas here for Gaussian models. Extensions to spanned or unspanned non-Gaussian models are straightforward.

The Gaussian state vector can be separated into sub-vectors $g_t = \left(g'_{1,t}, g'_{2,t}\right)'$ whose dimensions are $G_1 \times 1$ and $G_2 \times 1$, respectively. The dynamics under the $\mathbb{P}$ measure are

$$g_{1,t+1} = \mu_{g,1} + \Phi_{g,11} g_{1t} + \Phi_{g,12} g_{2t} + \varepsilon_{1,t+1} \qquad \varepsilon_{1,t+1} \sim \mathrm{N}\left(0, \Sigma_{0,g}\Sigma_{0,g}'\right) \tag{23}$$

$$g_{2,t+1} = \mu_{g,2} + \Phi_{g,21} g_{1t} + \Phi_{g,22} g_{2t} + \varepsilon_{2,t+1} \qquad \varepsilon_{2,t+1} \sim \mathrm{N}\left(0, I_{G_2}\right) \tag{24}$$

The dynamics of $g_{1,t}$ are the same under the $\mathbb{Q}$ measure but with the restrictions that $\Phi_{g,12}^{\mathbb{Q}} = 0$ and the last G_2 entries of δ_{1g} are zero. These restrictions imply that only $g_{1,t}$ directly impacts yields as the bond loadings on $g_{2,t}$ are zero by construction.

Given the parameters that enter the bond loadings, the factors that price bonds are conditionally observable $g_{1,t} = B_1^{-1}\left(Y_t^{(1)} - A_1\right)$ just as in step $(i.)$ of Procedure 1. We can treat $g_{1,t}$ as the observed data and (23) is the new observation equation for a linear, Gaussian state space model. The remaining state variables g_{2t} have transition equation (24) and are serially correlated shocks to the factors g_{1t} that price bonds. We can use the Kalman filter to estimate this model, which is equivalent to a GLS regression where the errors are serially correlated. To concentrate the parameters $(\mu_{g,1}, \mu_{g,2}, \Phi_{g,11}, \Phi_{g,21})$ out of the likelihood as in step (ii) of Procedure 1, we can either place these parameters in the state vector or use the augmented Kalman filter of de Jong(1991), see also Chapter 5 of Durbin and Koopman(2012).

Example #3: parameter constraints

In our approach, a researcher can impose constraints – such as in Ang and Piazzesi(2003), Kim and Wright(2005), Cochrane and Piazzesi(2008) and Bauer(2014) – and still concentrate out parameters by linear regression. We denote the penalized or constrained log-likelihood function $\ell_p(\theta)$ as

$$\ell_p(\theta) = \log p\left(Y_{1:T}; \theta\right) + p(\theta),$$

where $p(\theta)$ is the penalty term. If the constraints are only on the $\mathbb{Q}$ parameters, a researcher can directly apply our Procedures 1 and 2. If the goal is to constrain either the $\mathbb{P}$ parameters or the relationship between the $\mathbb{P}$ and $\mathbb{Q}$ parameters, the penalty term is a vector of Lagrange multipliers times the constraints. Step (*ii.*) of Procedure 1 or 2 can be replaced by constrained GLS. If the constraints are linear in μ_g, Φ_g, Φ_{gh}, there is a unique solution. Popular restrictions in the literature are all in this category. If the penalty term $p(\theta)$ is a quadratic function of μ_g, Φ_g, Φ_{gh}, step (*ii.*) of Procedures 1 or 2 reduces to ridge regression and the parameters can be shrunk to a pre-specified value similar to a Bayesian VAR. A researcher may want to shrink the $\mathbb{P}$ parameters toward the $\mathbb{Q}$ parameters, which are often measured more precisely.

Example #4: serially correlated measurement errors

Although often assumed to be i.i.d. normal in the literature, several authors have found the measurement errors η_t in (17) to be serially correlated; see, e.g. Hamilton and Wu(2014). Procedures 1 and 2 can be extended to cover models with serially correlated errors without increasing the dimension of the numerical optimization. For example, if η_t has VAR(1) dynamics $\eta_t = \Phi_\eta \eta_{t-1} + \kappa_t$ with $\kappa_t \sim \mathrm{N}(0, \Omega)$, then we can concentrate out $\theta_c = (\mu_g, \Phi_g, \Phi_{gh}, \Omega, \Phi_\eta)$ by least squares. Conditional on θ_m, we calculate the bond loadings A and B and factors x_t and, from these, the values of η_t and η_{t-1} which are implicitly a function of θ_m. To concentrate Φ_η and Ω out of the objective function, we compute their least squares estimates from the regression $\eta_t = \Phi_\eta \eta_{t-1} + \kappa_t$ and plug these values back into the log-likelihood.

Relation to Joslin, Singleton, and Zhu(2011) and Hamilton and Wu(2012)

Both Joslin, Singleton, and Zhu(2011) and Hamilton and Wu(2012) use linear regression to estimate some parameters of Gaussian models.[6] In the special case of Gaussian models with observable factors ($A_1 = 0$ and $B_1 = I$), our method is identical to the ML estimator of Joslin, Singleton, and Zhu(2011). Like the procedure in this paper, the approach of

[6]The work by Adrian, Crump, and Moench(2012) and Diez de Los Rios(2013) are also similar in spirit to these methods. They focus only on Gaussian models. Our discussion here applies to these papers as well.

Hamilton and Wu(2012) works for a wider range of Gaussian models. Their minimum chi-square estimator is asymptotically equivalent to the ML estimator in this paper.

The critical difference is that our procedures are designed for non-Gaussian models. Leveraging the analytical solution for linear regressions has not been explored in this area. For spanned models with both Gaussian and non-Gaussian factors, being able to rotate the factors is important. If a researcher takes an arbitrary basis of yields (such as principal components) and assumes that they can be separated a priori into observable Gaussian and non-Gaussian factors, it will restrict the likelihood. Our approach lets the data decide what linear combination of yields are the factors.

4 Data and parameter restrictions

4.1 Data

We use the Fama and Bliss(1987) zero coupon bond data available from the Center for Research in Securities Prices (CRSP). The data is monthly and spans from June 1952 through June 2012 for a total of $T = 721$ observations with maturities of $(1, 3, 12, 24, 36, 48, 60)$ months. For three factor models, the yields measured without error $Y_t^{(1)}$ include the (1, 12, 60) month maturities. In models with four spanned factors, $Y_t^{(1)}$ are the (1, 12, 24, 60) month maturities.

4.2 Parameter restrictions

4.2.1 Identifying restrictions for spanned models

We impose the following restrictions for identification. For the Gaussian part, these are: (i) $\mu_g^{\mathbb{Q}} = 0$; (ii) $\Phi_g^{\mathbb{Q}}$ in ordered Jordan form;[7] (iii) $\delta_{1g} = \iota$ is a column vector of ones; and (iv) $\Sigma_{i,g}$ is lower triangular. For the non-Gaussian part, (i) $\mu_h = 0$. (ii) $\Phi_{gh}^{\mathbb{Q}} = 0$. (iii) Elements

[7]For the case where $\Phi_g^{\mathbb{Q}}$ has real distinct eigenvalues, it is a diagonal matrix with diagonal elements in descending order.

of the vector $\delta_{1h} = \pm 1$ can take either sign[8], which unlike Gaussian-only models will lead to inequivalent maxima as we explain in Section 5.2; (iv) Σ_h is diagonal.

To guarantee non-negativity and admissibility of the factors, we also impose the discrete-time equivalent of the Feller condition $\nu_{h,i} > 1$ and $\nu^{\mathbb{Q}}_{h,i} > 1$ for $i = 1, \ldots, H$. The matrices Σ_h, $\Sigma_h^{-1}\Phi_h\Sigma_h$ and $\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}\Sigma_h$ must also be non-negative.

4.2.2 USV restrictions

We focus on discrete-time USV models similar to the continuous-time models presented in Collin-Dufresne, Goldstein, and Jones(2009) and Joslin(2010).[9] We label these models $\mathbb{U}_1(4)$ because they have one unspanned volatility factor and three Gaussian factors. USV restrictions are not unique. We present several models whose restrictions under $\mathbb{Q}$ result in non-Gaussian loadings where $b_{n,h} = 0$ for all maturities.

The first model, labeled $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$, has the following set of restrictions: (1) $\delta_{1,h} = 0$ and $\Sigma_{gh} = 0$. (2) $\Phi_g^{\mathbb{Q}}$ is a diagonal matrix with eigenvalues ϕ, ϕ^2, ψ. (3) All entries of $\Sigma_{1,g}\Sigma_{1,g}'$ are zero but the $(1,1)$ element. This entry is $\Sigma^2_{1,g,11}$. (4) $\Phi^{\mathbb{Q}}_{gh,3} = 0; \Phi^{\mathbb{Q}}_{gh,1} = \frac{\delta_{1,g,1}}{(1-\phi)}\Sigma^2_{1,g,11}; \Phi^{\mathbb{Q}}_{gh,2} = -\frac{(1-\phi^2)\delta^2_{1,g,1}}{2(1-\phi)^2\delta_{1,g,2}}\Sigma^2_{1,g,11}$. In this model, only the Gaussian factor associated with the eigenvalue ϕ has stochastic volatility as the remaining entries of $\Sigma_{1,g}\Sigma_{1,g}'$ are zero for all the other Gaussian factors. The USV restrictions force two of the eigenvalues of $\Phi_g^{\mathbb{Q}}$ to be related as ϕ and ϕ^2. These restrictions summarize three different USV models depending on the relative size of ϕ and ψ. We label the models $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$, $\mathbb{U}_1(4)(\phi > \psi > \phi^2)$, $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$. Each one of them is identified after imposing an ordering on the eigenvalues.[10]

[8]In theory, $\delta_{1h} = 0$ is also admissible and creates additional local maxima. We would like to thank an anonymous referee for pointing this out. Unlike a Gaussian model, the Jensen's inequality term prevents $b_{n,h}$ from being zero for maturities $n \geq 2$ even when $\delta_{1h} = 0$. In practice, this model is not well identified, as the factor loadings coming from the Jensen's inequality term are extremely small and close to 0. We found that the likelihood values under this restriction are significantly smaller in such models.

[9]Papers that document the potential existence of USV factors include Heidari and Wu(2003), Li and Zhao(2006), Trolle and Schwartz(2009), and Andersen and Benzoni(2010).

[10]We do not consider the case where the eigenvalues may be identical and we restrict our attention to eigenvalues that are less than one.

A second model, labeled $\mathbb{U}_1(4)(\phi, \phi^2, \phi^4)$, allows two out of the three Gaussian factors to share a common stochastic volatility factor. The restrictions for this model are (1) $\delta_{1,h} = 0$ and $\Sigma_{gh} = 0$. (2) $\Phi_g^{\mathbb{Q}}$ is a diagonal matrix with eigenvalues ϕ, ϕ^2, ϕ^4. (3) All entries of $\Sigma_{1,g}\Sigma_{1,g}'$ are zero but the $(1,1)$ and $(2,2)$ elements. These entries are $\Sigma_{1,g,11}^2$ and $\Sigma_{1,g,22}^2$. (4) $\Phi_{gh,1}^{\mathbb{Q}} = \frac{\delta_{1,g,1}}{1-\phi}\Sigma_{1,g,11}^2$; $\Phi_{gh,2}^{\mathbb{Q}} = \frac{\delta_{1,g,2}}{(1-\phi^2)}\Sigma_{1,g,22}^2 - \frac{(1-\phi^2)\delta_{1,g,1}^2}{2(1-\phi)^2\delta_{1,g,2}}\Sigma_{1,g,11}^2$; $\Phi_{gh,3}^{\mathbb{Q}} = -\frac{(1-\phi^4)\delta_{1,g,2}^2}{2(1-\phi^2)^2\delta_{1,g,3}}\Sigma_{1,g,22}^2$. In this model, two diagonal entries in the covariance matrix $\Sigma_{1,g}\Sigma_{1,g}'$ are non-zero. The additional flexibility in $\Sigma_{1,g}\Sigma_{1,g}'$ comes at the expense of another restriction on the diagonal components of $\Phi_g^{\mathbb{Q}}$. This model is unique because $\phi > \phi^2 > \phi^4$.

These restrictions lead to the following proposition

Proposition 4 *If the model has risk neutral dynamics (1)-(4) and satisfies the restrictions of $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$ or $\mathbb{U}_1(4)(\phi, \phi^2, \phi^4)$, then the model exhibits USV where $b_{n,h} = 0 \ \forall \ n$.*

Proof: see Appendix G.

Intuitively, USV models free up the volatility factors to fit the heteroskedasticity of yields because they still enter the covariance matrix of the Gaussian factors in their $\mathbb{P}$ dynamics. The cost of adding USV factors comes from the constraints they place on $\Phi_g^{\mathbb{Q}}$. These constraints can sacrifice the cross-sectional fit of the model.

We impose the identifying restrictions of Section 4.2.1 for the Gaussian factors. For the non-Gaussian portion of the model, we need an additional restriction on the scale parameters $\Sigma_{1,g}$ or Σ_h, as they enter the likelihood in the same way. We set $\Sigma_h = 0.01$.

5 A three factor model

In this section, we use a three factor model with one spanned volatility factor $\mathbb{A}_1(3)$ (in the Dai and Singleton(2000) notation) to demonstrate the performance of our method and to discuss the local maxima that arise in spanned non-Gaussian models. This model has been the preferred model by many researchers in the literature.[11] For an $\mathbb{A}_1(3)$ model, the

[11] This model has been widely considered as the benchmark non-Gaussian ATSM, see Dai and Singleton(2000), Cheridito, Filipovic, and Kimmel(2007), Collin-Dufresne, Goldstein, and Jones(2008), and Aït-

concentrated likelihood drops the number of parameters by one-third from 24 parameters to 16 parameters.

5.1 Performance comparison

To illustrate the mileage we gain from using our method, we compare our approach to the conventional method that does not concentrate out $(\mu_g, \Phi_g, \Phi_{gh})$ or use analytical gradients. We perform an experiment where we estimate the $\mathbb{A}_1(3)$ model on the CRSP dataset 100 times from 100 different starting values using both methods.[12] We compare our method and the direct approach along two dimensions: convergence and speed. To measure the former, we use the likelihood ratio (two times the difference in log-likelihoods).

The global solution found by our method has a log-likelihood of 36647.69 (estimates and quasi-maximum likelihood standard errors can be found on the right hand side of Table 2). We achieve an identical value for all 17 random starting values whenever the parameters were initialized in this region or mode of the parameter space.[13] Seventeen equals the number of times (one-sixth) that it started in this region. Conversely, the conventional method does not find this log-likelihood once nor does the method reproduce the same (incorrect) estimates for each of these 17 starting values. The highest log-likelihood value found by the standard approach is 36645.29. The differences between the two methods is significant across these 17 starting values. The conventional method yields log-likelihood values ranging between 36645.29 to 36636.82, and it only achieves the former for one starting value. With our method producing the same number repeatedly, we can conclude that it is a maximum.

An immediate benefit of the stable behavior of our method is that we are able to find that the $\mathbb{A}_1(3)$ model has 6 local modes with three well-behaved local maxima and three regions of the log-likelihood that appear to be locally unidentified. The three well-behaved local

Sahalia and Kimmel(2010) for examples.

[12] To make the comparison as parallel as we can, we write the likelihood function the same way, impose the same identifying restrictions, and use the same scaling and initial values for the parameters except that the conventional method has additional parameters entering the numerical optimizer.

[13] We consider two log-likelihoods to be numerically identical if they agree up to 2 decimal points. In practice, the log likelihood values are identical up to 8 decimal points.

maxima are listed in Table 1 and we will discuss the properties of the model that create these local modes in Section 5.2. Our method converges 17/100 times to Local 1, 14/100 times to Local 2, and 17/100 times to Local 3. The median likelihood ratio between our procedure and the un-concentrated log-likelihood with no analytical gradient is 29.5 indicating a substantial difference between the two procedures. The conventional method, even if it gets close to a local maximum, always stops before it fully converges. This makes it difficult for researchers to differentiate between points that are near a well-behaved local maximum that have the same economic meaning and locations corresponding to local maxima that are economically different.

Estimation time is another important dimension along which we compare our approach to the conventional method. The median estimation time for our procedure to estimate from a random starting value is less than 2 minutes, whereas the conventional approach of directly maximizing the log-likelihood function takes more than 2 hours. To perform our study with 100 starting values, it takes our method about 4 hours, whereas it takes roughly 9 days to complete the same exercise with the conventional method.

In summary, our method addresses all of the following problems with the conventional method. The conventional method is painfully slow. It does not achieve the global maximum. And, it is extremely hard to assess convergence behavior and the number of local maxima because conventional approaches do not repeatedly find the same local maximum even when started in that region of the parameter space.

5.2 Local maxima

Using our approach simplifies estimation and helps uncover some features of the log-likelihood surface that may be obscured by directly maximizing the log-likelihood. In this section, we discuss the characteristics of the model that create local maxima and their economic consequences.

In Gaussian models, a change in the sign of δ_{1g} rotates the factors from g_t to $-g_t$.

Table 1: Local maxima in the $\mathbb{A}_1(3)$ model

	Local 1	Local 2	Local 3
h_t	level	slope	curvature
$\Phi_h^{\mathbb{Q}}$	0.9961	0.9528	0.5812
$\Phi_g^{\mathbb{Q}}$	0.9514	1.0001	0.9983
	0.5358	0.5881	0.9389
$\delta_{1,h}$	1	1	-1
LLF	36647.69	36482.56	36530.19

Estimates of $\Phi_h^{\mathbb{Q}}$ and $\Phi_g^{\mathbb{Q}}$ from the $\mathbb{A}_1(3)$ model with corresponding log-likelihood. Each value is a different local maximum depending on the sign of $\delta_{1,h}$ and whether the non-Gaussian factor is the level, slope, or curvature.

This rotation is economically irrelevant because the estimated model switches between two global maximums. Unlike Gaussian models, fixing the sign of δ_{1h} in spanned models is not inconsequential, because the state variable h_t is positive by definition. Changing the sign of δ_{1h} does not rotate h_t to $-h_t$. Therefore, there can exist inequivalent local maxima for each combination of different signs of δ_{1h}. For each of the local maxima, the estimated latent factor h_t is different, which changes the conditional variance of g_t and consequently the log-likelihood.

