Analysis of Multifactor Affine Yield Curve Models

Siddhartha CHIB and Bakhodir ERGASHEV

In finance and economics much work has been done on the theoretical modeling and statistical estimation of the yield curve, defined as the relationship between $-\frac{1}{\tau}\log p_t(\tau)$ and τ, where $p_t(\tau)$ is the time t price of a zero-coupon bond with payoff 1 at maturity date $t+\tau$. Of considerable current interest are models of the yield curve in which a collection of observed and latent factors determine the market price of factor risks, the stochastic discount factor, and the arbitrage-free bond prices. The model is particularly interesting from a statistical perspective, because the yields are complicated nonlinear functions of the underlying parameters (e.g., those appearing in the evolution dynamics of the factors and those appearing in the model of the factor risks). This nonlinearity tends to produce a likelihood function that is multimodal. In this article we revisit the question of how such models should be fit from the Bayesian viewpoint. Key aspects of the inferential framework include (a) a prior on the parameters of the model that is motivated by economic considerations, in particular, those involving the slope of the implied yield curve; (b) posterior simulation of the parameters in ways to improve the efficiency of the MCMC output, for example, through sampling of the parameters marginalized over the factors and tailoring of the proposal densities in the Metropolis–Hastings steps using information about the mode and curvature of the current target based on the output of a simulating annealing algorithm; and (c) measures to mitigate numerical instabilities in the fitting through reparameterizations and square root filtering recursions. We apply the techniques to explain the monthly yields on nine U.S. Treasury Bills (with maturities ranging from 1 month to 120 months) over the period January 1986–December 2005. The model contains three factors, one latent and two observed. We also consider the problem of predicting the nine yields for each month of 2006. We show that the (multi-step-ahead) prediction regions properly bracket the actual yields in those months, thus highlighting the practical value of the fitted model.

KEY WORDS: Bayesian analysis; Forecasting; Kalman filtering; Markov chain Monte Carlo; Metropolis–Hastings algorithm; No-arbitrage condition; Simulated annealing; Square root filter; Tailored proposal density; Term structure; Zero-coupon bond.

1. INTRODUCTION

In finance and economics much attention has been devoted to understanding the pricing of default-free zero-coupon bonds, such as U.S. Treasury Bills that have no risk of default and that provide a single payment—typically normalized to 1—at a future date when the bond matures, and are sold before the maturity date at a discount from the face value of 1. For bonds in general, and zero-coupon bonds in particular, a central quantity of interest is the *yield to maturity*, which is the internal rate of return of the payoffs, or the interest rate that equates the present-value of the bond payoffs (a single payoff in the case of zero-coupon bonds) to the current price. Letting τ denote the time to maturity of the bond, and $p_t(\tau)$ the price of the bond that matures at time $t+\tau$, then the yield of the bond, $z_{t\tau}$, is essentially equal to $-\frac{1}{\tau}\log p_t(\tau)$. Of crucial interest in this context is the so-called "yield curve," the set of yields that differ only in their time to maturity τ. This yield curve is generally plotted with the yields to maturity $z_{t\tau}$ against the time to maturity τ and in practice can be upward-sloping (the normal case), downward-sloping, flat, or of some other shape. A central question is to model both the determinants of the yield curve and its evolution over time. Although this modeling can be approached in different ways, ranging from the purely theoretical (i.e., with heavy reliance on economic principles) to the purely statistical (i.e., modeling the yields as a vector time series process with little connection to the underlying economics), over the past decade interest has grown in a middle ground, involving models that have a statistical orientation, and at the same time are connected to economics through the enforcement of a no-arbitrage condition on bond prices. This no-arbitrage condition is principally the statement that the expected return from the bond, net of the risk premium, at each time to maturity is equal to the risk-free rate.

The models with the foregoing features that have attracted the most attention are the multifactor affine yield curve models. This class of models was introduced in an important article by Duffie and Kan (1996). Here the modeling strategy is to explain the yield curve in terms of a collection of factors that are assumed to follow a stationary vector Markov process. These factors, along with a vector of variables that represent the market price of factor risks γ_t, are then assumed to determine the so-called "pricing kernel," or *stochastic discount factor*, $\kappa_{t,t+1}$. The market price of factor risks γ_t are in turn modeled as an affine function of the factors. The no-arbitrage condition is enforced automatically by pricing the τ period bond (which becomes a $\tau-1$ period bond next period) according to the rule that $p_t(\tau)=\mathbb{E}_t[\kappa_{t,t+1}p_{t+1}(\tau-1)]$, where $\mathbb{E}_t$ is the expectation conditioned on time t information. Duffie and Kan (1996) showed that the resulting prices, $\{p_t(\tau),\tau=1,2,3,\ldots\}$, are an exponential affine function of the factors, where the parameters of this affine function, which are a function of the deep parameters of the model, can be obtained by iterating a set of vector difference equations. Then, on taking logs and dividing by minus τ, the yields become an affine function of the factors.

The framework of Duffie and Kan provides a general approach for modeling the yield curve. Ang and Piazzesi (2003) enhanced its practical value by incorporating macroeconomic variables in the list of factors that drive the dynamics of the model. In particular, one factor is taken to be latent and two are observed macroeconomic variables. We refer to this model

Siddhartha Chib is Harry C. Hartkopf Professor, Olin Business School, Washington University in St. Louis, St. Louis, MO 63130 (E-mail: *chib@wustl.edu*). Bakhodir Ergashev is Financial Economist, Federal Reserve Bank of Richmond, Charlotte Branch, Charlotte, NC 28230 (E-mail: *bakhodir.ergashev@rich.frb.org*). The views in this article are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Federal Reserve Bank of Richmond or the Board of the Governors of the Federal Reserve System. The authors thank the editor, the referees, Kyu Ho Kang, and Srikanth Ramamurthy for their insightful and constructive comments on previous versions of the manuscript.

Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08029

as the L1M2 model. A version of this model was systematically examined by Ang, Dong, and Piazzesi (2007). A convenient statistical aspect of this multifactor affine model is that it can be expressed in linear state-space form with the transition equation consisting of the evolution process of the factors and the observation model comprising the set of yields derived from the pricing model. This model is particularly interesting from a statistical perspective because the parameters in the observation equation are highly nonlinear functions of the underlying deep parameters (e.g., the parameters appearing in the evolution dynamics of the factors and those appearing in the model of γ_t). This nonlinearity is quite severe and can produce a likelihood function that is multimodal, as we show later.

To deal with the estimation challenges, Ang, Dong, and Piazzesi (2007) adopted a Bayesian approach. One reason for pursuing the Bayesian approach is that it provides the means to introduce prior information that can be helpful in estimating parameters that are otherwise ill-determined. However, Ang, Dong, and Piazzesi (2007) used diffuse priors and thus did not fully exploit this aspect of the Bayesian approach. Another reason is because summaries of the posterior distribution, such as the posterior expectations and posterior credibility intervals of parameters, can be easier to interpret than the (local) mode of an irregular likelihood function. Ang, Dong, and Piazzesi (2007) demonstrated the value of the Bayesian approach by estimating the L1M2 model on quarterly data and yields of maturities up to 20 quarters. They used a specific variant of a Markov chain Monte Carlo (MCMC) method (in particular a random-walk–based Metropolis–Hastings sampler) to sample the posterior distribution of the parameters. For the most part, they concentrated on the finance implications of the fitting and did not discuss how well their MCMC approach actually performed in terms of the metrics common in the Bayesian literature. For instance, they did not provide inefficiency factors and other related measures that can be useful in evaluating the efficiency of MCMC sampling (Chib 2001; Liu 2001; Robert and Casella 2004).

In this article we continue the Bayesian study of the L1M2 multifactor affine yield curve model. Our contributions address several interrelated issues. First, we formulate our prior distribution to incorporate the belief of a positive term premium, because a diffuse or vague prior on the parameters can imply a yield curve that is a priori unreasonable. In our view, it is important that the prior be formulated with the yield curve in mind. Such a prior is easier to motivate and defend and in practice is helpful in model estimation, because it tends to smooth out and diminish the importance of regions of the parameter space that are a priori unlikely. Second, in an attempt to deal with the complicated posterior distribution, we pursue a careful MCMC strategy in which the parameters of the model are grouped into blocks and each block is sampled in turn within each sweep of the MCMC algorithm with the help of the Metropolis–Hastings (MH) algorithm, with proposal densities that are tailored to the conditional posterior distribution of that block, along the lines of Chib and Greenberg (1994). A noteworthy point is that these tailored proposal densities are based on the output of the simulated annealing method. Third, we sample the parameters marginalized over the factors, because factors and the parameters are confounded in such models (Chib, Nardari, and Shephard 2006). Finally, we consider the problem of forecasting the yield curve. In the context of our model and data, we generate 1- to 12-month-ahead Bayesian predictive densities of the yield curve. For each month in the forecast period, the observed yield curve is properly bracketed by the 95% prediction region. We take this as evidence that the L1M2 model is useful for applied work.

The rest of the article is organized as follows. Section 2 introduces the arbitrage-free model, the identification restrictions, and the data used in the empirical analysis. Section 3 presents the state-space form of the model, the likelihood function, and the prior distribution. It then discusses how the resulting posterior distribution is summarized by MCMC methods. Section 4 presents results from our analysis of the L1M2 model. Section 5 summarizes our conclusions. Appendixes provide details, including those related to the instability of the coefficients in the state-space model to changes in the parameter values.

2. ARBITRAGE–FREE YIELD CURVE MODELING

Suppose that in a given market at some discrete time t, we are interested in pricing a family of default-free zero-coupon bonds that provide a payoff of 1 unit at (time to) maturity τ (measured in, say, months). As is well known, arbitrage opportunities across bonds of different maturities are precluded if the price $p_t(\tau)$ of the bond maturing in period $(t+\tau)$, which becomes a $(\tau-1)$ period bond at time $(t+1)$, satisfies the conditions

$$p_t(\tau) = \mathbb{E}_t[\kappa_{t+1} p_{t+1}(\tau-1)], \qquad t = 1, 2, \ldots, n, \tau = 1, 2, \ldots, \tau^*, \quad (2.1)$$

where $\mathbb{E}_t$ is the expectation conditioned on time t information and $\kappa_{t+1} > 0$ is the so-called "pricing kernel" (i.e., the stochastic discount factor). The goal is to model the yields

$$z_{t\tau} = -\frac{1}{\tau}\log p_t(\tau), \qquad t = 1, 2, \ldots, n, \tau = 1, 2, \ldots, \tau^*,$$

for each time t and each maturity τ.