Reordering the eigenvalues in $\Phi^{\mathbb{Q}}$ has completely different implications for spanned non-Gaussian models than for Gaussian models.[14] If the eigenvalues are reordered in a multi-factor Gaussian model, it implies equivalent global maxima with the same economic implication. However, with non-Gaussian spanned factors, they can yield inequivalent local maxima. Using the $\mathbb{A}_1(3)$ model as an example, the factors are labeled as level, slope and curvature. Reordering the eigenvalues *across* $\Phi_g^{\mathbb{Q}}$ and $\Phi_h^{\mathbb{Q}}$ does not generally change the shape of the factors but it does change whether h_t is the level, slope, or curvature. Any change in h_t from one type of factor (level) to another (slope/curvature) implies a different conditional variance for g_t making the likelihood no longer equivalent. Changing the order of the

[14] We collect the autoregressive parameters together in matrices as

$$\Phi = \begin{pmatrix} \Phi_h & 0 \\ \Phi_{gh} & \Phi_g \end{pmatrix} \qquad \Phi^{\mathbb{Q}} = \begin{pmatrix} \Phi_h^{\mathbb{Q}} & 0 \\ \Phi_{gh}^{\mathbb{Q}} & \Phi_g^{\mathbb{Q}} \end{pmatrix}$$

eigenvalues *within* $\Phi_g^{\mathbb{Q}}$ and/or $\Phi_h^{\mathbb{Q}}$ results in an equivalent global maximum. The intuition is the same as re-ordering of the factors g_t within a Gaussian ATSM.

In an ATSM with spanned non-Gaussian factors, it is not clear a priori which local maximum created by these characteristics of the model will be the global maximum. One must intentionally search each region that potentially has a local maximum and compare their likelihood values. To illustrate this idea, we present different local maxima for the $\mathbb{A}_1(3)$ model corresponding to different signs of $\delta_{1,h}$ and different orderings of the eigenvalues. We report $\Phi^{\mathbb{Q}}$, $\delta_{1,h}$, and log-likelihood values in Table 1. In the first column, h_t is the level factor and $\delta_{1,h}$ is positive. This is the global maximum in this case. In our sample, volatility is high during episodes where interest rates are high, so the level factor tends to explain the volatility best and $\delta_{1,h}$ is positive. The next two columns present what happens when h_t is the slope or curvature factor. Due to the nature of the data we are using, the likelihood function drops significantly from the global maximum to these alternative local maxima. In theory, there are six potentially different local maxima for each combination of eigenvalues and sign of $\delta_{1,h}$ but in practice there are only three well-behaved local maxima. For the rest, we observe parameters hitting a boundary or eigenvalues of the autoregressive parameters being numerically 1. Each of these local maxima correspond to models where volatility declines in the 1970's, which contradicts the data. Those points can be locally unidentified, meaning that there exists a region of the parameter space where a subset of the parameters are unidentified. For models with unspanned volatility factors, the issues of multiple modes do not appear, because the likelihood surface of USV models is similar to spanned Gaussian models.

6 Model comparison

In this section, we use our methodology to estimate a collection of popular and prominent ATSMs. The models that we estimate include three factor spanned models $\mathbb{A}_H(3)$ with

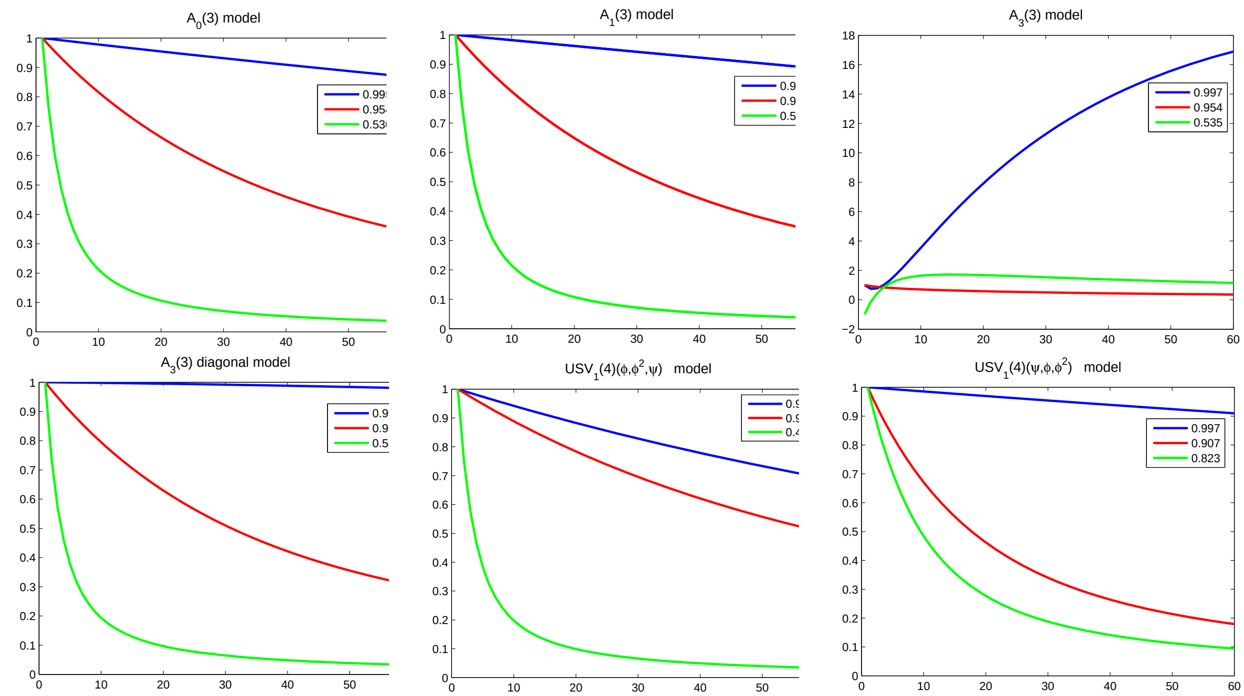
Table 2: Maximum likelihood estimates for the $\mathbb{A}_0(3)$ and $\mathbb{A}_1(3)$ models.

$G=3, H=0$	LLF =	37080.94		$G=2, H=1$	LLF =	36647.69		
μ_g				$\Sigma_h \nu_h$	μ_g			ν_h
6.97e-05	-4.85e-05	-3.37e-04		3.00e-05	-1.34e-05	3.32e-05		1.9332
(4.92e-05)	(1.12e-04)	(7.06e-05)		—	(5.14e-05)	(2.78e-05)		(2.0989)
Φ_g				Φ_h				
1.0074	0.0475	0.0665		0.9943				
(0.0084)	(0.0137)	(0.0316)		(0.0061)				
				Φ_{gh}	Φ_g			
-0.0115	0.9375	0.0192		0.0077	0.9854	0.0657		
(0.0184)	(0.0309)	(0.0607)		(0.0132)	(0.0224)	(0.0462)		
-0.0366	-0.0585	0.6306		-0.0405	-0.0729	0.6434		
(0.0111)	(0.0183)	(0.0535)		(0.0077)	(0.0152)	(0.0426)		
$\mu_g^{\mathbb{Q}}$			δ_0	$\Sigma_h \nu_h^{\mathbb{Q}}$	$\mu_g^{\mathbb{Q}}$		δ_0	$\nu_h^{\mathbb{Q}}$
0	0	0	0.0083	4.09e-05	0	0	-0.0011	2.6371
—	—	—	(0.0005)	—	—	—	(0.0002)	(0.2879)
$\Phi_g^{\mathbb{Q}}$				$\Phi_h^{\mathbb{Q}}$	$\Phi_g^{\mathbb{Q}}$			
0.9950	0.9538	0.5299		0.9961	0.9514	0.5358		
(0.0007)	(0.0031)	(0.0295)		(0.0006)	(0.0029)	(0.0293)		
$\Sigma_{0,g}$				Σ_h				
3.99e-04	0	0		1.55e-05				
(2.62e-05)	—	—		(1.68e-06)				
				Σ_{gh}	$\Sigma_{0,g}$		$\Sigma_{1,g}$	
-3.09e-04	5.09e-04	0		-0.8932	3.61e-11	0	0.0063	0
(4.54e-05)	(3.66e-05)	—		(0.0587)	(1.11e-11)	—	(0.0004)	—
-4.50e-06	-2.52e-04	3.78e-04		0.0538	-1.89e-11	-5.80e-12	-0.0035	0.0046
(2.23e-05)	(2.82e-05)	(2.37e-05)		(0.0397)	(8.52e-12)	(1.30e-12)	(0.0003)	(0.0003)
$\delta_{1,g}$				$\delta_{1,h}$	$\delta_{1,g}$			
1	1	1		1	1	1		
—	—	—		—	—	—		
$\sqrt{\text{diag}(\Omega)} \times 1200$				$\sqrt{\text{diag}(\Omega)} \times 1200$				
3 m	2 yr	3 yr	4 yr	3 m	2 yr	3 yr	4 yr	
0.2243	0.1251	0.1235	0.1077	0.2245	0.1248	0.1236	0.1078	

Maximum likelihood estimates with quasi-maximum likelihood standard errors. Left: Gaussian $\mathbb{A}_0(3)$ model. Right: non-Gaussian $\mathbb{A}_1(3)$ model. The identifying restrictions $\mu_g^{\mathbb{Q}} = 0, \delta_{1,g} = \iota$, and $\delta_{1,h} = 1$ are imposed during estimation.

$H = 0, 1, 2, 3$ volatility factors and four factor spanned models $\mathbb{A}_H(4)$ with $H = 0, 1$. We also estimate four USV models from Section 4.2.2: three versions of $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$ as well as the $\mathbb{U}_1(4)(\phi, \phi^2, \phi^4)$ model. We report the estimates as well as quasi-maximum likelihood standard errors as in White(1982) for eight models that have better empirical performance, leaving out the estimates for models with lower likelihoods for brevity.

Figure 1: Bond loadings for six different three factor models.



Bond loadings B as a function of maturity n. Top row from left to right: $\mathbb{A}_0(3)$ model, $\mathbb{A}_1(3)$ model, $\mathbb{A}_3(3)$ model. Second row: $\mathbb{A}_3(3)$ model with diagonal $\Phi_h^{\mathbb{Q}}$, $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model, $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ model. Reported in each graph are the eigenvalues of the risk-neutral feedback matrix $\Phi^{\mathbb{Q}}$. To make the restricted $\mathbb{A}_3(3)$ model easily comparable, we report the absolute value of the loadings for this model.

6.1 Estimates and model fit

Three factor spanned models Estimates with standard errors for the $\mathbb{A}_0(3)$ and $\mathbb{A}_1(3)$ models are included in Table 2. These two models have historically drawn most of the attention in the term structure literature. For the $\mathbb{A}_0(3)$ model, the concentrated likelihood drops the number of parameters entering the numerical optimizer from 22 to 10 (excluding Ω), while this number drops from 24 to 16 for the $\mathbb{A}_1(3)$ model. The likelihood for the $\mathbb{A}_0(3)$ model at 37080.94 is significantly higher than the $\mathbb{A}_1(3)$ model at 36647.69.

Surprisingly, among the models that we estimated, the $\mathbb{A}_3(3)$ model has the highest likelihood with a value of 37385.28, substantially higher than the $\mathbb{A}_0(3)$ model. We report estimates for this model on the right panel of Table 3. In order to satisfy the admissibility restrictions, all values in Σ_h, $\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}\Sigma_h$ and $\Sigma_h^{-1}\Phi_h\Sigma_h$ must be non-negative.[15] We also impose the discrete-time equivalent of the Feller condition, i.e. $\nu_{h,i} > 1$ and $\nu_{h,i}^{\mathbb{Q}} > 1$ for $i = 1, \ldots, H$.[16] In the $\mathbb{A}_3(3)$ model, we found that some of these parameters were near their boundaries. We fix them at the boundary when calculating the standard errors. For comparison purposes, we also report on the left panel of Table 3 estimates of an $\mathbb{A}_3(3)$ model where the matrices Φ_h and $\Phi_h^{\mathbb{Q}}$ are restricted to be diagonal in which case the estimated parameters are not close to the boundaries. Finally, we note that the $\mathbb{A}_2(3)$ model (estimates not reported for brevity) had the lowest likelihood among the models we estimated with a value of 36120.34.

The bond loadings for the $\mathbb{A}_0(3)$, $\mathbb{A}_1(3)$, and diagonal $\mathbb{A}_3(3)$ models are almost the same, see Figure 1. This happens for two reasons. First, the functional form of the bond loading recursions are the same up to Jensen's inequality. Second, the size of the measurement errors in the cross-section of yields are small relative to the magnitude of the time series shocks causing an efficient estimator (like maximum likelihood) to emphasize the fit of the cross section. The estimated factors of these models have high correlations (ranging from 0.95 to 1). Interestingly, the bond loadings for the unrestricted $\mathbb{A}_3(3)$ model appear to be non-stationary, even though the eigenvalues of $\Phi_h^{\mathbb{Q}}$ for this model are inside the unit circle and are nearly identical to the other models.[17] The average pricing errors are similar in these models, with the more restricted diagonal $\mathbb{A}_3(3)$ model having larger values. The measurement errors are about 22 basis points for 3 months, 12 for 2 years, 12 for 3 years and 11 for 4 years. These errors have the same magnitude as those reported in Ang and Piazzesi(2003).

Differences between these models are largely driven by their ability to fit the time series

[15] Multiple non-Gaussian factors tend to make the inversion problem more complicated. We solve this with randomized starting values, and use several iterations of "fminsearch" in MATLAB for each starting value to get an initial value that does not produce negative volatility. It takes roughly 20 trials (within a minute) to locate a starting value for the $\mathbb{A}_3(3)$ model.

[16] No arbitrage requires that the stochastic discount factor M_{t+1} is strictly positive which is guaranteed by the Feller condition, see Appendix D for details.

[17] Unlike Gaussian models, the bond loading recursions $\bar{b}_{n,h}$ for multi-factor non-Gaussian models in (6) are non-linear difference equations. The stability conditions for these equations appear to be determined by more than the eigenvalues of $\Phi_h^{\mathbb{Q}}$. We leave the conditions for stability of $\bar{b}_{n,h}$ for future research.

component of the likelihood. For example, in the $\mathbb{A}_1(3)$ model, the conditional mean under $\mathbb{P}$ is more restricted than the $\mathbb{A}_0(3)$ model as the Gaussian factors cannot enter the conditional mean of the non-Gaussian factors. The economic implication of this restriction for the $\mathbb{A}_1(3)$ model is that the level factor does not depend on the past values of slope and curvature factors. This is apparently counterintuitive. If the slope is high in the last period, i.e., the long rate is much higher than the short rate (more than explained by compensating for risk), then it means the market expects the short rate will increase in the future. On average, the next periods' short rate or level will increase.

Four factor USV models The two best fitting USV models are the $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ and $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ models, whose estimates are in Table 4. See Section 4.2.2 for the definition of the models. We calculate the likelihood for USV models by the particle filter, see Creal(2012). The likelihood for the best fitting $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model is 37331.25, which is lower than the unrestricted $\mathbb{A}_3(3)$ model but substantially higher than other spanned models.

The bond loadings for USV models are not as flexible compared to spanned models due to the USV restrictions explained in Section 4.2.2. The first USV model constrains the two largest eigenvalues of $\Phi^{\mathbb{Q}}$ for the level and slope factors to be related as ϕ and ϕ^2. Without any restrictions on the eigenvalues of $\Phi^{\mathbb{Q}}$ as in the $\mathbb{A}_0(3)$ and $\mathbb{A}_1(3)$ models, these are estimated to be 0.995 and 0.954. In the $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model, if the level factor has $\phi = 0.995$, the USV restriction would require the slope factor to be more persistent $\phi^2 = 0.990$ than it would be without the restriction (0.954). Conversely, if the model tried to fit the slope factor first, the USV restriction would not allow the level factor to be persistent enough. As a compromise, the two eigenvalues in the $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model are closer to each other than what the data would like with estimated values $\phi = 0.9868$ and $\phi^2 = 0.9738$. Consequently, the loadings on the level and slope factors in this model lie in between the loadings for spanned models, see Figure 1. Due to the more restrictive loadings, it is not

Table 3: Maximum likelihood estimates for two $\mathbb{A}_3(3)$ models.

diagonal	LLF =	36949.94		flexible	LLF =	37385.28	
$\Sigma_h \nu_h$				$\Sigma_h \nu_h$			
5.48e-05	2.75e-04	3.27e-04		5.35e-07	4.45e-05	9.35e-06	
—	—	—		—	—	—	
ν_h				ν_h			
7.2325	10.7472	5.3456		1	3.1176	1	
(5.2324)	(4.9148)	(1.2905)		—	(2.7588)	—	
Φ_h				Φ_h			
0.9918	0	0		0.9587	0.0010	0	
(0.0056)	—	—		(0.0102)	(0.0003)	—	
0	0.9511	0		0.3109	0.9133	0.5975	
—	(0.0136)	—		(0.2807)	(0.0259)	(0.1841)	
0	0	0.8133		0	0.0226	0.7876	
—	—	(0.0298)			(0.0018)	(0.0408)	
$\Sigma_h \nu_h^{\mathbb{Q}}$				$\Sigma_h \nu_h^{\mathbb{Q}}$			
2.01e-05	3.10e-04	6.63e-04		8.10e-07	1.43e-05	9.35e-06	
—	—	—		—	—	—	
$\nu_h^{\mathbb{Q}}$			δ_0	$\nu_h^{\mathbb{Q}}$			δ_0
2.6464	12.0937	10.8396	-0.0070	1.5151	1	1	-9.12e-04
(0.7721)	(5.2723)	(1.9412)	(0.0016)	(1.7822)	—	—	(4.82e-04)
$\Phi_h^{\mathbb{Q}}$				$\Phi_h^{\mathbb{Q}}$			
0.9996	0	0		0.9890	0.0002	0	
(0.0005)	—	—		(0.0014)	(0.0000)	—	
0	0.9479	0		0	0.8850	1.3287	
—	(0.0018)	—		—	(0.0333)	(0.2776)	
0	0	0.4837		0.4968	0.0206	0.6122	
—	—	(0.0239)		(0.0964)	(0.0047)	(0.0490)	
Σ_h				Σ_h			
7.58e-06	0	0		5.35e-07	0	0	
(1.27e-06)	—	—		(5.84e-08)	—	—	
0	2.56e-05	0		0	1.43e-05	0	
—	(6.60e-06)	—		—	(3.05e-06)	—	
0	0	6.12e-05		0	0	9.35e-06	
—	—	(5.95e-06)		—	—	(2.43e-06)	
$\delta_{1,h}$				$\delta_{1,h}$			
1	1	-1		1	1	-1	
—	—	—		—	—	—	
$\sqrt{\mathrm{diag}(\Omega)} \times 1200$				$\sqrt{\mathrm{diag}(\Omega)} \times 1200$			
3 m	2 yr	3 yr	4 yr	3 m	2 yr	3 yr	4 yr
0.2247	0.1324	0.1286	0.1112	0.2230	0.1274	0.1236	0.1077

Maximum likelihood estimates with asymptotic quasi-maximum likelihood standard errors. Left: $\mathbb{A}_3(3)$ model with diagonal matrices $\Phi_h^{\mathbb{Q}}$ and Φ_h. Right: $\mathbb{A}_3(3)$ model. The identifying restrictions $\mu_h = 0$, and $\delta_{1,h} = (1, 1, -1)$ are imposed during estimation.

surprising that the average pricing errors for this model are larger than for spanned models. Moreover, the Gaussian factors whose eigenvalues share the relationship ϕ and ϕ^2 are highly correlated with a correlation of -0.928. This number is 0.43 in absolute value for the $\mathbb{A}_0(3)$ and $\mathbb{A}_1(3)$ models for example.