Let u_t be a latent variable, $\mathbf{m}_t = (m_{1t}, m_{2t})'$ be a 2-vector of observed macroeconomic variables, and $\mathbf{f}_t = (u_t, \mathbf{m}_t')'$ be the stacked vector of latent and observed factors. In the affine model, it is assumed that these factors follow the vector Markov process,

$$\underbrace{\begin{pmatrix} u_t \\ \mathbf{m}_t \end{pmatrix}}_{\mathbf{f}_t} - \underbrace{\begin{pmatrix} \mu_u \\ \boldsymbol{\mu}_m \end{pmatrix}}_{\boldsymbol{\mu}} = \underbrace{\begin{pmatrix} G_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}}_{\mathbf{G}} \left(\begin{pmatrix} u_{t-1} \\ \mathbf{m}_{t-1} \end{pmatrix} - \begin{pmatrix} \mu_u \\ \boldsymbol{\mu}_m \end{pmatrix} \right) + \underbrace{\begin{pmatrix} \eta_{ut} \\ \boldsymbol{\eta}_{mt} \end{pmatrix}}_{\boldsymbol{\eta}_t}, \quad (2.2)$$

where $\mathbf{G}$ is a matrix with eigenvalues < 1 in absolute value

$$\boldsymbol{\eta}_t | \boldsymbol{\Omega} \sim \text{iid}\, \mathcal{N}_{k+m}(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = \begin{pmatrix} \Omega_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}' & \boldsymbol{\Omega}_{22} \end{pmatrix},$$

and $\mathcal{N}_{k+m}(\mathbf{0}, \boldsymbol{\Omega})$ is the 3-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$.

Next suppose, in the manner of Duffie and Kan (1996), Dai and Singleton (2000), Dai and Singleton (2003), and Ang and Piazzesi (2003), that the stochastic discount factor is given by

$$\kappa_{t,t+1} = \exp\{-\delta_1 - \boldsymbol{\delta}_2'\mathbf{f}_t - \tfrac{1}{2}\boldsymbol{\gamma}_t'\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_t'\mathbf{L}^{-1}\boldsymbol{\eta}_{t+1}\}, \quad (2.3)$$

where δ_1 and $\boldsymbol{\delta}_2$ are constants, $\mathbf{L}$ is a lower triangular matrix such that $\mathbf{L}\mathbf{L}' = \boldsymbol{\Omega}$, and $\boldsymbol{\gamma}_t$ is a vector of time-varying market prices of factor risks that is an affine function of the factors

$$\boldsymbol{\gamma}_t = \boldsymbol{\gamma} + \boldsymbol{\Phi}\mathbf{f}_t. \quad (2.4)$$

In the sequel, we call $\boldsymbol{\gamma}: 3 \times 1$ and $\boldsymbol{\Phi}: 3 \times 3$ the risk premia parameters.

Under these conditions, following Duffie and Kan (1996), the arbitrage-free bond prices are given by

$$p_t(\tau) = \exp\{-a_\tau - \mathbf{b}_\tau'\mathbf{f}_t\},$$

where a_τ and $\mathbf{b}_\tau$ are obtained from the following set of vector difference equations:

$$a_{j+1} = a_j + \mathbf{b}_j'\{(\mathbf{I} - \mathbf{G})\boldsymbol{\mu} - \mathbf{L}\boldsymbol{\gamma}\} - \tfrac{1}{2}\mathbf{b}_j'\boldsymbol{\Omega}\mathbf{b}_j + \delta_1, \quad (2.5)$$

$$\mathbf{b}_{j+1} = (\mathbf{G} - \mathbf{L}\boldsymbol{\Phi})'\mathbf{b}_j + \boldsymbol{\delta}_2, \quad j = 1, 2, \ldots, \tau, \tau = 1, 2, \ldots, \tau^*. \quad (2.6)$$

In practice, the recursions that we work with take a slightly different form,

$$a_{j+1} = a_j + \mathbf{b}_j'\{(\mathbf{I} - \mathbf{G})\boldsymbol{\mu} - \mathbf{L}\mathbf{H}^{-1}\boldsymbol{\gamma}\} - \tfrac{1}{2}\mathbf{b}_j'\boldsymbol{\Omega}\mathbf{b}_j/1{,}200 + \delta_1, \quad (2.7)$$

$$\mathbf{b}_{j+1} = (\mathbf{G} - \mathbf{L}\mathbf{H}^{-1}\boldsymbol{\Phi})'\mathbf{b}_j + \boldsymbol{\delta}_2, \quad j = 1, 2, \ldots, \tau, \tau = 1, 2, \ldots, \tau^*. \quad (2.8)$$

In these revised expressions, the number 1200 comes from multiplying the original yields (which are small numbers and thus can cause problems in the fitting) by 1200 to convert the yields to annualized percentages. The matrix $\mathbf{H}$, which is diagonal, is given by

$$\mathbf{H} = \operatorname{diag}(100, 100, 1200).$$

This arises from a similar conversion applied to the factors. In particular, because one of the macroeconomic factors that we specify later (namely capacity utilization) is expressed as a monthly proportion, while the other factor (namely inflation) is a monthly decimal increment, we multiply capacity utilization by 100 to convert it to a percentage, and multiply inflation by 1200 to convert it to an annualized percentage. We also multiply the latent factor by 100 to make the three factors comparable.

One should note that a_τ and $\mathbf{b}_\tau$ are highly nonlinear functions of the unknown parameters of the factor evolution and stochastic discount factor specifications. It is this complicated nonlinearity that makes the analysis of this model difficult.

If we now assume that each yield is subject to measurement or pricing error, then the theoretical model of the yield curve each time t can be expressed as

$$z_{t\tau} = \frac{1}{\tau}a_\tau + \frac{1}{\tau}\mathbf{b}_\tau'\mathbf{f}_t + \varepsilon_{t\tau}, \quad t = 1, 2, \ldots, n, \tau = 1, 2, \ldots, \tau^*, \quad (2.9)$$

where the first equation in this system is the short rate equation

$$z_{t1} = \delta_1 + \boldsymbol{\delta}_2\mathbf{f}_t + \varepsilon_{t1} \quad (2.10)$$

and the errors $\varepsilon_{t\tau}|\sigma_\tau \sim \text{iid}\,\mathcal{N}(0, \sigma_\tau^2)$.

2.1 Identification Restrictions

As is well known in the context of factor models, rotations and linear transformations applied to the latent factors result in observationally equivalent systems. Thus, for identification purposes, we impose some restrictions on the parameters. Following Dai and Singleton (2000), we assume that G_{11} is positive, the first element of $\boldsymbol{\delta}_2$ (that corresponding to the latent factor) is positive, μ_u is 0, and $\boldsymbol{\Omega}_{11}$ is 1. Although not strictly necessary, we further assume that $\boldsymbol{\Omega}_{12}$ is the 0 row vector. These additional restrictions are not particularly strong, but they have the effect of improving inferences about the remaining parameters.

In addition, we require that all eigenvalues of the matrix $\mathbf{G}$ be <1 in absolute value. This constraint is the stationarity restriction on the factor evolution process. We also impose a similar eigenvalue restriction on the matrix,

$$\mathbf{G} - \mathbf{L}\mathbf{H}^{-1}\boldsymbol{\Phi},$$

to ensure that the no-arbitrage recursions are stable. Under these assumptions, and following the approach of Dai and Singleton (2000), it can be shown that the preceding model is identified.

2.2 Empirical State-Space Formulation

A useful feature of affine models for statistical analysis is that they can be cast in linear state-space form, comprising the measurement equations for the yields and the evolution equations of the factors. To begin, we need to fix the maturities of interest. Suppose that interest centers on the maturities in the set $A = \{\tau_1, \tau_2, \ldots, \tau_p\}$ where, for example, $A = \{1, 3, 6, 12, 24, 36, 60, 84, 120\}$ as in our empirical example. In that case, the yields of interest at each time t are given by $\mathbf{z}_t = (z_{t1}, \ldots, z_{tp})'$, where $z_{ti} \equiv z_{t\tau_i}$ with $\tau_i \in A$, $i = 1, 2, \ldots, p$, are expressed in (2.9).

Starting with the measurement equations, let $\bar{\mathbf{a}} = (\bar{a}_{\tau_1}, \ldots, \bar{a}_{\tau_p})': p \times 1$ and $\bar{\mathbf{B}} = (\bar{\mathbf{b}}_{\tau_1}, \ldots, \bar{\mathbf{b}}_{\tau_p})': p \times 3$ such that $\bar{a}_{\tau_i} = a_{\tau_i}/\tau_i$ and $\bar{\mathbf{b}}_{\tau_i} = \mathbf{b}_{\tau_i}/\tau_i$, where a_{τ_i} and $\mathbf{b}_{\tau_i}$ are obtained by iterating the recursions sequentially in (2.7) and (2.8) from $j = 1$ to τ_i. Then, from (2.9), it follows that, conditioned on the factors and the parameters, we have

$$\mathbf{z}_t = \bar{\mathbf{a}} + \bar{\mathbf{B}}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t|\boldsymbol{\Sigma} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}), t = 1, 2, \ldots, n,$$

where $\boldsymbol{\Sigma}$ is diagonal with unknown elements given by $(\sigma_1^2, \ldots, \sigma_p^2)$. Note that $\bar{\mathbf{a}}$ and $\bar{\mathbf{B}}$ must be recalculated for every new value of the parameters.

Because the factors in this case contain some observed components (namely $\mathbf{m}_t$), we must ensure that these are inferred without error. We can achieve this by defining the outcome as

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{z}_t \\ \mathbf{m}_t \end{pmatrix}$$

and then letting the measurement equations of the state-space model take the form

$$\underbrace{\begin{pmatrix} \mathbf{z}_t \\ \mathbf{m}_t \end{pmatrix}}_{\mathbf{y}_t} = \underbrace{\begin{pmatrix} \bar{\mathbf{a}} \\ \mathbf{0}_{3\times 1} \end{pmatrix}}_{\mathbf{a}} + \underbrace{\begin{pmatrix} \bar{\mathbf{B}}_{p\times 3} \\ \mathbf{J}_{2\times 3} \end{pmatrix}}_{\mathbf{B}}\mathbf{f}_t + \underbrace{\begin{pmatrix} \mathbf{I}_p \\ \mathbf{0}_{2\times p} \end{pmatrix}}_{\mathbf{T}}\boldsymbol{\varepsilon}_t, \quad (2.11)$$

where $\mathbf{J} = (\mathbf{0}_{2\times 1}, \mathbf{I}_2) : 2 \times 3$. The rest of the state-space model is given by the set of evolution equations in (2.2).