Estimation of an $\mathbb{A}_0(3)$ that includes the same restrictions on $\Phi_g^{\mathbb{Q}}$ as the $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model but no stochastic volatility has a likelihood of 37040.5. This indicates that this

Table 4: Maximum likelihood estimates for two $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$ models.

$G = 3, H = 1$	LLF =	37331.25		$G = 3, H = 1$	LLF =	37241.05	
μ_g				μ_g			
2.11e-04	-1.88e-04	-2.37e-04		7.42e-04	-1.17e-03	8.49e-06	
(5.04e-07)	(3.66e-07)	(2.39e-06)		(8.80e-05)	(8.42e-05)	(5.76e-06)	
Φ_g				Φ_g			
1.0724	0.1181	0.1107		1.1316	0.3251	0.0520	
(0.0020)	(0.0008)	(0.0009)		(0.0427)	(0.1009)	(0.0161)	
-0.0788	0.8736	-0.0555		-0.2881	0.4463	-0.0923	
(0.0008)	(0.0015)	(0.0005)		(0.0213)	(0.0715)	(0.0149)	
-0.0321	-0.0372	0.6347		0.0348	0.0349	0.9995	
(0.0000)	(0.0001)	(0.0070)		(0.0162)	(0.0233)	(0.0029)	
$\mu_g^{\mathbb{Q}}$			δ_0	$\mu_g^{\mathbb{Q}}$			δ_0
0	0	0	0.0061	0	0	0	0.0115
—	—	—	(0.0000)	—	—	—	(0.0007)
$\Phi_g^{\mathbb{Q}}$				$\Phi_g^{\mathbb{Q}}$			
0.9868	0.9738	0.4931		0.9073	0.8231	0.9967	
(0.0004)	—	(0.0019)		(0.0045)	—	(0.0004)	
$\Sigma_{0,g}$				$\Sigma_{0,g}$			
1.16e-03	0	0		1.21e-03	0	0	
(2.41e-06)	—	—		(7.24e-05)	—	—	
-1.23e-03	2.44e-04	0		-1.30e-03	2.13e-05	0	
(2.83e-06)	(1.14e-07)	—		(4.63e-06)	(4.40e-05)	—	
5.04e-05	-4.27e-04	1.66e-10		-3.67e-04	7.05e-05	3.31e-04	
(9.23e-08)	(2.10e-06)	(3.90e-13)		(1.06e-04)	(1.43e-04)	(9.73e-05)	
$\Sigma_{1,g}$				$\Sigma_{1,g}$			
7.05e-04	0	0		8.35e-04	0	0	
(1.44e-06)	—	—		(4.03e-05)	—	—	
0	0	0		0	0	0	
—	—	—		—	—	—	
0	0	0		0	0	0	
—	—	—		—	—	—	
ν_h	Φ_h	Σ_h		ν_h	Φ_h	Σ_h	
1	0.9598	0.0100		1.0658	0.9539	0.0100	
—	(0.0086)	—		(0.0319)	(0.0101)	—	
$\delta_{1,g}$				$\delta_{1,g}$			
1	1	1		1	1	1	
—	—	—		—	—	—	
$\sqrt{\text{diag}(\Omega)} \times 1200$				$\sqrt{\text{diag}(\Omega)} \times 1200$			
3 m	2 yr	3 yr	4 yr	3 m	2 yr	3 yr	4 yr
0.2266	0.1298	0.1316	0.1101	0.2266	0.1298	0.1316	0.1101

Maximum likelihood estimates with quasi-maximum likelihood standard errors for two unspanned models. Left: $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model. Right: $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ model. The USV and the identifying restrictions $\Sigma_h = 0.01$, $\mu_g^{\mathbb{Q}} = 0$, and $\delta_{1,g} = (1, 1, 1)$ are imposed during estimation.

single USV restriction is rejected relative to the benchmark Gaussian $\mathbb{A}_0(3)$ model of Table 2 by a likelihood ratio test. On the other hand, the addition of unspanned stochastic volatility factors increases the likelihood by 37331.25 - 37040.5 = 290.75. The primary source is a significantly better fit of the dynamics of volatility, as we discuss further below.

When the USV restriction is imposed on the second and third largest eigenvalues as in the $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ model of Table 4, it constrains the bond loadings of the slope and curvature factors because the eigenvalues in $\Phi_g^{\mathbb{Q}}$ associated with these factors now share the relationship ϕ and ϕ^2. This restriction causes them to be closer than they would otherwise have been if left unrestricted. The $\mathbb{U}_1(4)(\phi, \phi^2, \phi^4)$ model imposes even stronger restrictions on $\Phi_g^{\mathbb{Q}}$ because it only has a single free parameter. It has likelihood 36889.44 (parameter estimates not reported), which is well below the benchmark $\mathbb{A}_0(3)$ model.

Four factor spanned models Finally, we consider two four factor spanned models: the Gaussian $\mathbb{A}_0(4)$ model and the non-Gaussian $\mathbb{A}_1(4)$ model. There are a total of 35 parameters in the $\mathbb{A}_0(4)$ model and only 20 of these parameters enter the numerical optimizer, while there are 39 parameters in the $\mathbb{A}_1(4)$ model and 15 of these can be concentrated out. Parameter estimates and standard errors for both models are in Table 5. While adding another Gaussian factor increases the likelihood relative to the $\mathbb{A}_0(3)$ model to 37194.34 for the $\mathbb{A}_0(4)$, the additional factor is not as important as adding an USV factor.

Repeated eigenvalues When we estimated the $\mathbb{A}_1(4)$ model, we found that it had repeated eigenvalues and our estimates in Table 5 have imposed them using the Jordan decomposition. If we use a diagonal matrix for $\Phi^{\mathbb{Q}}$, the matrix B_1 will be singular and one element in $\Phi^{\mathbb{Q}}$ is unidentified.[18] To illustrate this point, Table 6 reports different sets of parameter values all with the same likelihood. Across the four local maxima, the values of $\Phi_h^{\mathbb{Q}}$, the last eigenvalue of $\Phi_g^{\mathbb{Q}}$, and the log-likelihood function are almost identical but the first two eigenvalues of $\Phi_g^{\mathbb{Q}}$ vary across different optima. The last column of Table 6 shows the results when we impose repeated eigenvalues (from the model reported in Table 2). The estimates of $\Phi_h^{\mathbb{Q}}$, the last eigenvalue of $\Phi_g^{\mathbb{Q}}$ and the likelihood function have the same values as before. However, the first two eigenvalues of $\Phi_g^{\mathbb{Q}}$ are identical by definition and are equal to the average of the first two eigenvalues in those local maxima. The log-likelihood value also does not change.

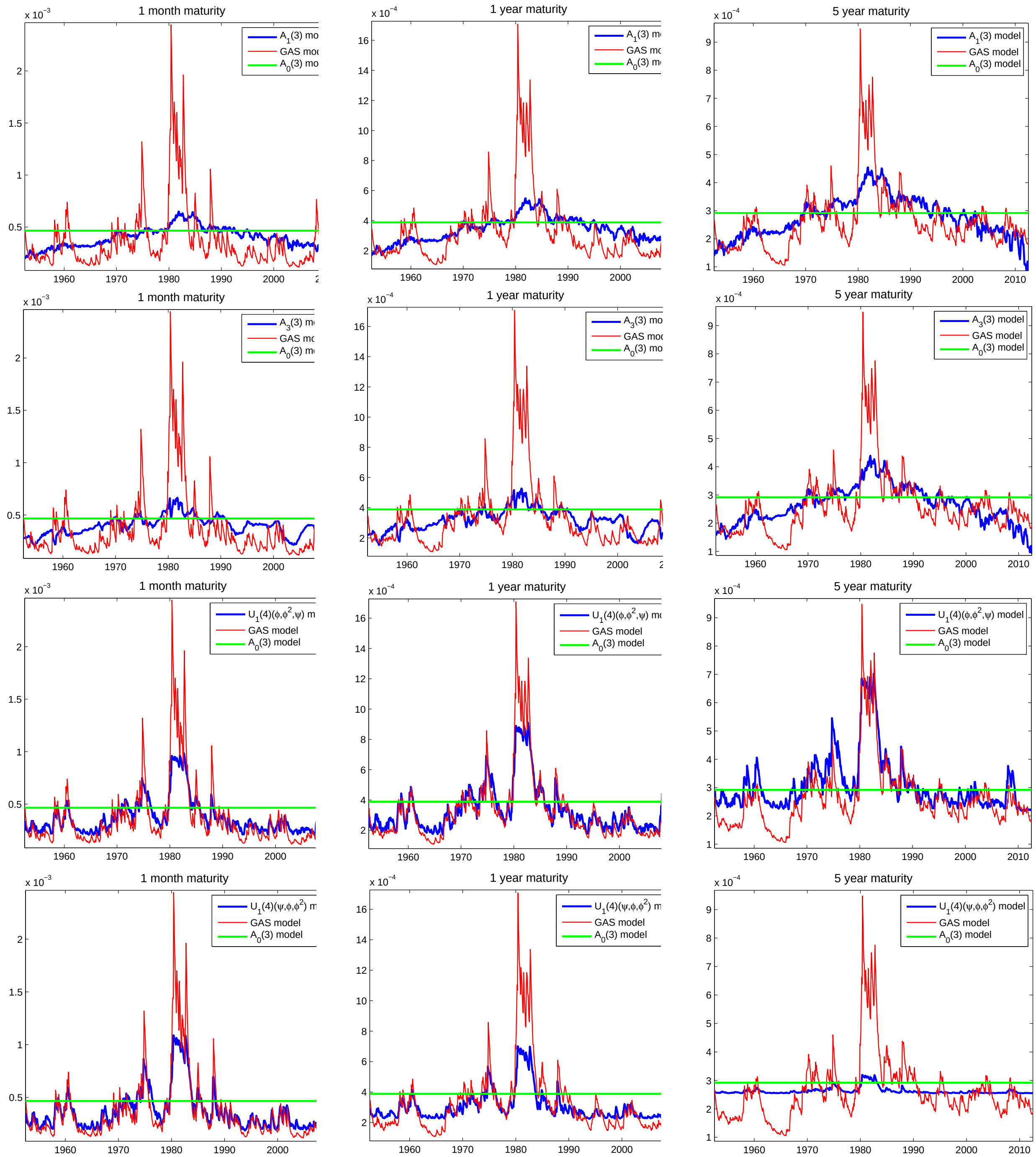
[18]For a matrix with repeated eigenvalues, the Jordan decomposition imposes the additional restrictions necessary to obtain identification. With two repeated eigenvalues in $\Phi_g^{\mathbb{Q}}$, the Jordan decomposition is $\text{vec}\left(\Phi_g^{\mathbb{Q}}\right)' = (\lambda_1 \ 1 \ 0 \ ; 0 \ \lambda_1 \ 0 \ ; 0 \ 0 \ \lambda_2)$ where λ_1 and λ_2 are the unique eigenvalues.

Table 5: Maximum likelihood estimates for the $\mathbb{A}_0(4)$ and $\mathbb{A}_1(4)$ models.

$G=4, H=0$	LLF =	37195.84		$G=3, H=1$	LLF =	36729.22		
μ_g				$\Sigma_h \nu_h$	μ_g			ν_h
3.89e-04	-9.73e-04	1.28e-03	-8.19e-04	5.64e-05	3.07e-04	-3.76e-05	-3.10e-04	2.6780
(1.73e-04)	(2.11e-04)	(3.07e-04)	(3.68e-04)	—	(1.76e-04)	(1.84e-05)	(1.79e-04)	(1.8951)
Φ_g				Φ_h				
1.0336	0.0874	-0.0142	0.0908	0.9901				
(0.0352)	(0.0536)	(0.0475)	(0.0646)	(0.0073)				
				Φ_{gh}	Φ_g			
-0.0783	0.8338	0.2044	-0.0788	0.0037	0.8668	1.1695	0.0911	
(0.0626)	(0.0893)	(0.1298)	(0.1734)	(0.0260)	(0.0644)	(0.4487)	(0.0974)	
0.0889	0.1541	0.6935	0.1926	0.0016	0.0153	0.8156	0.0028	
(0.0766)	(0.1312)	(0.1731)	(0.2259)	(0.0031)	(0.0062)	(0.0456)	(0.0097)	
-0.0851	-0.1405	-0.0398	0.5456	-0.0352	-0.0095	-0.1235	0.6508	
(0.0528)	(0.0901)	(0.0924)	(0.1377)	(0.0257)	(0.0555)	(0.4361)	(0.0974)	
$\mu_g^{\mathbb{Q}}$				$\Sigma_h \nu_h^{\mathbb{Q}}$	$\mu_g^{\mathbb{Q}}$			$\nu_h^{\mathbb{Q}}$
0	0	0	0	2.71E-05	0	0	0	1.2839
—	—	—	—	—	—	—	—	(0.3516)
$\Phi_g^{\mathbb{Q}}$				$\Phi_h^{\mathbb{Q}}$	$\Phi_g^{\mathbb{Q}}$			
0.9922	0.9604	0.8764	0.6964	0.9952	0.9121	—	0.7021	
(0.0024)	(0.0116)	(0.0303)	(0.0504)	(0.0011)	(0.0075)	—	(0.0324)	
$\Sigma_{0,g}$				Σ_h				
6.94e-04	0	0	0	2.11e-05				
(2.33e-04)	—	—	—	(3.75e-06)				
				Σ_{gh}	$\Sigma_{0,g}$			
-1.47e-03	9.77e-04	0	0	0.9248	8.37e-04	0	0	
(2.80e-04)	(3.98e-04)	—	—	(0.4737)	(2.90e-04)	—	—	
1.66e-03	-1.39e-03	8.74e-04	0	-0.2171	-9.03e-05	7.43e-13	0	
(1.96e-04)	(4.33e-04)	(3.21e-04)	—	(0.0333)	(2.69e-05)	(4.22e-12)	—	
-8.06e-04	5.65e-04	-7.18e-04	4.04e-04	-1.5635	-7.96e-04	9.41e-11	4.03e-10	
(3.69e-04)	(2.68e-04)	(3.59e-04)	(2.68e-05)	(0.4198)	(2.72e-04)	(1.75e-11)	(6.48e-11)	
					$\Sigma_{1,g}$			
					1.04e-02	0	0	
					(2.48e-03)	—	—	
					-6.75e-04	8.15e-04	0	
					(2.75e-04)	(7.96e-05)	—	
					-9.01e-03	1.12e-03	4.68e-03	
					(2.69e-03)	(4.10e-04)	(3.04e-04)	
δ_0				δ_0				
3.92e-03				-4.44e-04				
(6.45e-04)				(3.67e-04)				
$\delta_{1,g}$				$\delta_{1,h}$	$\delta_{1,g}$			
1	1	1	1	1	1	1	1	
—	—	—	—	—	—	—	—	
$\sqrt{\text{diag}(\Omega)} \times 1200$				$\sqrt{\text{diag}(\Omega)} \times 1200$				
3 m	3 yr	4 yr		3 m	3 yr	4 yr		
0.2390	0.1013	0.0982		0.2408	0.1005	0.0981		

Maximum likelihood estimates with quasi-maximum likelihood standard errors. Left: Gaussian $\mathbb{A}_0(4)$ model. Right: non-Gaussian $\mathbb{A}_1(4)$ model. The identifying restrictions $\mu_g^{\mathbb{Q}} = 0, \delta_{1,g} = \iota$, and $\delta_{1,h} = 1$ are imposed during estimation.

Figure 2: Estimated conditional volatility of yields from four different affine models.



Estimated conditional volatility of yields from different affine models compared to the multivariate generalized autoregressive score model and the $\mathbb{A}_0(3)$ *with constant volatility. Across the columns are the one month, one year, and 5 year maturities. Top row:* $A_1(3)$ *model. Second row:* $A_3(3)$ *model. Third row:* $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ *model. Bottom row:* $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ *model.*

Table 6: Repeated Eigenvalues

	Local 1	Local 2	Local 3	Repeated
$\Phi_h^{\mathbb{Q}}$	0.9952	0.9952	0.9952	0.9952
$\Phi_g^{\mathbb{Q}}$	0.9130	0.9164	0.9126	0.9121
	0.9112	0.9075	0.9115	–
	0.7021	0.7025	0.7021	0.7021
LLF	36729.22	36729.20	36729.22	36729.22

Estimates from the $\mathbb{A}_1(4)$ model when repeated eigenvalues are not imposed compared to when they are. The table illustrates how this can create identification problems in affine models.