In practice, it is helpful to parameterize the factors in terms of deviations from $\boldsymbol{\mu}$ as

$$\breve{\mathbf{f}}_t = (\mathbf{f}_t - \boldsymbol{\mu}),$$

in which case the model of interest becomes

$$\mathbf{y}_t = \mathbf{a} + \mathbf{B}(\breve{\mathbf{f}}_t + \boldsymbol{\mu}) + \mathbf{T}\boldsymbol{\varepsilon}_t, \tag{2.12}$$

$$\breve{\mathbf{f}}_t = \mathbf{G}\breve{\mathbf{f}}_{t-1} + \boldsymbol{\eta}_t, \qquad t \le n, \tag{2.13}$$

where, at $t = 0$, $\breve{\mathbf{f}}_0 = (u_0, \mathbf{m}_0 - \boldsymbol{\mu})$ with the parameter $\boldsymbol{\mu}$ present in $\breve{\mathbf{f}}_0$. We now assume that $\mathbf{m}_0$ is known from the data and that u_0, independent of $\mathbf{m}_0$, follows the stationary distribution

$$u_0 \sim \mathcal{N}(0, V_u), \tag{2.14}$$

where $V_u = (1 - G_{11}^2)^{-1}$.

Thus, (2.7), (2.8), (2.12), (2.13), and (2.14) define the model that we study in this article.

2.3 Data

The term structure data used in this study is the collection of historical yields of Constant Maturity Treasury (CMT) securities that are computed by the U.S. Treasury and published in the Federal Reserve Statistical Release H.15. It is available online from the Federal Reserve Bank of St. Louis FREDII database. The data set covers the period between January 1986 and December 2006 (for a sample size of 252) on nine yields of 1-, 3-, 6-, 12-, 24-, 36-, 60-, 84-, and 120-month maturities. We use this time span because monetary policy remained relatively stable during this period.

The model is estimated on data up to December 2005. The last 12 months of the sample is used for prediction and validation purposes. Our proxy for the 1-month yield is the federal funds rate (FFR), as suggested by Duffee (1996) and Piazzesi (2003), among others. Note that Treasury bonds of over 1 year pay semiannual coupon payments, whereas Treasury bills (of maturities of 1 year or less) pay no coupons. We extract the implied zero-coupon yield curves by the interpolation method used by the U.S. Treasury.

The macroeconomic factors in this study are the manufacturing capacity utilization (CU) and the annual price inflation (Infl) rates, both measured in percentages (as in, e.g., Ang and Piazzesi 2003). These data are taken from the Federal Reserve Bank of St. Louis' FRED II database.

Figure 1 provides a graphical representation of our data. The top panel presents the time series plots of the yields in three and two dimensions, and the bottom panel presents time series plots of our macroeconomic factors. Table 1 provides a descriptive summary of these data.

3. PRIOR–POSTERIOR ANALYSIS

3.1 Preliminaries

When doing inference about the unknown parameters, it is helpful (both for specifying the prior distribution and for conducting the subsequent MCMC simulations) to assemble the unknowns into separate blocks. To begin, we let

$$\boldsymbol{\theta}_1 = (g_{11}, g_{22}, g_{33})' \qquad \text{and}$$

$$\boldsymbol{\theta}_2 = (g_{12}, g_{13}, g_{21}, g_{31}, g_{23}, g_{32})',$$

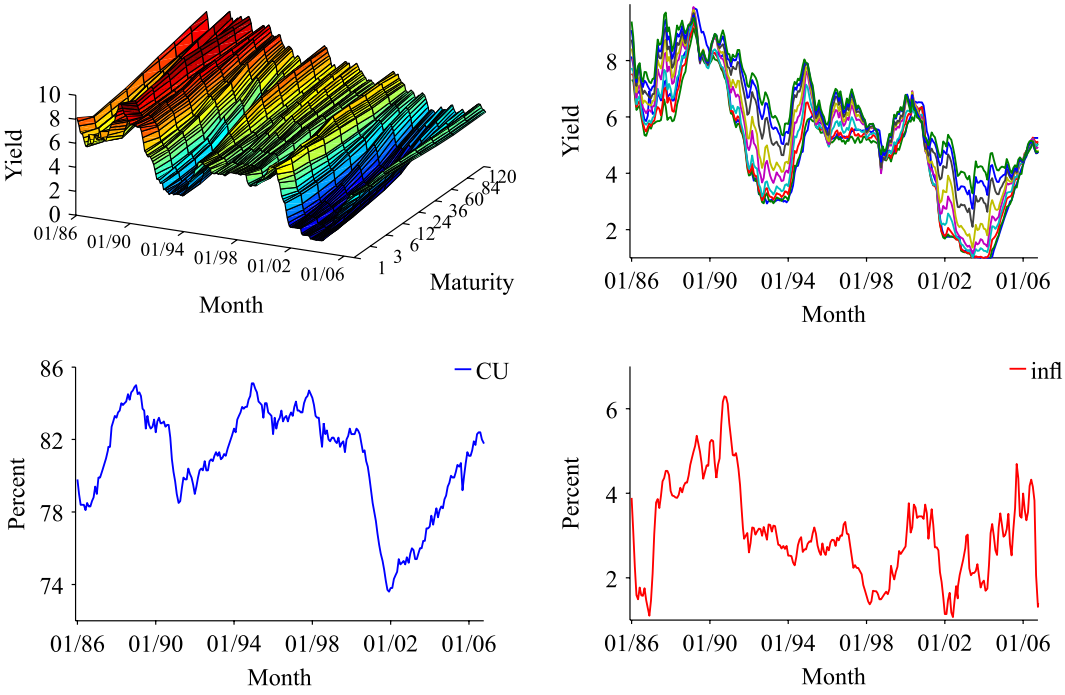

Figure 1. Term structure of U.S. Treasury interest rates and macroeconomic variables. The data cover the period between January 1986 and December 2006. The yields data consist of nine time series of length 252 on the short rate (approximated by the federal funds rate) and the yields of the following maturities: 3, 6, 12, 24, 36, 60, 84, and 120 months. These data are presented in the top two graphs in the form of three-dimensional and two-dimensional plots. The macroeconomic variables are the Manufacturing Capacity Utilization (CU) and the Consumer Price Index (Infl). (Source: Federal Reserve Bank of St. Louis FREDII database.)