6.2 Volatility

The blue lines in Figure 2 are the conditional volatilities from four different models; across the rows are the $\mathbb{A}_1(3), \mathbb{A}_3(3)$, $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ and $\mathbb{U}_1(4)(\psi > \phi > \psi^2)$ models.[19] The three columns represent maturities of 1 month, 1 year, and 5 years. The volatilities for the USV models are the filtered (one-sided) estimates calculated from the particle filter. To provide a point of comparison, we also plot in these graphs the unconditional volatility from the $\mathbb{A}_0(3)$ model in green and in red estimates of the conditional volatilities from the multivariate generalized autoregressive score model of Creal, Koopman, and Lucas(2011) and Creal, Koopman, and Lucas(2013).[20]

The estimated volatilities from spanned models (first two rows) are much less volatile than GAS volatility. A similar observation was made by Collin-Dufresne, Goldstein, and Jones(2009). With only a single volatility factor, the $\mathbb{A}_1(3)$ model does not fit yield volatility at any maturity. The $\mathbb{A}_3(3)$ model adds flexibility with two more factors, which improves the fit for volatility at shorter horizons, especially for the recent episode of low volatility. At longer horizons, the fit to volatility appears to be the same as the $\mathbb{A}_1(3)$ model. This is because the estimated level (volatility) factor is similar across both the $\mathbb{A}_1(3)$ and $\mathbb{A}_3(3)$ models, and the long term volatility loads mostly on it. Ultimately, spanned models lack

[19]The volatilities from the remaining models have the same qualitative features and are not shown.

[20]The generalized autoregressive score model with time-varying covariance matrix is similar to a multivariate GARCH model. To make the volatilities of yields comparable across models, we use a VAR(1) for the conditional mean of yields $Y_t^{(1)}$ and allow the errors to have time-varying volatilities and correlations.

flexibility because the non-Gaussian state variables h_t serve a dual role: they must simultaneously fit the conditional mean and variance. The maximum likelihood estimator chooses the parameter vector θ to fit the conditional mean first before fitting the conditional variance.

USV models are designed to fit volatility by separating the role of Gaussian and non-Gaussian factors. The $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ model performs much better than spanned models at fitting the volatility across different maturities, although it only has a single stochastic volatility factor. It does a particularly good job for short and medium maturities. At longer maturities, it fits the high volatility periods of the early 1980's well, although misses the period of low volatility early in the sample.

Our results suggest, however, that not all USV models fit yield volatility equally well and researchers must be careful when choosing USV restrictions. We demonstrate this point by comparing the performance of the two best USV models $\mathbb{U}_1(4)(\phi > \phi^2 > \psi)$ and $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$. Instead of having stochastic volatility on the level factor, the $\mathbb{U}_1(4)(\psi > \phi > \phi^2)$ model has stochastic volatility on the slope factor. It fits the volatility at shorter maturities equally well but, at longer maturities, it exhibits almost no stochastic volatility. This is because the eigenvalue associated with the slope factor that has stochastic volatility is $\phi = 0.9073$ as opposed to $\phi = 0.9868$ in the first USV model. The bond loadings on this factor decay rapidly as maturity increases, meaning that long maturities have no stochastic volatility. To summarize, although USV models are designed to fit the volatility, the restrictions needed to impose USV are not unique and the choice of which restriction to impose is not innocuous.

7 Conclusion

We provide new estimation procedures for non-Gaussian affine term structure models with spanned or unspanned stochastic volatility. The new estimation approach for spanned models leverages the fact that many of the parameters can be concentrated out of the likelihood function. By optimizing the concentrated likelihood, it provides exactly the same solution as maximizing the original likelihood. But, it improves the estimation dramatically by reducing the number of parameters that need to be maximized numerically. We demonstrate the improvement in performance with our method. Using our procedure, we show what characteristics of spanned non-Gaussian models cause local maxima to exist and how alternative local maxima may have dramatically different economic implications. We apply a similar idea to the concentrated likelihood to estimate unspanned models.

Estimating a wide range of popular models, we find that models with spanned volatility have similar cross sectional fit for yields. They fit better than the USV models, because the latter impose restrictions on the cross section in order to introduce USV factors. Models with USV fit the volatility better by design. The choice of how to impose USV restrictions is not innocuous as some USV models can severely limit yield volatility at particular maturities.

USV models make an effort to fit the volatility of yields. Future work on term structure models aiming to fit both the conditional mean and volatility of yields simultaneously will likely require (1) multiple unspanned volatility factors, and (2) the ability to relax the restrictions that USV impose on the cross section.

References

Abramowitz, Milton, and Irene A. Stegun (1964) *Handbook of Mathematical Functions* Dover Publications Inc, New York, NY.

Adrian, Tobias, Richard K. Crump, and Emanuel Moench (2012) "Pricing the term structure with linear regressions." *Journal of Financial Economics* 110, 110–138.

Aït-Sahalia, Yacine (2008) "Closed-form likelihood expansions for multivariate diffusions." *The Annals of Statistics* 36, 906–937.

Aït-Sahalia, Yacine, and Robert L. Kimmel (2010) "Estimating affine multifactor term structure models using closed-form likelihood expansions." *Journal of Financial Eco-*

nomics 98, 113–144.

Andersen, Torben, and Luca Benzoni (2010) "Do bonds span volatility risk in the U.S. treasury market? A specification test for affine term structure models." *The Journal of Finance* 65, 603–653.

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

Bauer, Michael D. (2014) "Bayesian Estimation of Dynamic Term Structure Models under Restrictions on Risk Pricing" Federal Reserve Bank of San Francisco Working Paper.

Bauer, Michael D., Glenn D. Rudebusch, and Jing Cynthia Wu (2012) "Correcting estimation bias in dynamic term structure models." *Journal of Business and Economic Statistics* 30, 454–467.

Cheridito, Patrick, Damir Filipovic, and Robert L. Kimmel (2007) "Market price of risk specifications for affine models: theory and evidence" *Journal of Financial Economics* 84, 123–170.

Christensen, Jens H.E., Francis X. Diebold, and Glenn D. Rudebusch (2011) "The affine arbitrage-free class of Nelson-Siegel term structure models." *Journal of Econometrics* 164, 4–20.

Cochrane, John H., and Monika Piazzesi (2008) "Decomposing the yield curve" Unpublished manuscript, Booth School of Business, University of Chicago.

Collin-Dufresne, Pierre, and Robert S. Goldstein (2002) "Do bonds span the fixed income markets? Theory and evidence for unspanned stochastic volatility." *The Journal of Finance* 57, 1685–1730.

Collin-Dufresne, Pierre, Robert S. Goldstein, and Charles Jones (2008) "Identification of maximal affine term structure models." *The Journal of Finance* 63, 743–795.

Collin-Dufresne, Pierre, Robert S. Goldstein, and Charles Jones (2009) "Can the volatility of interest rates be extracted from the cross section of bond yields? An investigation of unspanned stochastic volatility." *Journal of Financial Economics* 94, 47–66.

Cox, John C., Jonathan E. Ingersoll, and Stephen A. Ross (1985) "A theory of the term structure of interest rates" *Econometrica* 53, 385–407.

Creal, Drew D. (2012) "A survey of sequential Monte Carlo methods for economics and finance." *Econometric Reviews* 31, 245–296.

Creal, Drew D., Siem Jan Koopman, and André Lucas (2011) "A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations." *Journal of Business and Economic Statistics* 29, 552–563.

Creal, Drew D., Siem Jan Koopman, and André Lucas (2013) "Generalized autoregressive score models with applications." *Journal of Applied Econometrics* 28, 777–795.

Dai, Qiang, and Kenneth J. Singleton (2000) "Specification analysis of affine term structure models." *The Journal of Finance* 55, 1943–1978.

de Jong, Piet (1991) "The diffuse Kalman filter" *The Annals of Statistics* 19, 1073–83.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977) "Maximum likelihood from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society, Series B* 39, 1–38.

Diebold, Francis X., and Glenn D. Rudebusch (2013) *Yield Curve Modeling and Forecasting.* Princeton University Press, Princeton, NJ.

Diez de Los Rios, Antonio (2013) "A new linear estimator for Gaussian dynamic term structure models." Unpublished manuscript, Bank of Canada.

Duffee, Gregory R. (2002) "Term premia and interest rate forecasts in affine models" *The Journal of Finance* 57, 405–443.

Duffee, Gregory R. (2011) "Information in (and not in) the term structure" *The Review of Financial Studies* 24, 2895–2934.

Duffee, Gregory R. (2012) "Bond pricing and the macroeconomy" Working paper, Johns Hopkins University.

Duffie, Darrell, and Rui Kan (1996) "A yield factor model of interest rates" *Mathematical Finance* 6, 379–406.

Durbin, James, and Siem Jan Koopman (2012) *Time Series Analysis by State Space Methods* Oxford University Press, Oxford, UK 2 edition.

Fama, Eugene F., and Robert R. Bliss (1987) "The information in long maturity forward rates" *American Economic Review* 77, 680–692.

Godsill, Simon J., Arnaud Doucet, and Michael West (2004) "Monte Carlo smoothing for nonlinear time series." *Journal of the American Statistical Association* 99, 156–168.

Gouriéroux, Christian, and Joann Jasiak (2006) "Autoregressive gamma processes." *Journal of Forecasting* 25, 129–152.

Gouriéroux, Christian, Joann Jasiak, and Razvan Sufana (2009) "The Wishart autoregressive process of multivariate stochastic volatility." *Journal of Econometrics* 150, 167–181.

Gouriéroux, Christian, Alain Monfort, and Vassilis Polimenis (2002) "Affine term structure models." Institut National de La Statistique et des Etudes Economiques.

Gürkaynak, Refet S., and Jonathan H. Wright (2012) "Macroeconomics and the term structure" *Journal of Economic Literature* 50, 331–367.

Hamilton, James D., and Jing Cynthia Wu (2012) "Identification and estimation of Gaussian affine term structure models." *Journal of Econometrics* 168, 315–331.

Hamilton, James D., and Jing Cynthia Wu (2014) "Testable implications of affine term structure models." *Journal of Econometrics* 178, 231–242.

Heidari, Massoud, and Liuren Wu (2003) "Are interest rate derivatives spanned by the term structure of interest rates?" *The Journal of Fixed Income* 13, 75–86.

Jacquier, Eric, Michael Johannes, and Nicholas G. Polson (2007) "MCMC maximum likelihood for latent state models" *Journal of Econometrics* 137, 615–640.

Joslin, Scott (2010) "Can unspanned stochastic volatility models explain the cross section of bond volatilities?" Working paper, University of Southern California, Marshall School of Business.

Joslin, Scott, Kenneth J. Singleton, and Haoxiang Zhu (2011) "A new perspective on Gaussian affine term structure models" *The Review of Financial Studies* 27, 926–970.

Kim, Don H., and Athanasios Orphanides (2005) "Term structure estimation with survey data on interest rate forecasts." Federal Reserve Board, Finance and Economics Discussion Series 2005-48.

Kim, Don H., and Jonathan H. Wright (2005) "An arbitrage-free three-factor term structure model and the recent behavior of long-term yields and distant-horizon forward rates." Federal Reserve Board, Finance and Economics Discussion Series 2005-33.

Le, Anh, Kenneth J. Singleton, and Qiang Dai (2010) "Discrete-time affine term structure models with generalized market prices of risk." *The Review of Financial Studies* 23, 2184–2227.

Li, Haitao, and Feng Zhao (2006) "Unspanned stochastic volatility: evidence from hedging interest rate derivatives." *The Journal of Finance* 61, 341–378.

Malik, Sheheryar, and Michael K. Pitt (2011) "Particle filters for continuous likelihood evaluation and maximisation." *Journal of Econometrics* 165, 190–209.

Meng, Xiao-Li, and Donald B. Rubin (1993) "Maximum likelihood estimation via the ECM algorithm: A general framework." *Biometrika* 80, 267–278.

Piazzesi, Monika (2010) "Affine term structure models" in *Handbook of Financial Econometrics*, edited by Y. Ait-Sahalia and L. P. Hansen Elsevier, New York pages 691–766.

Richard, Jean Francois, and Wei Zhang (2007) "Efficient high-dimensional importance sampling" *Journal of Econometrics* 141, 1385–1411.

Trolle, Anders B., and Eduardo S. Schwartz (2009) "A general stochastic volatility model for the pricing of interest rate derivatives." *The Review of Financial Studies* 22, 2007–2057.

Wei, Greg C. G., and Martin A. Tanner (1990) "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms." *Journal of the American Statistical Association* 85, 699–704.

White, Halbert L. (1982) "Maximum likelihood estimation of misspecified models." *Econometrica* 50, 1–25.

Appendix A Distributions

We define the distributions found in the paper. The notation for them is local to the appendix.

Appendix A.1 Gamma and multivariate gamma distributions

A gamma r.v. $w_{t+1} \sim \text{Gamma}(\nu_h, \kappa)$ has p.d.f $p(w_{t+1}|\nu_h, \kappa) = \frac{1}{\Gamma(\nu_h)} w_{t+1}^{\nu_h - 1} \kappa^{-\nu_h} \exp\left(-\frac{w_{t+1}}{\kappa}\right)$ and $\mathbb{E}(w_{t+1}) = \nu_h \kappa$ and $\mathbb{V}(w_{t+1}) = \nu_h \kappa^2$. The Laplace transform is $\mathbb{E}[\exp(u w_{t+1})] = \left(\frac{1}{1-\kappa u}\right)^{\nu_h}$, which exists only if $\kappa u < 1$.

A multivariate gamma random vector $h_{t+1} \sim \text{Mult. Gamma}(\nu_h, \Sigma_h, \mu_h)$ can be obtained by shifting and rotating a vector of uncorrelated gamma r.v.'s. Let $h_{t+1} = \mu_h + \Sigma_h w_{t+1}$ where w_{t+1} is an $H \times 1$ vector with elements $w_{i,t+1} \sim \text{Gamma}(\nu_{h,i}, 1)$ for $i = 1, \ldots, H$. The $H \times 1$ vector of (non-negative) location parameters is μ_h, Σ_h is a full rank $H \times H$ matrix of (non-negative) scale parameters, and $\nu_h > 0$ is a $H \times 1$ vector of shape parameters. The p.d.f of h_{t+1} can be determined by a standard change-of-variables

$$p(h_{t+1}|\nu_h, \Sigma_h, \mu_h) = |\Sigma_h^{-1}| \prod_{i=1}^{H} \frac{1}{\Gamma(\nu_{h,i})} \left(\mathrm{e}_i' \Sigma_h^{-1} [h_{t+1} - \mu_h]\right)^{\nu_{h,i} - 1} \exp\left(-\mathrm{e}_i' \Sigma_h^{-1} [h_{t+1} - \mu_h]\right)$$

where e_i is an $H \times 1$ unit vector. The mean and variance are $\mathbb{E}[h_{t+1}] = \mu_h + \Sigma_h \nu_h$ and $\mathbb{V}[h_{t+1}] = \Sigma_h \text{diag}(\nu_h) \Sigma_h'$. The Laplace transform is

$$\begin{aligned}
\mathbb{E}[\exp(u' h_{t+1})] &= \int_0^\infty \exp(u' h_{t+1})\, p(h_{t+1}|\nu_h, \Sigma_h, \mu_h)\, dh_{t+1} \\
&= \exp(u'\mu_h) \int_0^\infty \exp(u' \Sigma_h w_{t+1}) \prod_{i=1}^{H} \frac{1}{\Gamma(\nu_{h,i})} w_{i,t+1}^{\nu_{h,i}-1} \exp(-w_{t+1})\, dw_{t+1} \\
&= \exp(u'\mu_h) \prod_{i=1}^{H} \left(\frac{1}{1 - \mathrm{e}_i' \Sigma_h' u}\right)^{\nu_{h,i}} = \exp\left(u'\mu_h - \sum_{i=1}^{H} \nu_{h,i} \log[1 - \mathrm{e}_i' \Sigma_h' u]\right)
\end{aligned}$$

The Laplace transform exists only if $\mathrm{e}_i' \Sigma_h' u < 1$ for $i = 1, \ldots, H$.

Appendix A.2 Multivariate non-central gamma distributions

A $H \times 1$ non-central gamma (NCG) random vector $h_{t+1} \sim \text{Mult.-N.C.G.}(\nu_h, \Phi_h h_t, \Sigma_h, \mu_h)$ is a Poisson mixture of multivariate gamma r.v.'s as in (12)-(14). The process h_t remains positive and well-defined as long as $\mu_h \geq 0$, $\Sigma_h^{-1} \Phi_h \Sigma_h \geq 0$, $\nu_h > 1$, and elements of Σ_h cannot be negative. A standard multivariate NCG random variable (the discrete-time CIR process) is obtained by setting $\mu_h = 0$ and letting Σ_h be a diagonal matrix. Further properties of the univariate NCG process are described in Gouriéroux and Jasiak(2006).

As long as Σ_h has full rank, the p.d.f. can be found by integrating out the Poisson r.v.'s

$$\begin{aligned} p\left(h_{t+1}|\nu_h,\Phi_h h_t,\Sigma_h,\mu_h\right) &= |\Sigma_h^{-1}|\exp\left(-\sum_{i=1}^{H}\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]+\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right) \\ & \prod_{i=1}^{H}\left(\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]\right)^{\nu_{h,i}-1} \\ & \sum_{z_{i,t+1}=0}^{\infty}\frac{1}{\Gamma\left(\nu_{h,i}+z_{i,t+1}\right)}\frac{1}{z_{i,t+1}!}\left[\left(\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]\right)\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right)\right]^{z_{i,t+1}} \end{aligned}$$

Using the definition of the modified Bessel function of the first kind[21], the p.d.f. can be expressed as

$$\begin{aligned} p\left(h_{t+1}|\nu_h,\Phi_h h_t,\Sigma_h,\mu_h\right) &= |\Sigma_h^{-1}|\exp\left(-\sum_{i=1}^{H}\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]+\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right) \\ & \prod_{i=1}^{H}\left(\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]\right)^{\frac{\nu_{h,i}-1}{2}}\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right)^{-\frac{\nu_{h,i}-1}{2}} \\ & I_{\nu_{h,i}-1}\left(2\sqrt{\left(\mathrm{e}_i'\Sigma_h^{-1}\left[h_{t+1}-\mu_h\right]\right)\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right)}\right). \end{aligned}$$

The Laplace transform can be derived from the law of iterated expectations

$$\begin{aligned} \mathbb{E}\left[\exp\left(u'h_{t+1}\right)\right] &= \mathbb{E}_z\left(\mathbb{E}_{h|z}\left[\exp\left(u'h_{t+1}\right)\right]\right) = \mathbb{E}_z\left(\exp\left(u'\mu_h\right)\prod_{i=1}^{H}\left(\frac{1}{1-\mathrm{e}_i'\Sigma_h'u}\right)^{\nu_{h,i}+z_i}\right) \\ &= \exp\left(u'\mu_h\right)\prod_{i=1}^{H}\left(\frac{1}{1-\mathrm{e}_i'\Sigma_h'u}\right)^{\nu_{h,i}}\mathbb{E}_z\left(\prod_{i=1}^{H}\left(\frac{1}{1-\mathrm{e}_i'\Sigma_h'u}\right)^{z_i}\right) \\ &= \exp\left(u'\mu_h\right)\prod_{i=1}^{H}\left(\frac{1}{1-\mathrm{e}_i'\Sigma_h'u}\right)^{\nu_{h,i}}\prod_{i=1}^{H}\exp\left(\frac{\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_t-\mu_h\right]\right)\mathrm{e}_i'\Sigma_h'u}{1-\mathrm{e}_i'\Sigma_h'u}\right) \\ &= \exp\left(u'\mu_h+\sum_{i=1}^{H}\frac{\mathrm{e}_i'\Sigma_h'u}{1-\mathrm{e}_i'\Sigma_h'u}\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left(h_t-\mu_h\right)-\sum_{i=1}^{H}\nu_{h,i}\log\left(1-\mathrm{e}_i'\Sigma_h'u\right)\right) \end{aligned}$$

where $\mathrm{e}_i'\Sigma_h'u$ denotes the i-th element of the $H\times 1$ vector $\Sigma_h'u$. The Laplace transform exists only if $\mathrm{e}_i'\Sigma_h'u<1$ for $i=1,\ldots,H$.

Appendix A.3 Mixture of Gaussian and mult. NCG distributions

From standard results in statistics, the multivariate $(G\times 1)$ Gaussian r.v. $g_{t+1}\sim \mathrm{N}\left(g_{t+1}|\mu_g,\Sigma_g\Sigma_g'\right)$ has Laplace transform $\mathbb{E}\left[\exp\left(u'g_{t+1}\right)\right] = \exp\left(\mu_g'u+\frac{1}{2}u'\Sigma_g\Sigma_g'u\right)$ for any real $(G\times 1)$ vector u. Consider a $(G+H)\times 1$ vector $x_{t+1}=(h_{t+1}',g_{t+1}')'$ where h_{t+1} is an $H\times 1$ vector having a multivariate NCG

[21] This is defined as $I_\lambda(x)=\left(\frac{x}{2}\right)^\lambda\sum_{z=0}^{\infty}\frac{1}{\Gamma(\lambda+z+1)z!}\left(\frac{x^2}{4}\right)^z$, see Abramowitz and Stegun(1964).

distribution $p\left(h_{t+1}|\nu_h, \Phi_h h_t, \Sigma_h, \mu_h\right)$ and g_{t+1} is a $G \times 1$ vector of conditionally Gaussian r.v. $g_{t+1} \sim \mathrm{N}\left(\mu_g + \Sigma_{gh} h_{t+1}, \Sigma_g \Sigma_g'\right)$. Let $u = (u_h', u_g')'$ where u_h and u_g are $H \times 1$ and $G \times 1$ vectors, respectively. Using the law of iterated expectations, the Laplace transform is

$$\begin{aligned}
\mathbb{E}\left[\exp\left(u' x_{t+1}\right)\right] &= \mathbb{E}\left[\exp\left(u_g' g_{t+1}\right)\exp\left(u_h' h_{t+1}\right)\right] = \mathbb{E}_h\left[\mathbb{E}_{g|h}\left[\exp\left(u_g' g_{t+1}\right)\right]\exp\left(u_h' h_{t+1}\right)\right] \\
&= \mathbb{E}_h\left[\exp\left((\mu_g + \Sigma_{gh} h_{t+1})' u_g + \frac{1}{2} u_g' \Sigma_g \Sigma_g' u_g\right)\exp\left(u_h' h_{t+1}\right)\right] \\
&= \exp\left(u_g' \mu_g + \frac{1}{2} u_g' \Sigma_g \Sigma_g' u_g\right)\mathbb{E}_h\left[\exp\left(\left[u_g' \Sigma_{gh} + u_h'\right] h_{t+1}\right)\right] \\
&= \exp\left(u_g' \mu_g + \frac{1}{2} u_g' \Sigma_g \Sigma_g' u_g + u_{gh}' \mu_h - \sum_{i=1}^{H} \nu_{h,i} \log\left(1 - \mathrm{e}_i' \Sigma_h' u_{gh}\right)\right. \\
&\quad \left. + \sum_{i=1}^{H} \frac{\mathrm{e}_i' \Sigma_h' u_{gh}}{1 - \mathrm{e}_i' \Sigma_h' u_{gh}} \mathrm{e}_i' \Sigma_h^{-1} \Phi_h \left(h_t - \mu_h\right)\right)
\end{aligned}$$

where $u_{gh} = \Sigma_{gh}' u_g + u_h$ is an $H \times 1$ vector. The Laplace transform exists only if $\mathrm{e}_i' \Sigma_h' u_{gh} < 1$ for $i = 1, \ldots, H$. This is the key expression for solving for closed-form zero-coupon bond prices.

Appendix B Bond pricing

Bond prices can be solved by induction. Guess that bond prices are $P_t^n = \exp\left(\bar{a}_n + \bar{b}_{n,h}' h_t + \bar{b}_{n,g}' g_t\right)$ for some coefficients $\bar{a}_n$, $\bar{b}_{n,h}$, and $\bar{b}_{n,g}$. At maturity $n = 1$ when the payoff is $P_{t+1}^0 = 1$, we find

$$P_t^1 = \mathrm{E}_t^{\mathbb{Q}}\left[\exp\left(-r_t\right) P_{t+1}^0\right] = \exp\left(-\delta_0 - \delta_{1,h}' h_t - \delta_{1,g}' g_t\right)$$

such that $\bar{a}_1 = -\delta_0, \bar{b}_{1,g} = -\delta_{1,g}$ and $\bar{b}_{1,h} = -\delta_{1,h}$. Next, consider an n-period bond whose price in the next period is P_{t+1}^{n-1}. We find

$$\begin{aligned}
P_t^n &= \mathrm{E}_t^{\mathbb{Q}}\left[\exp\left(-r_t\right) P_{t+1}^{n-1}\right] = \mathrm{E}_t^{\mathbb{Q}}\left[\exp\left(-\delta_0 - \delta_{1,h}' h_t - \delta_{1,g}' g_t\right)\exp\left(\bar{a}_{n-1} + \bar{b}_{n-1,h}' h_{t+1} + \bar{b}_{n-1,g}' g_{t+1}\right)\right] \\
&= \exp\left(-\delta_0 - \delta_{1,h}' h_t - \delta_{1,g}' g_t + \bar{a}_{n-1}\right)\mathrm{E}_t^{\mathbb{Q}}\left[\exp\left(\bar{b}_{n-1,h}' h_{t+1} + \bar{b}_{n-1,g}' g_{t+1}\right)\right]
\end{aligned}$$

where the expectation is taken with respect to the distribution of the random vector $x_{t+1} = \left(h_{t+1}', g_{t+1}'\right)'$

$$\begin{aligned}
h_{t+1} &\overset{\mathbb{Q}}{\sim} \text{Mult-NCG}\left(\nu_{h,i}^{\mathbb{Q}}, \Phi_h^{\mathbb{Q}} h_t, \Sigma_h, \mu_h\right) \\
g_{t+1} &\overset{\mathbb{Q}}{\sim} \mathrm{N}\left(\mu_g^{\mathbb{Q}} + \Phi_g^{\mathbb{Q}} g_t + \Phi_{gh}^{\mathbb{Q}} h_t + \Sigma_{gh}\left[h_{t+1} - \left((I_H - \Phi_h^{\mathbb{Q}})\mu_h + \Sigma_h \nu_h^{\mathbb{Q}} + \Phi_h^{\mathbb{Q}} h_t\right)\right], \Sigma_{g,t} \Sigma_{g,t}'\right)
\end{aligned}$$

This expectation has the same form as the Laplace transform provided in Appendix A. We find

$$
\begin{aligned}
P_t^n &= \exp\Bigg(-\delta_0 - \delta_{1,g}' g_t - \delta_{1,h}' h_t + \bar{a}_{n-1} + \frac{1}{2} b_{n-1,g}' \Sigma_{g,t}\Sigma_{g,t}' \bar{b}_{n-1,g} \\
&\quad + \left[\mu_g^{\mathbb{Q}} + \Phi_g^{\mathbb{Q}} g_t + \Phi_{gh}^{\mathbb{Q}} h_t - \Sigma_{gh}\left((I_H - \Phi_h^{\mathbb{Q}})\mu_h + \Sigma_h \nu_h^{\mathbb{Q}} + \Phi_h^{\mathbb{Q}} h_t\right)\right]' \bar{b}_{n-1,g} \\
&\quad + \left[\Sigma_{gh}' \bar{b}_{n-1,g} + b_{n-1,h}\right]' \mu_h + \sum_{i=1}^{H} \frac{\mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}}{1 - \mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}} \mathrm{e}_i' \Sigma_h^{-1} \Phi_h^{\mathbb{Q}} \left[h_t - \mu_h\right] - \sum_{i=1}^{H} \nu_{h,i}^{\mathbb{Q}} \log\left(1 - \mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}\right)\Bigg) \\
&= \exp\Bigg(-\delta_0 + \bar{a}_{n-1} + \mu_g^{\mathbb{Q}\prime} \bar{b}_{n-1,g} + \mu_h' \left[b_{n-1,h} + \Phi_h^{\mathbb{Q}\prime} \Sigma_{gh}' \bar{b}_{n-1,g}\right] - \nu_h^{\mathbb{Q}\prime} \Sigma_h' \Sigma_{gh}' \bar{b}_{n-1,g} \\
&\quad + \frac{1}{2} b_{n-1,g}' \Sigma_{0,g} \Sigma_{0,g}' \bar{b}_{n-1,g} - \sum_{i=1}^{H} \nu_{h,i}^{\mathbb{Q}} \log\left(1 - \mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}\right) - \sum_{i=1}^{H} \frac{\mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}}{1 - \mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}} \mathrm{e}_i' \Sigma_h^{-1} \Phi_h^{\mathbb{Q}} \mu_h \\
&\quad + \left[\bar{b}_{n-1,g}' \Phi_g^{\mathbb{Q}} - \delta_{1,g}'\right] g_t \\
&\quad + \Bigg[\sum_{i=1}^{H} \frac{\mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}}{1 - \mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh}} \mathrm{e}_i' \Sigma_h^{-1} \Phi_h^{\mathbb{Q}} + \bar{b}_{n-1,g}' \left(\Phi_{gh}^{\mathbb{Q}} - \Sigma_{gh} \Phi_h^{\mathbb{Q}}\right) - \delta_{1,h}' \\
&\quad + \frac{1}{2} \left(I_H \otimes \bar{b}_{n-1,g}\right)' \Sigma_g \Sigma_g' \left(\iota_H \otimes \bar{b}_{n-1,g}\right)\Bigg] h_t\Bigg)
\end{aligned}
$$

where $\Sigma_g \Sigma_g'$ is a $GH \times GH$ matrix with diagonal elements $\Sigma_{i,g} \Sigma_{i,g}'$ for $i = 1, \ldots, H$ and $\bar{b}_{n-1,gh} = \Sigma_{gh}' \bar{b}_{n-1,g} + \bar{b}_{n-1,h}$. The Laplace transform exists only if $\mathrm{e}_i' \Sigma_h' \bar{b}_{n-1,gh} < 1$ for $i = 1, \ldots, H$.

Appendix C Factor rotations

Appendix C.1 Proof of Proposition 1

The admissibility restrictions needed to keep h_t positive are: (1.) $C_{hg} = 0$; (2.) C_{hh} is such that all elements $C_{hh}\Sigma_h$ are non-negative; (3.) c_h is such that $c_h + C_{hh}\mu_h$ is non-negative; (4.) C_{hh} and C_{gg} are full rank. For some values of θ, these restrictions may allow c_h and C_{hh} to be negative.

Under these restrictions, the process $\left(\tilde{h}_t', \tilde{g}_t'\right)'$ is a member of the same family of distributions as $(h_t', g_t')'$ only under a new parameters $\tilde{\theta}$. The proof is immediate by comparing the Laplace transform of these random variables before and after rotating them. The mapping between the parameters $\tilde{\theta}$ and θ is

$$
\tilde{\mu}_h = c_h + C_{hh}\mu_h \qquad \tilde{\Phi}_h = C_{hh}\Phi_h C_{hh}^{-1} \qquad \tilde{\Sigma}_h = C_{hh}\Sigma_h \qquad \tilde{\Phi}_g = C_{gg}\Phi_g C_{gg}^{-1}
$$

$$
\tilde{\mu}_g = c_g + C_{gg}\mu_g - C_{gg}\Phi_g C_{gg}^{-1} c_g + C_{gh}\left(\left[I_H - \Phi_h\right]\mu_h + \Sigma_h \nu_h\right) - \left(C_{gh}\Phi_h - C_{gg}\Phi_g C_{gg}^{-1} C_{gh} + C_{gg}\Phi_{gh}\right) C_{hh}^{-1} c_h
$$

$$
\tilde{\Phi}_{gh} = \left(C_{gh}\Phi_h - C_{gg}\Phi_g C_{gg}^{-1} C_{gh} + C_{gg}\Phi_{gh}\right) C_{hh}^{-1} \qquad \tilde{\Sigma}_{gh} = \left(C_{gh} + C_{gg}\Sigma_{gh}\right) C_{hh}^{-1}
$$

$$\tilde{\Sigma}_{0,g}\tilde{\Sigma}_{0,g}' = C_{gg}\Sigma_{0,g}\Sigma_{0,g}'C_{gg}' - \sum_{i=1}^{H} C_{gg}\Sigma_{i,g}\Sigma_{i,g}'C_{gg}'\mathrm{e}_i'C_{hh}^{-1}c_h \qquad \tilde{\Sigma}_{i,g}\tilde{\Sigma}_{i,g}' = \sum_{j=1}^{H} C_{gg}\Sigma_{j,g}\Sigma_{j,g}'C_{gg}'\mathrm{e}_j'C_{hh}^{-1}\mathrm{e}_i$$

■

Appendix D Stochastic discount factor

We define the stochastic discount factor as

$$M_{t+1} = \frac{\exp(-r_t)\, p(g_{t+1}|\mathcal{I}_t, h_{t+1}, z_{t+1}; \theta, \mathbb{Q})\, p(h_{t+1}|\mathcal{I}_t, z_{t+1}; \theta, \mathbb{Q})\, p(z_{t+1}|\mathcal{I}_t; \theta, \mathbb{Q})}{p(g_{t+1}|\mathcal{I}_t, h_{t+1}, z_{t+1}; \theta, \mathbb{P})\, p(h_{t+1}|\mathcal{I}_t, z_{t+1}; \theta, \mathbb{P})\, p(z_{t+1}|\mathcal{I}_t; \theta, \mathbb{P})}$$

where the distributions are Gaussian, gamma, and Poisson. The exact SDF is

$$M_{t+1} = \frac{\exp\left(-r_t - \frac{1}{2}\varepsilon_{g,t+1}^{\mathbb{Q}\prime}\varepsilon_{g,t+1}^{\mathbb{Q}} - \sum_{i=1}^{H} \mathrm{e}_i'\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}[h_t - \mu_h]\right)}{\exp\left(-\frac{1}{2}\varepsilon_{g,t+1}'\varepsilon_{g,t+1} - \sum_{i=1}^{H} \mathrm{e}_i'\Sigma_h^{-1}\Phi_h[h_t - \mu_h]\right)} \prod_{i=1}^{H} w_{i,t+1}^{\nu_{h,i}^{\mathbb{Q}} - \nu_{h,i}} \frac{\Gamma(\nu_{h,i} + z_{i,t+1})}{\Gamma\left(\nu_{h,i}^{\mathbb{Q}} + z_{i,t+1}\right)} \left(\frac{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}[h_t - \mu_h]}{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h[h_t - \mu_h]}\right)^{z_{i,t+1}}$$

where $\varepsilon_{g,t+1} = \Sigma_{g,t}^{-1}(g_{t+1} - \mu_g - \Phi_g g_t - \Phi_{gh}h_t - \Sigma_{gh}\varepsilon_{h,t+1})$ and $w_{t+1} = \Sigma_h^{-1}[h_{t+1} - \mu_h]$. No arbitrage requires that $M_{t+1} > 0$ which implies that $w_{t+1} > 0$. The Feller conditions $\nu_h > 1$ and $\nu_h^{\mathbb{Q}} > 1$ guarantee that $w_{t+1} > 0$ under both probability measures.

For intuition, break the log-SDF $m_{t+1} = \log(M_{t+1})$ into three terms; one for each of the shocks

$$m_{t+1} = -r_t + m_{g,t+1} + m_{h,t+1} + m_{z,t+1}$$

where $m_{i,t+1}$ is the compensation for risk i. Let $\lambda_g = \mu_g - \mu_g^{\mathbb{Q}}$, $\lambda_h = \nu_h - \nu_h^{\mathbb{Q}}$, $\Lambda_g = \Phi_g - \Phi_g^{\mathbb{Q}}$, $\Lambda_h = \Phi_h - \Phi_h^{\mathbb{Q}}$, and $\Lambda_{gh} = \Phi_{gh} - \Phi_{gh}^{\mathbb{Q}}$.