Table 1. Descriptive statistics for the macro factors and the yields

Macro variables	CU	Infl							
Sample average (%)	80.92	3.04							
Standard deviation	(2.82)	(1.11)							
Bond maturity (month)	1	3	6	12	24	36	60	84	120
Average yield (%)	4.96	4.71	4.89	5.06	5.39	5.61	5.96	6.22	6.41
Standard deviation	(2.17)	(1.97)	(2.00)	(1.99)	(2.01)	(1.93)	(1.77)	(1.68)	(1.63)

NOTE: This table presents the descriptive statistics for the macro factors, the short rate (approximated by the FFR) that corresponds to the yield on 1 month, and eight yields on the constant maturity Treasury securities for the period January 1986–December 2006. The macro factors are manufacturing capacity utilization (CU) and inflation (Infl), as measured by the Consumer Price Index. (Source: Federal Reserve Bank of St. Louis FREDII database.)

collecting in $\boldsymbol{\theta}_1$ those elements of $\mathbf{G}$ that are likely to be large in magnitude. We also let

$$\boldsymbol{\theta}_3 = (\phi_{11}, \phi_{22}, \phi_{23}, \phi_{32}, \phi_{33})' \qquad \text{and}$$
$$\boldsymbol{\theta}_4 = (\phi_{12}, \phi_{13}, \phi_{21}, \phi_{31})'$$

for the elements of $\boldsymbol{\Phi}$. Next, we express $\boldsymbol{\Omega}$ as $\mathbf{L}\mathbf{L}'$ and denote the three free elements of the lower-triangular $\mathbf{L}$ as

$$\boldsymbol{\theta}_5 = (l^*_{22}, l_{32}, l^*_{33}),$$

where $l_{22} = \exp(l^*_{22})$ and $l_{33} = \exp(l^*_{33})$, so that any value of $\boldsymbol{\theta}_5$ leads to a positive definite $\boldsymbol{\Omega}$ in which $\boldsymbol{\Omega}_{12}$ is zero. Also, we let

$$\boldsymbol{\theta}_6 = \boldsymbol{\delta} \qquad \text{and}$$
$$\boldsymbol{\theta}_7 = (\boldsymbol{\mu}, \boldsymbol{\gamma}).$$

Finally, because the elements σ_i^2 of the matrix $\boldsymbol{\Sigma}$ are liable to be small and to have a U-shape, with relatively larger values at the low- and high-maturity ends, we reparameterize the variances and let

$$\boldsymbol{\theta}_8 = (\sigma_1^{2*}, \ldots, \sigma_p^{2*}),$$

where $\sigma_i^{2*} = d_i\sigma_i^2$, $d_1 = d_2 = d_7 = d_8 = 10$, $d_3 = d_5 = d_6 = 100$, and $d_4 = 2000$. The choice of these d_i's is not particularly important; what is important is that we do inferences about σ_i^2 indirectly, through the much larger σ_i^{2*}. These transformations of the variances are introduced because the inverse-gamma distribution (the traditional distribution for representing beliefs about variances) is not very flexible when dealing with small quantities.

With these definitions, the unknown parameters of the model are given by $\boldsymbol{\psi} = (\boldsymbol{\theta}, u_0)$, where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i\}_{i=1}^8$. In a model with $p = 9$ yields, the dimension of each block in $\boldsymbol{\psi}$ is 3, 6, 5, 4, 3, 4, 5, 9, and 1. In addition, the parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$, $\boldsymbol{\theta}_4$, $\boldsymbol{\theta}_5$, and $\boldsymbol{\theta}_6$ are constrained to lie in the set $\mathcal{S} = \mathcal{S}1 \cap \mathcal{S}2 \cap \mathcal{S}3$, where $\mathcal{S}1 = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 : \text{abs}(\text{eig}(\mathbf{G})) < 1\}$, $\mathcal{S}2 = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4, \boldsymbol{\theta}_5 : \text{abs}(\text{eig}(\mathbf{G} - \mathbf{L}\mathbf{H}^{-1}\boldsymbol{\Phi})) < 1\}$, and $\mathcal{S}3 = \{\boldsymbol{\theta}_6 : \delta_{2u} \in R_+\}$.