For the Gaussian part, we find $m_{g,t+1} = -\frac{1}{2}\lambda_{gt}'\lambda_{gt} - \lambda_{gt}'\epsilon_{g,t+1}$ where $\epsilon_{g,t+1} = \Sigma_{g,t}^{-1}\varepsilon_{g,t+1}$ is a standard, zero mean Gaussian shock. The price of Gaussian risk is

$$\lambda_{gt} = \Sigma_{g,t}^{-1}\{(\lambda_g + \Lambda_g g_t + \Lambda_{gh}h_t) - \Sigma_{gh}[\Sigma_h\lambda_h + \Lambda_h(h_t - \mu_h)]\}$$

This generalizes the expression for Gaussian ATSMs, with a time-varying quantity of risk $\Sigma_{g,t}$.

Recall that $w_{t+1} = \Sigma_h^{-1}(h_{t+1} - \mu_h)$. We will write risk compensation in terms of w_{t+1}.

$$\begin{aligned}
m_{h,t+1} &= \sum_{i=1}^{H} -\log\Gamma\left(\nu_{h,i}^{\mathbb{Q}} + z_{i,t+1}\right) + \left(\nu_{h,i}^{\mathbb{Q}} + z_{i,t+1} - 1\right)\log\left(\mathrm{e}_i'\Sigma_h^{-1}(h_{t+1} - \mu_h)\right) - \mathrm{e}_i'\Sigma_h^{-1}(h_{t+1} - \mu_h) \\
& \sum_{i=1}^{H} \log\Gamma(\nu_{h,i} + z_{i,t+1}) - (\nu_{h,i} + z_{i,t+1} - 1)\log\left(\mathrm{e}_i'\Sigma_h^{-1}(h_{t+1} - \mu_h)\right) + \mathrm{e}_i'\Sigma_h^{-1}(h_{t+1} - \mu_h) \\
&= \sum_{i=1}^{H} \log\left(\frac{\Gamma(\nu_{h,i} + z_{i,t+1})}{\Gamma\left(\nu_{h,i}^{\mathbb{Q}} + z_{i,t+1}\right)}\right) - \left(\nu_{h,i} - \nu_{h,i}^{\mathbb{Q}}\right)\log\left(\mathrm{e}_i'\Sigma_h^{-1}(h_{t+1} - \mu_h)\right) \\
&\approx \sum_{i=1}^{H} \log\left([\nu_{h,i} + z_{i,t+1}]^{\nu_{h,i} - \nu_{h,i}^{\mathbb{Q}}}\right) - \lambda_{h,i}\log(w_{i,t+1}) \;=\; \sum_{i=1}^{H} -\lambda_{h,i}\left[\log(w_{i,t+1}) - \log(\nu_{h,i} + z_{i,t+1})\right] \\
&= \sum_{i=1}^{H} -\lambda_{h,i}\left[\log\left(1 + \frac{w_{i,t+1} - \nu_{h,i} - z_{i,t+1}}{\nu_{h,i} + z_{i,t+1}}\right)\right] \;\approx\; \sum_{i=1}^{H} -\frac{\lambda_{h,i}}{\sqrt{\nu_{h,i} + z_{i,t+1}}}\frac{w_{i,t+1} - \nu_{h,i} - z_{i,t+1}}{\sqrt{\nu_{h,i} + z_{i,t+1}}}
\end{aligned}$$

The compensation for gamma risks is approximately $m_{h,t+1} \approx -\lambda_{wt}'\epsilon_{w,t+1}$ where $\epsilon_{w,t+1,i} = \frac{w_{i,t+1} - \nu_{h,i} - z_{i,t+1}}{\sqrt{\nu_i + z_{i,t+1}}}$ is a gamma r.v. standardized to have mean 0 and variance 1.[22] The market price of risk is $\lambda_{wt,i} = \frac{\lambda_{h,i}}{\sqrt{\nu_{h,i} + z_{i,t+1}}}$. We note that $\mathbb{V}(w_{it}|z_t) = \nu_{h,i} + z_{i,t}$.

Consider the non-Gaussian part due to the Poisson distribution

$$\begin{aligned}
m_{z,t+1} &= \sum_{i=1}^{H} z_{i,t+1}\log\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}(h_t - \mu_h)\right) - \log(z_{i,t+1}!) - \mathrm{e}_i'\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}(h_t - \mu_h) \\
& -z_{i,t+1}\log\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h(h_t - \mu_h)\right) + \log(z_{i,t+1}!) + \mathrm{e}_i'\Sigma_h^{-1}\Phi_h(h_t - \mu_h) \\
&= \sum_{i=1}^{H} z_{i,t+1}\log\left(1 - \frac{\mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t}{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}\right) + \mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t \\
&\approx \sum_{i=1}^{H} -z_{i,t+1}\frac{\mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t}{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t} + \mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t \\
&= \sum_{i=1}^{H} -\mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t\left(\frac{z_{i,t+1} - \mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}\right) \;=\; \sum_{i=1}^{H} -\frac{\mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t}{\sqrt{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}}\epsilon_{z,t+1,i}
\end{aligned}$$

The log SDF is $m_{z,t+1} \approx -\lambda_{zt}'\epsilon_{z,t+1}$ where $\epsilon_{z,t+1} = \frac{z_{i,t+1} - \mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}{\sqrt{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}}$ is Poisson r.v. standardized to have mean 0 and variance 1 and $\lambda_{zt,i} = \frac{\mathrm{e}_i'\Sigma_h^{-1}\Lambda_h\Sigma_h w_t}{\sqrt{\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\Sigma_h w_t}}$.

[22]Our derivation uses the approximation that $\frac{\Gamma(a+x)}{\Gamma(b+x)} \propto x^{a-b}\left(1 + O\left(\frac{1}{x}\right)\right)$ for large x. We also use the fact that $\log(1+x) = x$ for small x.

Appendix E Log-likelihood for spanned models

Dropping the initial condition, the conditional log-likelihood for the general affine model is given by

$$
\begin{aligned}
\ell(\theta) = {} & \text{CONST} - (T-1)\log|B_1| - \frac{T-1}{2}\log|\Omega| - \frac{1}{2}\sum_{t=2}^{T}\text{tr}\left(\Omega^{-1}\eta_t\eta_t'\right) - \frac{1}{2}\sum_{t=2}^{T}\log\left|\Sigma_{g,t-1}\Sigma_{g,t-1}'\right| \\
& -\frac{1}{2}\sum_{t=2}^{T}\text{tr}\left(\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\varepsilon_{gt}\varepsilon_{gt}'\right) \\
& -(T-1)\log|\Sigma_h| - \sum_{t=2}^{T}\sum_{i=1}^{H}\text{e}_i'\Sigma_h^{-1}\left(h_t-\mu_h\right) - \sum_{t=2}^{T}\sum_{i=1}^{H}\text{e}_i'\Sigma_h^{-1}\Phi_h\left(h_{t-1}-\mu_h\right) \\
& +\sum_{t=2}^{T}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2}\log\left(\text{e}_i'\Sigma_h^{-1}\left[h_t-\mu_h\right]\right) - \sum_{t=2}^{T}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2}\log\left(\text{e}_i'\Sigma_h^{-1}\Phi_h\left[h_{t-1}-\mu_h\right]\right) \\
& +\sum_{t=2}^{T}\sum_{i=1}^{H}\log I_{\nu_{h,i}-1}\left(2\sqrt{\left(\text{e}_i'\Sigma_h^{-1}\left[h_t-\mu_h\right]\right)\left(\text{e}_i'\Sigma_h^{-1}\Phi_h\left[h_{t-1}-\mu_h\right]\right)}\right)
\end{aligned}
$$

where $I_\lambda(x)$ is the modified Bessel function of the first kind and e_i is the $H\times 1$ unit vector.

Appendix F Proofs and Analytical Derivatives

In this appendix, we prove the propositions in the paper. We begin with a preliminary lemma.

Lemma 1 *The derivative of the concentrated log-likelihood function can be computed as the partial derivative of the log-likelihood function:* $\frac{d\ell\left(\hat{\theta}_c(\theta_m),\theta_m\right)}{d\theta_m'} = \frac{\partial\ell\left(\hat{\theta}_c,\theta_m\right)}{\partial\theta_m'}$, *where* $\hat{\theta}_c(\theta_m) = \arg\max_{\theta_c}\ell(\theta_c,\theta_m)$

Proof: $\frac{d\ell\left(\hat{\theta}_c(\theta_m),\theta_m\right)}{d\theta_m'} = \frac{\partial\ell\left(\hat{\theta}_c,\theta_m\right)}{\partial\theta_m'} + \frac{\partial\ell(\theta_c,\theta_m)}{\partial\theta_c'}|_{\theta_c=\hat{\theta}_c}\frac{d\hat{\theta}_c(\theta_m)}{d\theta_m'} = \frac{\partial\ell\left(\hat{\theta}_c,\theta_m\right)}{\partial\theta_m'}$, where $\frac{\partial\ell(\theta_c,\theta_m)}{\partial\theta_c'}|_{\theta_c=\hat{\theta}_c} = 0$ by the definition of $\hat{\theta}_c$. ■

Appendix F.1 Proof of Proposition 2

Proof First, we show that optimizing the original likelihood and optimizing the concentrated likelihood lead to equivalent solutions. Solving $\max_\theta \ell(\theta)$ requires $\frac{\partial\ell(\theta_c,\theta_m)}{\partial\theta_m'} = 0$ and $\frac{\partial\ell(\theta_m,\theta_c)}{\partial\theta_c'} = 0$. The solution to $\max_{\theta_m}\ell\left(\hat{\theta}_c(\theta_m),\theta_m\right)$ has the property $\frac{d\ell\left(\hat{\theta}_c(\theta_m),\theta_m\right)}{d\theta_m'} = 0$. By Lemma 1, $\frac{\partial\ell\left(\hat{\theta}_c,\theta_m\right)}{\partial\theta_m'} = 0$. This with $\frac{\partial\ell(\theta_c,\theta_m)}{\partial\theta_c'}|_{\theta_c=\hat{\theta}_c} = 0$ by the definition of $\hat{\theta}_c$ constitute the two F.O.C.'s for the original problem.

Second, showing $\hat{\theta}_c$ can be solved by least squares given θ_m is straightforward. A value for θ_m maps into a value for the loadings A and B though (9)-(10), and therefore x_t by (16). Given the factors, equation

(11) is a linear regression with heteroskedasticity, hence GLS. Given yields, factors and bond loadings, Ω in equation (17) is the variance-covariance matrix in a linear regression with homoskedasticity, hence OLS. ■

Appendix F.2 Proof of Proposition 3

We use the result in Lemma 1 to derive the gradients for the concentrated log likelihood based on the original log-likelihood, which makes the derivation easier.

Proof A direct result of Lemma 1 is $\frac{d\ell\left(\hat{\theta}_c(\theta_m),\theta_m,A(\theta_m),B(\theta_m)\right)}{d\theta_m'} = \frac{\partial\ell\left(\hat{\theta}_c,\theta_m,A(\theta_m),B(\theta_m)\right)}{\partial\theta_m'}$. Applying the chain rule,

$$\begin{aligned}\frac{\partial\ell\left(\hat{\theta}_c,\theta_m,A\left(\theta_m\right),B\left(\theta_m\right)\right)}{\partial\theta_m'} &= \frac{\partial\ell\left(\hat{\theta}_c,\theta_m,A,B\right)}{\partial\theta_m'}+\frac{\partial\ell\left(\hat{\theta}_c,\theta_m,A,B\right)}{\partial A'}\frac{\partial A\left(\theta_m\right)}{\partial\theta_m'}\\ &+\frac{\partial\ell\left(\hat{\theta}_c,\theta_m,A,B\right)}{\partial\mathrm{vec}\left(B'\right)'}\frac{\partial\mathrm{vec}\left(B\left(\theta_m\right)'\right)}{\partial\theta_m'}.\end{aligned}$$

■

Appendix F.3 Gradients

We now provide the analytical gradient. Note that $\hat{\theta}_c=(\hat{\mu}_g,\hat{\Phi}_g,\hat{\Phi}_{gh},\hat{\Omega})$ are optimized by $\hat{\theta}_c=\underset{\theta_c}{\mathrm{argmax}}\,\ell\left(\theta_c,\theta_m\right)$ detailed in Proposition 2. Let $\tilde{h}_{it}=2\sqrt{\left(\mathrm{e}_i'\Sigma_h^{-1}\left[h_t-\mu_h\right]\right)\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\left[h_{t-1}-\mu_h\right]\right)}$, $\hat{h}_t=\Sigma_h^{-1}\left(h_t-\mu_h\right)$, and $\hat{\hat{h}}_{t-1}=\Sigma_h^{-1}\Phi_h\left(h_{t-1}-\mu_h\right)$. The matrices S_g and S_h extract the Gaussian $g_t = S_g x_t$ and non-Gaussian factors $h_t = S_h x_t$ from x_t.

$$\begin{aligned}\frac{\partial\ell}{\partial\nu_{h,i}} &= -\sum_{t=2}^{T}\varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\Sigma_{gh}\Sigma_h\mathrm{e}_i+\frac{1}{2}\sum_{t=2}^{T}\log\left(\frac{\mathrm{e}_i'\hat{h}_t}{\mathrm{e}_i'\hat{\hat{h}}_{t-1}}\right)+\sum_{t=2}^{T}\frac{1}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\frac{\partial I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}{\partial\nu_{h,i}}\\ \frac{\partial\ell}{\partial\mathrm{vec}\left(\Phi_h\right)'} &= -\sum_{t=2}^{T}\mathrm{vec}\left(\Sigma_{gh}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\varepsilon_{gt}h_{t-1}'\right)'\\ &-\sum_{t=2}^{T}\sum_{i=1}^{H}\mathrm{vec}\left(\left(\Sigma_h'\right)^{-1}\mathrm{e}_i\left(h_{t-1}-\mu_h\right)'\right)'-\sum_{t=2}^{T}\sum_{i=1}^{H}\frac{\left(\nu_{h,i}-1\right)}{2\mathrm{e}_i'\hat{\hat{h}}_{t-1}}\mathrm{vec}\left(\left(\Sigma_h'\right)^{-1}\mathrm{e}_i\left(h_{t-1}-\mu_h\right)'\right)'\\ &+\sum_{t=2}^{T}\sum_{i=1}^{H}\left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}}+\frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right]\frac{2\mathrm{e}_i'\hat{h}_t}{\tilde{h}_{it}}\mathrm{vec}\left(\left(\Sigma_h'\right)^{-1}\mathrm{e}_i\left(h_{t-1}-\mu_h\right)'\right)'\end{aligned}$$

where we have used $\frac{\partial I_\lambda(x)}{\partial x} = \frac{\lambda}{x}I_\lambda(x)+I_{\lambda+1}(x)$, see, Abramowitz and Stegun(1964). The derivative $\frac{\partial I_\lambda(x)}{\partial\lambda}$ is a complicated expression that is easier to compute numerically. The derivatives for parameters entering the loadings are calculated by the chain rule. Take derivatives of ℓ w.r.t. the loadings A_1, A_2, B_1 and B_2. Then, take derivatives of the loadings w.r.t. the parameters.

$$\frac{\partial \ell}{\partial \delta_0} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \delta_0} \qquad \frac{\partial \ell}{\partial \delta_{1,g}'} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \delta_{1,g}'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \delta_{1,g}'} \qquad \frac{\partial \ell}{\partial \delta_{1,h}'} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \delta_{1,h}'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \delta_{1,h}'}$$

$$\frac{\partial \ell}{\partial \nu_h^{\mathbb{Q}\prime}} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \nu_h^{\mathbb{Q}\prime}} \qquad \frac{\partial \ell}{\partial \mu_g^{\mathbb{Q}\prime}} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \mu_g^{\mathbb{Q}\prime}} \qquad \frac{\partial \ell}{\partial \text{vec}\left(\Phi_g^{\mathbb{Q}}\right)'} = \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vec}\left(\Phi_g^{\mathbb{Q}}\right)'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vec}\left(\Phi_g^{\mathbb{Q}}\right)'}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \text{vec}\left(\Phi_{gh}^{\mathbb{Q}}\right)'} &= \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vec}\left(\Phi_{gh}^{\mathbb{Q}}\right)'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vec}\left(\Phi_{gh}^{\mathbb{Q}}\right)'} \\
\frac{\partial \ell}{\partial \text{vec}\left(\Phi_{h}^{\mathbb{Q}}\right)'} &= \frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vec}\left(\Phi_{h}^{\mathbb{Q}}\right)'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vec}\left(\Phi_{h}^{\mathbb{Q}}\right)'}
\end{aligned}$$