Now, if we let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the data, then we can write the density of $\mathbf{y}$ given $\boldsymbol{\psi}$ as

$$\begin{aligned}\log p(\mathbf{y}|\boldsymbol{\psi}) &= -\frac{np}{2}\log(2\pi) - \sum_{t=1}^{n}\big[\log\big(\det\big(\mathbf{R}_{t|t-1}\big)\big) \\ &\quad + \big(\mathbf{y}_t - \mathbf{a} - \mathbf{B}\big(\mathbf{f}_{t|t-1} + \boldsymbol{\mu}\big)\big)' \\ &\quad \times \big(\mathbf{R}_{t|t-1}\big)^{-1}\big(\mathbf{y}_t - \mathbf{a} - \mathbf{B}\big(\mathbf{f}_{t|t-1} + \boldsymbol{\mu}\big)\big)\big], \qquad (3.1)\end{aligned}$$

where $\mathbf{f}_{t|t-1} = \mathbb{E}(\breve{\mathbf{f}}_t|\mathbf{Y}_{t-1}, \boldsymbol{\psi})$ and $\mathbf{R}_{t|t-1} = \mathbb{V}(\mathbf{y}_t|\mathbf{Y}_{t-1}, \boldsymbol{\psi})$ are the one-step-ahead forecast of the state and the conditional variance of $\mathbf{y}_t$, given information $\mathbf{Y}_{t-1} = (\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$ up to time $(t-1)$. Generally, the latter quantities can be calculated by the Kalman filtering recursions (see, e.g., Harvey 1989); however, in this model, for some parameter values, the recursions in (2.12)–(2.13) can produce large values of a_i and $\mathbf{b}_i$ (Appendix A exemplifies this possibility), and $\mathbf{R}_{t|t-1}$ can become nonpositive definite. In such cases, we invoke the square root filter (Anderson and Moore 1979; Grewal and Andrews 2001), which tends to be more stable than the Kalman filter because the state covariance matrixes are propagated in square root form. Appendix B presents this filter in notation that corresponds to our model and with details that are missing in the aforementioned references.

Another issue is that the likelihood function can be multimodal. To illustrate this problem, we consider the posterior distribution of the parameters under a flat prior. Because the prior is flat, sampled variates drawn from this posterior (by the MCMC method given in the next section) effectively reveal features of the underlying likelihood function. Figure 2 presents graphs of the resulting likelihood surface for four pairs of the parameters. These graphs are kernel-smoothed plots computed from the sampled output of the parameters. The graphs show that the likelihood has multiple modes and other irregularities. As a result, locating the maximum of the likelihood is generally infeasible even with a stochastic optimization method, such as simulated annealing.

Given these facets of the likelihood surface, it is helpful to adopt the Bayesian perspective and shift the focus away from solely the likelihood to the posterior distribution. The goal now is to specify a prior distribution that downweights regions of the parameter space that are not economically meaningful. The posterior distribution under such a prior can be smoother and better behaved than the likelihood function. To show this possibility, Figure 3 gives the corresponding bivariate posterior densities from the informed prior described in the next section. As can be seen, the prior had some tangible effects. The bivariate posterior densities are considerably smoother, and the effective support of the last two distributions has narrowed.

3.2 Prior Distribution

We arrive at the prior distribution on $\boldsymbol{\theta}$ by reasoning in terms of the yield curve implied by the prior. Specifically, we formulate a prior which implies that the yield curve is upward-sloping

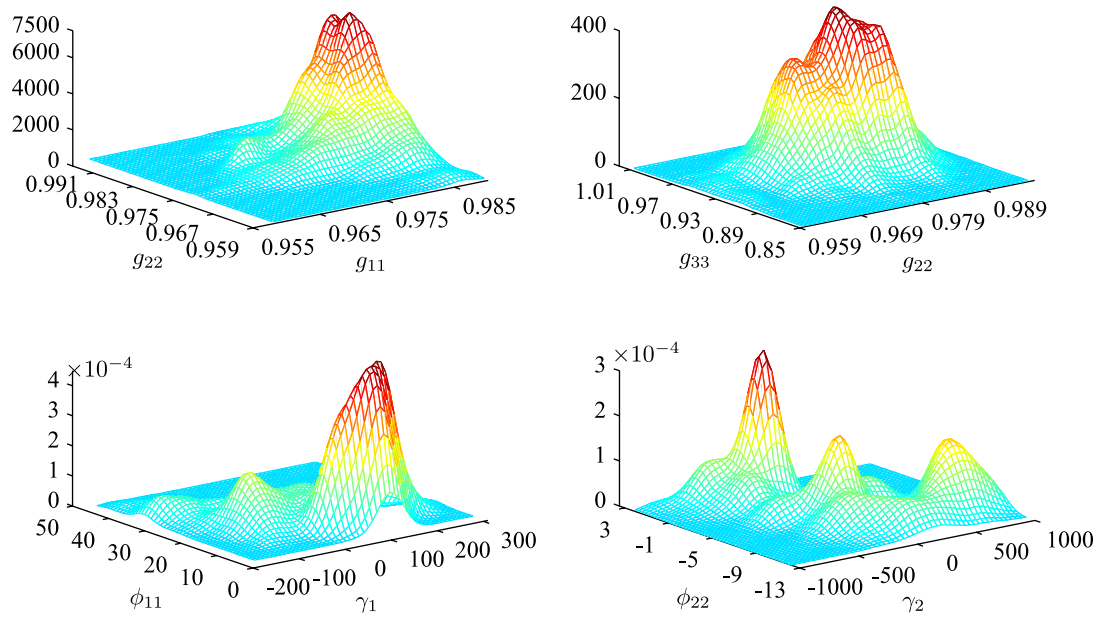

Figure 2. Kernel-smoothed likelihood surface plots for some pairs of parameters in the arbitrage-free model.

on average. The latter is of course a reasonable a priori assumption about the yield curve.