The derivatives of the remaining parameters that enter both the loadings and the $\mathbb{P}$ dynamics are

$$\frac{d\ell}{d\mu_h'} = \frac{\partial \ell}{\partial \mu_h'}+\frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \mu_h'} \qquad \frac{d\ell}{d\text{vech}\left(\Sigma_{0,g}\right)'} = \frac{\partial \ell}{\partial \text{vech}\left(\Sigma_{0,g}\right)'}+\frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vech}\left(\Sigma_{0,g}\right)'}$$

$$\begin{aligned}
\frac{d\ell}{d\text{vec}\left(\Sigma_h\right)'} &= \frac{\partial \ell}{\partial \text{vec}\left(\Sigma_h\right)'}+\frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vec}\left(\Sigma_h\right)'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vec}\left(\Sigma_h\right)'} \\
\frac{d\ell}{d\text{vec}\left(\Sigma_{gh}\right)'} &= \frac{\partial \ell}{\partial \text{vec}\left(\Sigma_{gh}\right)'}+\frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vec}\left(\Sigma_{gh}\right)'}+\frac{\partial \ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vec}\left(\Sigma_{gh}\right)'} \\
\frac{d\ell}{d\text{vech}\left(\Sigma_{i,g}\right)'} &= \frac{\partial \ell}{\partial \text{vech}\left(\Sigma_{i,g}\right)'}+\frac{\partial \ell}{\partial A'}\frac{\partial A}{\partial \text{vech}\left(\Sigma_{i,g}\right)'}+\frac{d\ell}{\partial \text{vec}\left(B'\right)'}\frac{\partial \text{vec}\left(B'\right)}{\partial \text{vech}\left(\Sigma_{i,g}\right)'}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \text{vec}\left(\Sigma_h\right)'} = & -\sum_{t=2}^{T}\text{vec}\left(\Sigma_{gh}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\varepsilon_{gt}\nu_h'\right)' - (T-1)\text{vec}\left(\left(\left(\Sigma_h^{-1}\right)'\right)\right)' \\
& +\sum_{t=2}^{T}\sum_{i=1}^{H}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{h}_t'\right)' + \sum_{t=2}^{T}\sum_{i=1}^{H}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{\hat{h}}_{t-1}'\right)' \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H}\frac{\left(\nu_{h,i}-1\right)}{2\mathrm{e}_i'\hat{h}_t}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{h}_t'\right)' + \sum_{t=2}^{T}\sum_{i=1}^{H}\frac{\left(\nu_{h,i}-1\right)}{2\mathrm{e}_i'\hat{\hat{h}}_{t-1}}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{\hat{h}}_{t-1}'\right)' \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H}\left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}}+\frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right]\frac{2\mathrm{e}_i'\hat{h}_t}{\tilde{h}_{it}}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{\hat{h}}_{t-1}'\right)' \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H}\left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}}+\frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right]\frac{2\mathrm{e}_i'\hat{\hat{h}}_{t-1}}{\tilde{h}_{it}}\text{vec}\left(\left(\Sigma_h^{-1}\right)'\mathrm{e}_i\hat{h}_t'\right)'
\end{aligned}$$

$$\frac{\partial \ell}{\partial \text{vec}\left(\Sigma_{gh}\right)'} = \sum_{t=2}^{T} \text{vec}\left(\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \varepsilon_{gt}\varepsilon_{ht}'\right)'$$

$$\frac{\partial \ell}{\partial \text{vech}\left(\Sigma_{0,g}\right)'} = \sum_{t=2}^{T} \text{vec}\left(\left[\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \varepsilon_{gt}\varepsilon_{gt}' - I_G\right] \left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \Sigma_{0,g}\right)' \mathcal{D}_G^L$$

$$\frac{\partial \ell}{\partial \text{vech}\left(\Sigma_{i,g}\right)'} = \sum_{t=2}^{T} \text{vec}\left(\left[\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \varepsilon_{gt}\varepsilon_{gt}' - I_G\right] \left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \Sigma_{i,g}h_{i,t-1}\right)' \mathcal{D}_G^L$$

$$\begin{aligned}
\frac{\partial \ell}{\partial \mu_h'} = & -\sum_{t=2}^{T} \varepsilon_{gt}' \left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1} \Sigma_{gh}\left(I_H - \Phi_h\right) + \sum_{t=2}^{T}\sum_{i=1}^{H} \text{e}_i'\Sigma_h^{-1} + \sum_{t=2}^{T}\sum_{i=1}^{H} \text{e}_i'\Sigma_h^{-1}\Phi_h \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H} \frac{\left(\nu_{h,i}-1\right)}{2\text{e}_i'\hat{h}_t}\text{e}_i'\Sigma_h^{-1} + \sum_{t=2}^{T}\sum_{i=1}^{H} \frac{\left(\nu_{h,i}-1\right)}{2\text{e}_i'\hat{\hat{h}}_{t-1}}\text{e}_i'\Sigma_h^{-1}\Phi_h \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H} \left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}} + \frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right] \frac{2}{\tilde{h}_{it}} \left[\text{e}_i'\hat{h}_t\text{e}_i'\Sigma_h^{-1}\Phi_h + \text{e}_i'\hat{\hat{h}}_{t-1}\text{e}_i'\Sigma_h^{-1}\right]
\end{aligned}$$

$$\frac{\partial \ell}{\partial A_2'} = \sum_{t=2}^{T} \eta_t'\Omega^{-1} \qquad \frac{\partial \ell}{\partial \text{vec}\left(B_2'\right)'} = \sum_{t=2}^{T} \text{vec}\left(x_t\eta_t'\Omega^{-1}\right)'$$

$$\begin{aligned}
\frac{\partial \ell}{\partial A_1'} = & \sum_{t=2}^{T} \left[-\eta_t'\Omega^{-1}B_2B_1^{-1} + \varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\left[\left(I_G - \Phi_g\right)S_g - \left(\Phi_{gh} + \Sigma_{gh} - \Sigma_{gh}\Phi_h\right)S_h\right]B_1^{-1}\right. \\
& +\frac{1}{2}\text{vec}\left(\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\left[I_G - \varepsilon_{gt}\varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\right]\right)' \sum_{i=1}^{H} \text{vec}\left(\Sigma_{i,g}\Sigma_{i,g}'\right) S_{h_i}B_1^{-1} \\
& +\sum_{t=2}^{T}\sum_{i=1}^{H} \text{e}_i'\Sigma_h^{-1}S_hB_1^{-1} - \sum_{t=2}^{T}\sum_{i=1}^{H} \frac{\left(\nu_{h,i}-1\right)}{2\text{e}_i'\hat{h}_t}\text{e}_i'\Sigma_h^{-1}S_hB_1^{-1} \\
& +\sum_{t=2}^{T}\sum_{i=1}^{H} \text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1} + \sum_{t=2}^{T}\sum_{i=1}^{H} \frac{\left(\nu_{h,i}-1\right)}{2\text{e}_i'\hat{\hat{h}}_{t-1}}\text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1} \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H} \left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}} + \frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right] \frac{2\text{e}_i'\hat{\hat{h}}_{t-1}}{\tilde{h}_{it}}\text{e}_i'\Sigma_h^{-1}S_hB_1^{-1} \\
& -\sum_{t=2}^{T}\sum_{i=1}^{H} \left[\frac{\left(\nu_{h,i}-1\right)}{\tilde{h}_{it}} + \frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right] \frac{2\text{e}_i'\hat{h}_t}{\tilde{h}_{it}}\text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1}
\end{aligned}$$

$$\frac{\partial \ell}{\partial \text{vec}\left(B_1'\right)'} = -(T-1)\text{vec}\left(B_1^{-1}\right)' + \sum_{t=2}^{T}\left[-\text{vec}\left(x_t\eta_t'\Omega^{-1}B_2B_1^{-1}\right)'\right.$$

$$+\text{vec}\left(x_t\varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\left(S_g - \Sigma_{gh}S_h\right)B_1^{-1}\right)'$$

$$-\text{vec}\left(x_{t-1}\varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\left(\Phi_g S_g + \left[\Phi_{gh} - \Sigma_{gh}\Phi_h\right]S_h\right)B_1^{-1}\right)'$$

$$+\frac{1}{2}\text{vec}\left(\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\left[I_G - \varepsilon_{gt}\varepsilon_{gt}'\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\right]\right)'\sum_{i=1}^{H}\text{vec}\left(\Sigma_{i,g}\Sigma_{i,g}'\right)\left[S_{h_i}B_1^{-1}\otimes x_{t-1}'\right]$$

$$+\sum_{t=2}^{T}\sum_{i=1}^{H}\text{vec}\left(x_t \text{e}_i'\Sigma_h^{-1}S_hB_1^{-1}\right)' - \sum_{t=2}^{T}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2\text{e}_i'\hat{h}_t}\text{vec}\left(x_t\text{e}_i'\Sigma_h^{-1}S_hB_1^{-1}\right)'$$

$$+\sum_{t=2}^{T}\sum_{i=1}^{H}\text{vec}\left(x_{t-1}\text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1}\right)' + \sum_{t=2}^{T}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2\text{e}_i'\hat{\hat{h}}_{t-1}}\text{vec}\left(x_{t-1}\text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1}\right)'$$

$$-\sum_{t=2}^{T}\sum_{i=1}^{H}\left[\frac{(\nu_{h,i}-1)}{\tilde{h}_{it}} + \frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right]\frac{2\text{e}_i'\hat{\hat{h}}_{t-1}}{\tilde{h}_{it}}\text{vec}\left(x_t\text{e}_i'\Sigma_h^{-1}S_hB_1^{-1}\right)'$$

$$-\sum_{t=2}^{T}\sum_{i=1}^{H}\left[\frac{(\nu_{h,i}-1)}{\tilde{h}_{it}} + \frac{I_{\nu_{h,i}}\left(\tilde{h}_{it}\right)}{I_{\nu_{h,i}-1}\left(\tilde{h}_{it}\right)}\right]\frac{2\text{e}_i'\hat{h}_t}{\tilde{h}_{it}}\text{vec}\left(x_{t-1}\text{e}_i'\Sigma_h^{-1}\Phi_hS_hB_1^{-1}\right)'$$

The derivatives of A and B w.r.t. the parameters can be computed recursively as a function of maturity along with $\bar{a}_n, \bar{b}_{n,g}$ and $\bar{b}_{n,h}$. The derivatives of B_g and B_h have separate recursions. Let $\bar{b}_{n,g,\psi}$ denote the derivatives of the Gaussian loadings $\bar{b}_{n,g}$ at maturity n w.r.t. a parameter ψ. All recursions are written assuming that ψ is a full vector/matrix of parameters with no restrictions. If the matrix has fewer parameters than entries, the user will have to multiply the respective recursion by a selection matrix. Let $\bar{d}_{n-1} = \text{diag}\left(\iota_H - \Sigma_h'\bar{b}_{n-1,gh}\right)$ be a diagonal $H \times H$ matrix. Let $\bar{c}_{n-1}' = \left(\nu^{\mathbb{Q}\prime}\bar{d}_{n-1}^{-1} - \mu_h^{\mathbb{Q}\prime}\Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\right)\Sigma_h'$ be an $1 \times H$ vector.

$$\underset{1\times G}{\bar{a}_{n,\mu_g^{\mathbb{Q}}}'} = \bar{a}_{n-1,\mu_g^{\mathbb{Q}}}' + \bar{b}_{n-1,g}'$$

$$\underset{1\times H}{\bar{a}_{n,\mu_h}'} = \bar{a}_{n-1,\mu_h}' + \bar{b}_{n-1,h}' + \bar{b}_{n-1,g}'\Sigma_{gh}\Phi_h^{\mathbb{Q}} - \bar{b}_{n-1,gh}'\Sigma_h\bar{d}_{n-1}^{-1}\Sigma_h^{-1}\Phi_h^{\mathbb{Q}}$$

$$\underset{1\times H}{\bar{a}_{n,\nu_h^{\mathbb{Q}}}'} = \bar{a}_{n-1,\nu_h^{\mathbb{Q}}}' - \log\left(\iota_H' - \bar{b}_{n-1,gh}'\Sigma_h\right) - \bar{b}_{n-1,g}'\Sigma_{gh}\Sigma_h$$

$$\underset{1\times G(G+1)/2}{\bar{a}_{n,\Sigma_{0,g}}'} = \bar{a}_{n-1,\Sigma_{0,g}}' + \text{vec}\left(\bar{b}_{n-1,g}\bar{b}_{n-1,g}'\Sigma_{0,g}\right)'\mathcal{D}_G^L$$

$$\underset{1\times H}{\bar{a}_{n,\delta_{1h}}'} = \bar{a}_{n-1,\delta_{1h}}' + \mu_h'\bar{b}_{n-1,h,\delta_{1h}} + \bar{c}_{n-1}'\bar{b}_{n-1,h,\delta_{1h}}$$

$$
\begin{aligned}
\underset{1\times GH}{\bar{a}'_{n,\Phi^{\mathbb{Q}}_{gh}}} &= \bar{a}'_{n-1,\Phi^{\mathbb{Q}}_{gh}} + \mu'_h \bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{gh}} + \bar{c}'_{n-1}\bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{gh}} \\
\underset{1\times G(G+1)/2}{\bar{a}'_{n,\Sigma_{i,g}}} &= \bar{a}'_{n-1,\Sigma_{i,g}} + \mu'_h \bar{b}_{n-1,h,\Sigma_{i,g}} + \bar{c}'_{n-1}\bar{b}_{n-1,h,\Sigma_{i,g}} \qquad i=1,\ldots,H \\
\underset{1\times H^2}{\bar{a}'_{n,\Phi^{\mathbb{Q}}_{h}}} &= \bar{a}'_{n-1,\Phi^{\mathbb{Q}}_{h}} + \mu'_h \bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{h}} + \bar{c}'_{n-1}\bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{h}} \\
&\quad + \mathrm{vec}\left(\left(\Sigma'_{gh}\bar{b}_{n-1,g} - \Sigma_h^{-1\prime}\bar{d}_{n-1}^{-1}\Sigma'_h\bar{b}_{n-1,gh}\right)\mu'_h\right)' \\
\underset{1\times GH}{\bar{a}'_{n,\Sigma_{gh}}} &= \bar{a}'_{n-1,\Sigma_{gh}} + \mu'_h \bar{b}_{n-1,h,\Sigma_{gh}} + \bar{c}'_{n-1}\bar{b}_{n-1,h,\Sigma_{gh}} + \mathrm{vec}\left[\bar{b}_{n-1,g}\left(\mu'_h\Phi_h^{\mathbb{Q}\prime} - \nu_h^{\mathbb{Q}\prime}\Sigma'_h + \bar{c}'_{n-1}\right)\right]' \\
\underset{1\times G}{\bar{a}'_{n,\delta_{1g}}} &= \bar{a}'_{n-1,\delta_{1g}} + \mu'_h \bar{b}_{n-1,h,\delta_{1g}} + \bar{c}'_{n-1}\bar{b}_{n-1,gh,\delta_{1g}} \\
&\quad + \left(\mu_g^{\mathbb{Q}\prime} + \mu'_h\Phi_h^{\mathbb{Q}\prime}\Sigma'_{gh} - \nu_h^{\mathbb{Q}\prime}\Sigma'_h\Sigma'_{gh}\right)\bar{b}_{n-1,g,\delta_{1g}} + \bar{b}'_{n-1,g}\Sigma_{0,g}\Sigma'_{0,g}\bar{b}_{n-1,g,\delta_{1g}} \\
\underset{1\times G^2}{\bar{a}'_{n,\Phi^{\mathbb{Q}}_{g}}} &= \bar{a}'_{n-1,\Phi^{\mathbb{Q}}_{g}} + \mu'_h \bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{g}} + \bar{c}'_{n-1}d\bar{b}_{n-1,gh,\Phi^{\mathbb{Q}}_{g}} \\
&\quad + \left(\mu_g^{\mathbb{Q}\prime} + \mu'_h\Phi_h^{\mathbb{Q}\prime}\Sigma'_{gh} - \nu_h^{\mathbb{Q}\prime}\Sigma'_h\Sigma'_{gh}\right)\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_{g}} + \bar{b}'_{n-1,g}\Sigma_{0,g}\Sigma'_{0,g}\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_{g}} \\
\underset{1\times H(H+1)/2}{\bar{a}'_{n,\Sigma_{h}}} &= \bar{a}'_{n-1,\Sigma_{h}} + \mu'_h \bar{b}_{n-1,h,\Sigma_{h}} + \bar{c}'_{n-1}\bar{b}_{n-1,h,\Sigma_{h}} + \mathrm{vec}\left(\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-1}\Sigma'_h\bar{b}_{n-1,gh}\mu'_h\Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\right)' \\
&\quad - \mathrm{vec}\left(\Sigma'_{gh}\bar{b}_{n-1,g}\nu'_h\right)' + \mathrm{vec}\left(\bar{b}_{n-1,gh}\bar{c}'_{n-1}\Sigma_h^{-1\prime}\right)'
\end{aligned}
$$

and $\frac{\partial \mathrm{vec}(A)}{\partial \delta_0} = \iota_N$ and with $\bar{b}_{1,g,\delta_{1g}} = -I_G, \bar{b}_{1,h,\delta_{1h}} = -I_H$. All other initial conditions start at zero.

$$
\begin{aligned}
\underset{G\times G}{\bar{b}_{n,g,\delta_{1g}}} &= \Phi_g^{\mathbb{Q}\prime}\bar{b}_{n-1,g,\delta_{1g}} - I_G \\
\underset{G\times G^2}{\bar{b}_{n,g,\Phi^{\mathbb{Q}}_{g}}} &= \Phi_g^{\mathbb{Q}\prime}\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_{g}} + \left(I_G\otimes\bar{b}'_{n-1,g}\right) \\
\underset{H\times H}{\bar{b}_{n,h,\delta_{1h}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\delta_{1h}} - I_H \\
\underset{H\times GH}{\bar{b}_{n,h,\Phi^{\mathbb{Q}}_{gh}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{gh}} + I_H\otimes\bar{b}'_{n-1,g} \\
\underset{H\times G(G+1)/2}{\bar{b}_{n,h,\Sigma_{i,g}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\Sigma_{gi}} + \mathrm{e}_i\mathrm{vec}\left(\bar{b}_{n-1,g}\bar{b}'_{n-1,g}\Sigma_{i,g}\right)'\mathcal{D}_G^L \qquad i=1,\ldots,H \\
\underset{H\times H^2}{\bar{b}_{n,h,\Phi^{\mathbb{Q}}_{h}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_{h}} - I_H\otimes\bar{b}'_{n-1,g}\Sigma_{gh} + I_H\otimes\bar{b}'_{n-1,gh}\Sigma_h\bar{d}_{n-1}^{-1}\Sigma_h^{-1} \\
\underset{H\times GH}{\bar{b}_{n,h,\Sigma_{gh}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\Sigma_{gh}} - \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\left(I_H - \bar{d}_{n-1}^{-2}\right)\Sigma'_h\otimes\bar{b}'_{n-1,g} \\
\underset{H\times G}{\bar{b}_{n,h,\delta_{1g}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,gh,\delta_{1g}} + \left(\Phi_{gh}^{\mathbb{Q}} - \Sigma_{gh}\Phi_h^{\mathbb{Q}}\right)'\bar{b}_{n-1,g,\delta_{1g}} \\
&\quad + \left(I_H\otimes\bar{b}'_{n-1,g}\right)\Sigma_g\Sigma'_g\left(\iota_H\otimes\bar{b}_{n-1,g,\delta_{1g}}\right) \\
\underset{H\times G^2}{\bar{b}_{n,h,\Phi^{\mathbb{Q}}_{g}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,gh,\Phi^{\mathbb{Q}}_{g}} + \left(\Phi_{gh}^{\mathbb{Q}} - \Sigma_{gh}\Phi_h^{\mathbb{Q}}\right)'\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_{g}} \\
&\quad + \left(I_H\otimes\bar{b}'_{n-1,g}\right)\Sigma_g\Sigma'_g\left(\iota_H\otimes\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_{g}}\right) \\
\underset{H\times H(H+1)/2}{\bar{b}_{n,h,\Sigma_{h}}} &= \Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\Sigma'_h\bar{b}_{n-1,h,\Sigma_{h}} - \left(\Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\otimes\bar{b}'_{n-1,gh}\Sigma_h\bar{d}_{n-1}^{-1}\Sigma_h^{-1}\right) + \left(\Phi_h^{\mathbb{Q}\prime}\Sigma_h^{-1\prime}\bar{d}_{n-1}^{-2}\otimes\bar{b}'_{n-1,gh}\right)
\end{aligned}
$$

where we also need to account for the derivatives of $\bar{b}_{n,gh} = \Sigma'_{gh}\bar{b}_{n,g} + \bar{b}_{n,h}$ as

$$\underset{H\times G}{\bar{b}_{n-1,gh,\delta_{1g}}} = \Sigma'_{gh}\bar{b}_{n-1,g,\delta_{1g}} + \bar{b}_{n-1,h,\delta_{1g}} \qquad \underset{H\times G^2}{\bar{b}_{n-1,gh,\Phi^{\mathbb{Q}}_g}} = \Sigma'_{gh}\bar{b}_{n-1,g,\Phi^{\mathbb{Q}}_g} + \bar{b}_{n-1,h,\Phi^{\mathbb{Q}}_g}$$

Many of the derivatives of the loadings are zero for all maturities including $\underset{G\times H}{\bar{b}_{n,g,\mu_h}}, \underset{H\times G}{\bar{b}_{n,h,\mu^{\mathbb{Q}}_g}}, \underset{G\times H}{\bar{b}_{n,g,\delta_{1h}}}, \underset{G\times H^2}{\bar{b}_{n,g,\Phi^{\mathbb{Q}}_h}}, \underset{G\times H^2}{\bar{b}_{n,g,\Sigma_h}}, \underset{G\times G(G+1)/2}{\bar{b}_{n,g,\Sigma_{0,g}}}, \underset{H\times G(G+1)/2}{\bar{b}_{n,h,\Sigma_{0,g}}}, \underset{G\times GH}{\bar{b}_{n,g,\Phi^{\mathbb{Q}}_{gh}}}, \underset{H\times G(G+1)/2}{\bar{b}_{n,g,\Sigma_{i,g}}}, \underset{G\times GH}{\bar{b}_{n,g,\Sigma_{gh}}}, \underset{G\times H}{\bar{b}_{n,g,\nu^{\mathbb{Q}}_h}}, \underset{H\times H}{\bar{b}_{n,h,\nu^{\mathbb{Q}}_h}}$.

Appendix G USV restrictions

Appendix G.1 Proof of Proposition 4

We provide proofs for the $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$ and $\mathbb{U}_1(4)(\phi, \phi^2, \phi^4)$ models.

Proof for $\mathbb{U}_1(4)(\phi, \phi^2, \psi)$: Showing $b_{n,h} = 0$ is equivalent to $\bar{b}_{n,h} = 0$. We prove $\bar{b}_{n,h} = 0$ by induction over n. At $n = 1$, we have $\bar{b}_{1,h} = -\delta_{1,h} = 0$. Next, suppose $\bar{b}_{n,h} = 0$, then under the restriction that $\Sigma_{gh} = 0$, we find that $\bar{b}_{n,gh} = 0$. Imposing the restrictions on $\Sigma_{1,g}\Sigma'_{1,g}$ and from (6), the loading recursion reduces to

$$\bar{b}_{n+1,h} = \Phi^{\mathbb{Q}}_{gh,1}\bar{b}_{n,g,1} + \Phi^{\mathbb{Q}}_{gh,2}\bar{b}_{n,g,2} + \Phi^{\mathbb{Q}}_{gh,3}\bar{b}_{n,g,3} + \frac{1}{2}\bar{b}^2_{n,g,1}\Sigma^2_{1,g,11}$$

The restrictions on $\Phi^{\mathbb{Q}}_g$ together with (7) implies the solution for $\bar{b}_{n,g}$: $\bar{b}_{n,g,1} = -\frac{1-\phi^n}{1-\phi}\delta_{1,g,1}, \bar{b}_{n,g,2} = -\frac{1-\phi^{2n}}{1-\phi^2}\delta_{1,g,2}, \bar{b}_{n,g,3} = -\frac{1-\psi^n}{1-\psi}\delta_{1,g,3}$. Substituting these above gives

$$\begin{aligned}\bar{b}_{n+1,h} &= -\Phi^{\mathbb{Q}}_{gh,1}\frac{1-\phi^n}{1-\phi}\delta_{1,g,1} - \Phi^{\mathbb{Q}}_{gh,2}\frac{1-\phi^{2n}}{1-\phi^2}\delta_{1,g,2} - \Phi^{\mathbb{Q}}_{gh,3}\frac{1-\psi^n}{1-\psi}\delta_{1,g,3} \\ &\quad + \frac{1}{2}\frac{(1-\phi^n)^2}{(1-\phi)^2}\delta^2_{1,g,1}\Sigma^2_{1,g,11}\end{aligned}$$

Collect terms related to $(\phi^n)^0, (\phi^n)^1, (\phi^n)^2, \psi^n$, we get the following three equations

$$\begin{aligned}\bar{b}_{n+1,h} &= \left(-\Phi^{\mathbb{Q}}_{gh,1}\frac{\delta_{1,g,1}}{1-\phi} - \Phi^{\mathbb{Q}}_{gh,2}\frac{\delta_{1,g,2}}{1-\phi^2} - \Phi^{\mathbb{Q}}_{gh,3}\frac{\delta_{1,g,3}}{1-\psi} + \frac{1}{2}\frac{\delta^2_{1,g,1}}{(1-\phi)^2}\Sigma^2_{1,g,11}\right) + \left(\Phi^{\mathbb{Q}}_{gh,3}\frac{\delta_{1,g,3}}{1-\psi}\right)\psi^n \\ &\quad + \left(\Phi^{\mathbb{Q}}_{gh,1}\frac{\delta_{1,g,1}}{1-\phi} - \frac{\delta^2_{1,g,1}}{(1-\phi)^2}\Sigma^2_{1,g,11}\right)\phi^n + \left(\Phi^{\mathbb{Q}}_{gh,2}\frac{\delta_{1,g,2}}{1-\phi^2} + \frac{1}{2}\frac{\delta^2_{1,g,1}}{(1-\phi)^2}\Sigma^2_{1,g,11}\right)\phi^{2n}\end{aligned}$$

The restrictions on $\Phi^{\mathbb{Q}}_{gh}$ guarantee that the coefficients in front of $(\phi^n)^0, (\phi^n)^1, (\phi^n)^2, \psi^n$ are all zero. Hence, $\bar{b}_{n+1,h} = 0$. ■

Proof for $\mathbb{U}_1(4)(\phi,\phi^2,\phi^4)$: We prove $\bar{b}_{n,h}=0$ by induction. First, at maturity $n=1$, we have $\bar{b}_{1,h}=-\delta_{1,h}=0$. Next, suppose $\bar{b}_{n,h}=0$, then $\bar{b}_{n,gh}=0$ because $\Sigma_{gh}=0$ under the USV restrictions. From (6), the non-Gaussian loading simplifies to

$$\bar{b}_{n+1,h}=\Phi_{gh}^{\mathbb{Q}\prime}\bar{b}_{n,g}+\frac{1}{2}\bar{b}_{n,g}^{\prime}\Sigma_{1,g}\Sigma_{1,g}^{\prime}\bar{b}_{n,g} \tag{G.1}$$

Substituting the parameter restrictions on $\Sigma_{1,g}\Sigma_{1,g}^{\prime}$ into the equation, we find

$$\bar{b}_{n+1,h}=\Phi_{gh,1}^{\mathbb{Q}}\bar{b}_{n,g,1}+\Phi_{gh,2}^{\mathbb{Q}}\bar{b}_{n,g,2}+\Phi_{gh,3}^{\mathbb{Q}}\bar{b}_{n,g,3}+\frac{1}{2}\bar{b}_{n,g,1}^{2}\Sigma_{1,g,11}^{2}+\frac{1}{2}\bar{b}_{n,g,2}^{2}\Sigma_{1,g,22}^{2}$$

The parameter restrictions on $\Phi_{g}^{\mathbb{Q}}$ together with (7) implies the solution for $\bar{b}_{n,g}$: $\bar{b}_{n,g,1}=-\frac{1-\phi^n}{1-\phi}\delta_{1,g,1}, \bar{b}_{n,g,2}=-\frac{1-\phi^{2n}}{1-\phi^2}\delta_{1,g,2}, \bar{b}_{n,g,3}=-\frac{1-\phi^{4n}}{1-\phi^4}\delta_{1,g,3}$. Substituting these in, we find

$$\begin{aligned}\bar{b}_{n+1,h} \quad = \quad & -\Phi_{gh,1}^{\mathbb{Q}}\frac{1-\phi^n}{1-\phi}\delta_{1,g,1}-\Phi_{gh,2}^{\mathbb{Q}}\frac{1-\phi^{2n}}{1-\phi^2}\delta_{1,g,2}-\Phi_{gh,3}^{\mathbb{Q}}\frac{1-\phi^{4n}}{1-\phi^4}\delta_{1,g,3}\\ & +\frac{1}{2}\frac{(1-\phi^n)^2}{(1-\phi)^2}\delta_{1,g,1}^2\Sigma_{1,g,11}^2+\frac{1}{2}\frac{\left(1-\phi^{2n}\right)^2}{(1-\phi^2)^2}\delta_{1,g,2}^2\Sigma_{1,g,22}^2\end{aligned}$$

Collect terms related to $(\phi^n)^0,(\phi^n)^1,(\phi^n)^2,(\phi^n)^4$

$$\begin{aligned}\bar{b}_{n+1,h} \quad = \quad & \left(-\Phi_{gh,1}^{\mathbb{Q}}\frac{\delta_{1,g,1}}{1-\phi}-\Phi_{gh,2}^{\mathbb{Q}}\frac{\delta_{1,g,2}}{1-\phi^2}-\Phi_{gh,3}^{\mathbb{Q}}\frac{\delta_{1,g,3}}{1-\phi^4}+\frac{1}{2}\frac{\delta_{1,g,1}^2}{(1-\phi)^2}\Sigma_{1,g,11}^2+\frac{1}{2}\frac{\delta_{1,g,2}^2}{(1-\phi^2)^2}\Sigma_{1,g,22}^2\right)\\ & +\left(\Phi_{gh,1}^{\mathbb{Q}}\frac{\delta_{1,g,1}}{1-\phi}-\frac{\delta_{1,g,1}^2}{(1-\phi)^2}\Sigma_{1,g,11}^2\right)\phi^n\\ & +\left(\Phi_{gh,2}^{\mathbb{Q}}\frac{\delta_{1,g,2}}{1-\phi^2}+\frac{1}{2}\frac{\delta_{1,g,1}^2}{(1-\phi)^2}\Sigma_{1,g,11}^2-\frac{\delta_{1,g,2}^2}{(1-\phi^2)^2}\Sigma_{1,g,22}^2\right)\phi^{2n}\\ & +\left(\Phi_{gh,3}^{\mathbb{Q}}\frac{\delta_{1,g,3}}{1-\phi^4}+\frac{1}{2}\frac{\delta_{1,g,2}^2}{(1-\phi^2)^2}\Sigma_{1,g,22}^2\right)\phi^{4n}\end{aligned}$$

The restrictions on $\Phi_{gh}^{\mathbb{Q}}$ guarantee that the coefficients in front of $(\phi^n)^0,(\phi^n)^1,(\phi^n)^2,(\phi^n)^4$ are all zero. Therefore, $\bar{b}_{n+1,h}=0$. ■

Appendix H EM algorithm

Appendix H.1 Intermediate quantity

Given the identifying restriction $\mu_h = 0$, we drop this parameter for convenience. The two components of the intermediate quantity $Q\left(\theta|\theta^{(i)}\right) = Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right) + Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$ are

$$\begin{aligned} Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right) &= -(T-1)\log|B_1| - \frac{T-1}{2}\log|\Omega| - \frac{1}{2}\sum_{t=2}^{T}\mathrm{tr}\left(\Omega^{-1}\eta_t\eta_t'\right) \\ &\quad -\frac{1}{2}\sum_{t=2}^{T}\mathbb{E}\left[\log\left|\Sigma_{g,t-1}\Sigma_{g,t-1}'\right|\right] - \frac{1}{2}\sum_{t=2}^{T}\mathrm{tr}\left(\mathbb{E}\left[\left(\Sigma_{g,t-1}\Sigma_{g,t-1}'\right)^{-1}\right]\varepsilon_{gt}\varepsilon_{gt}'\right) \end{aligned}$$

and

$$\begin{aligned} Q_2\left(\theta_{m,h}|\theta^{(i)}\right) &= -(T-1)\log|\Sigma_h| - \sum_{t=1}^{T-1}\sum_{i=1}^{H}\mathrm{e}_i'\Sigma_h^{-1}\mathbb{E}\left[h_t\right] - \sum_{t=1}^{T-1}\sum_{i=1}^{H}\mathrm{e}_i'\Sigma_h^{-1}\Phi_h\mathbb{E}\left[h_{t-1}\right] \\ &\quad + \sum_{t=1}^{T-1}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2}\mathbb{E}\left[\log\left(\mathrm{e}_i'\Sigma_h^{-1}h_t\right)\right] - \sum_{t=1}^{T-1}\sum_{i=1}^{H}\frac{(\nu_{h,i}-1)}{2}\mathbb{E}\left[\log\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h h_{t-1}\right)\right] \\ &\quad + \sum_{t=1}^{T-1}\sum_{i=1}^{H}\mathbb{E}\left[\log I_{\nu_{h,i}-1}\left(2\sqrt{\left(\mathrm{e}_i'\Sigma_h^{-1}h_t\right)\left(\mathrm{e}_i'\Sigma_h^{-1}\Phi_h h_{t-1}\right)}\right)\right] \end{aligned}$$

Maximizing $Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right)$ is similar to estimation of a Gaussian ATSM. When maximizing $Q_1\left(\theta_{m,b}, \theta_c|\theta^{(i)}\right)$, the analytical gradient of the intermediate quantity can be derived from the gradients in Appendix E.

There are several options for maximizing $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$, all of which lead to different types of EM algorithms. Each EM algorithm leads to the same maximum but they converge at different rates.

- Option #1: Maximize the intermediate quantity $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$ numerically over $\theta_{m,h}$ as above.
- Option #2: From the definition of h_{t+1} in (2)-(4), there is also the variable z_t. When calculating $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$, take the expectations of z_t and h_t. By introducing z_t, the maximization of Φ_h is analytical.
- Option #3: Instead of optimizing over $Q_2\left(\theta_{m,h}|\theta^{(i)}\right)$ at each iteration, a valid EM algorithm can optimize the log-likelihood over $\theta_{m,h}$. When $H = 1$, we can use the particle filter of Malik and Pitt(2011) to maximize the log-likelihood $\log p\left(Y_{1:T}^{(1)};\theta\right)$ over $\theta_{m,h}$.

For the results in the paper, we used the third option.

Appendix H.2 Particle filter

We implement a basic particle filter for the E-step in (a.). Let $q\left(h_t|h_{t-1}^{(m)}, g_{t+1}, g_t, \theta\right)$ be an importance density whose tails are heavier than the target distribution.

For $t = 1, \ldots, T$, run:

- For $m = 1, \ldots, M$, draw from a proposal distribution: $h_t^{(m)} \sim q\left(h_t|h_{t-1}^{(m)}, g_{t+1}, g_t, \theta\right)$.
- For $m = 1, \ldots, M$, calculate the importance weight: $w_t^{(m)} \propto \hat{w}_{t-1}^{(m)} \frac{p\left(g_{t+1}|g_t, h_t^{(m)}, \theta\right) p\left(h_t^{(m)}|h_{t-1}^{(m)}, \theta\right)}{q\left(h_t^{(m)}|h_{t-1}^{(m)}, g_{t+1}, g_t, \theta\right)}$
- For $m = 1, \ldots, M$, normalize the weights: $\hat{w}_t^{(m)} = \frac{w_t^{(m)}}{\sum_{m=1}^{M} w_t^{(m)}}$.
- Calculate the effective sample size: $\text{ESS}_t = \frac{1}{\sum_{m=1}^{M}\left(\hat{w}_t^{(m)}\right)^2}$
- If $\text{ESS}_t < 0.5M$ resample $\left\{h_t^{(m)}\right\}_{m=1}^{M}$ with probabilities $\left\{\hat{w}_t^{(m)}\right\}_{m=1}^{M}$ and set $\hat{w}_t = 1/M$.

At time $t = 1$, $q\left(h_1; \theta\right)$ does not depend on any previous particles. Simple proposal distributions are to draw from the transition density $p\left(h_{t+1}|h_t; \theta\right)$ of the model (2)-(4).

To calculate the expectations within $Q\left(\theta|\theta^{(i)}\right)$, we use the algorithm of Godsill, Doucet, and West(2004) that draws samples from the posterior. Store $\left\{\hat{w}_t^{(m)}, h_t^{(m)}\right\}_{m=1}^{M}$ for $t = 1, \ldots, T$ during the particle filter. On a backwards pass, sample $\tilde{h}_T = h_T^{(m)}$ with probability $\hat{w}_T^{(m)}$ and then for $t = T-1, \ldots, 1$

- For $m = 1, \ldots, M$, calculate backwards weights $w_{t+1|t}^{(m)} \propto w_t^{(m)} p\left(\tilde{h}_{t+1}|h_t^{(m)}; \theta\right)$.
- For $m = 1, \ldots, M$, normalize the weights: $\hat{w}_{t+1|t}^{(m)} = \frac{w_{t+1|t}^{(m)}}{\sum_{m=1}^{M} w_{t+1|t}^{(m)}}$.
- Sample $\tilde{h}_t = h_t^{(m)}$ with probability $\hat{w}_{t+1|t}^{(m)}$

To calculate the expectations $Q\left(\theta|\theta^{(i)}\right)$, we repeat this backwards pass a large number of times.

For the final climb after the EM algorithm, we optimize over all θ using the algorithm from Malik and Pitt(2011). To implement this particle filter, we resample at every time period and use the resampling algorithm described in their appendix.