The prior is developed as follows. We start with a distribution for each block of parameters, assume independence across blocks, and sample the parameters many times. For each drawing of the parameters, we generate the time series of factors and yields. We then check whether the yield curve is upward-sloping on average for each time period in the sample. If it is not, we revise the prior distribution and repeat the process until we are satisfied with the implied yield curve over time. It is important to note that this process does not involve the observed data in any way.

- $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4, \boldsymbol{\theta}_5$, and $\boldsymbol{\theta}_6)$: We suppose that the joint distribution of these parameters is proportional to

$$\mathcal{N}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{g}_0, \mathbf{V}_g)\mathcal{N}(\boldsymbol{\theta}_3, \boldsymbol{\theta}_4 | \boldsymbol{\phi}_0, \mathbf{V}_\phi) \times \mathcal{N}(\boldsymbol{\theta}_5 | \mathbf{l}_0, \mathbf{V}_l)\mathcal{N}(\boldsymbol{\theta}_6 | \boldsymbol{\delta}_0, \mathbf{V}_\delta) I_{\mathcal{S}}.$$

For the hyperparameters, we let

$$\mathbf{g}_0 = (0.95, 0.95, 0.95, 0, 0, 0, 0, 0, 0) \quad \text{and}$$

$$\mathbf{V}_g = \text{diag}(0.10, 0.10, 0.10, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20).$$

In terms of the untruncated distribution, these choices reflect the belief that (independently) the diagonal elements are centered at 0.95 with a standard deviation of 0.32 and the off-diagonal

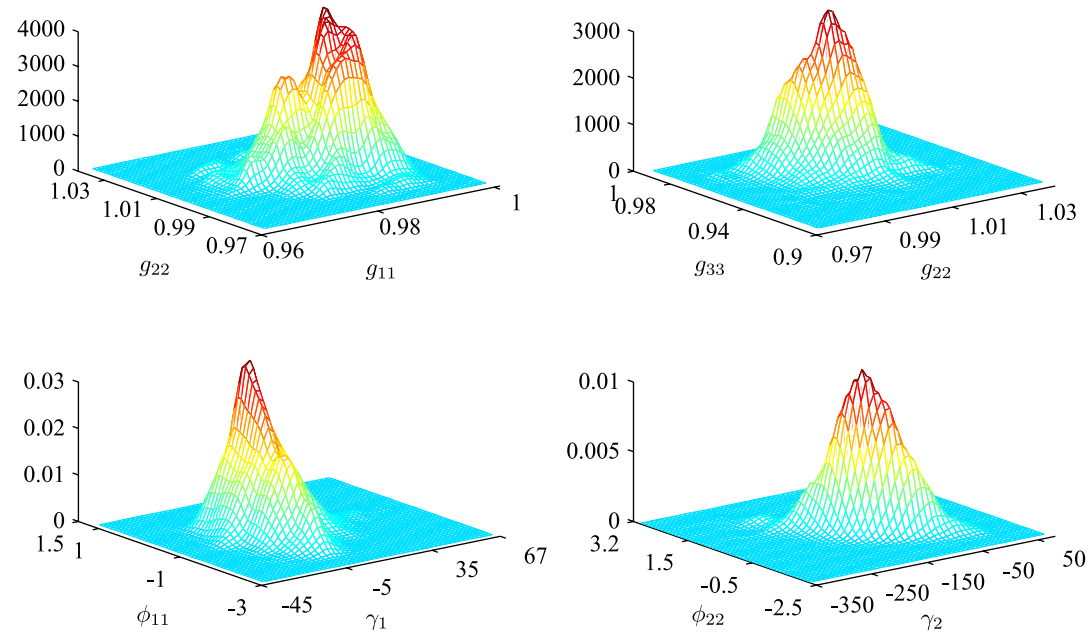

Figure 3. Kernel-smoothed posterior surface plots for some pairs of parameters in the arbitrage-free model.

elements at 0 with a standard deviation of 0.45. Given that $\mathbf{G}$ must satisfy the stationarity condition, and that the latent and macroeconomic factors can be expected to be highly persistent, the latter beliefs are both appropriate and diffuse. Next, we suppose that

$$\boldsymbol{\phi}_0 = (1, 1, 0, 0, 1, 0, 0, 0, 0) \qquad \text{and} \qquad \mathbf{V}_\phi = 2\mathbf{I}_9$$

because it can be inferred from the literature that time variation in the risk premia is driven mainly by the most persistent latent factor. In addition, we let

$$\mathbf{l}_0 = (-0.6, 0, -1) \qquad \text{and} \qquad \mathbf{V}_l = 0.30 \times \mathbf{I}_3$$

be the mean and covariance of $\boldsymbol{\theta}_5$. Thus the standard deviation of each element is 0.5, which implies a relatively diffuse prior assumption on these parameters. Finally, based on the Taylor rule intuition that high values of manufacturing capacity utilization and inflation should be associated with high short rate values, we let

$$\boldsymbol{\delta}_0 = (-3, 0.20, 0.10, 0.70)' \qquad \text{and}$$

$$\mathbf{V}_\delta = \operatorname{diag}(1, 0.2, 0.1, 0.2).$$

- $\boldsymbol{\theta}_7$: We assume that the joint distribution of these parameters is given by

$$\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \mathbf{V}_\mu)\mathcal{N}(\boldsymbol{\gamma}|\boldsymbol{\gamma}_0, \mathbf{V}_\gamma),$$

where

$$\boldsymbol{\mu}_0 = (75, 4)' \qquad \text{and} \qquad \mathbf{V}_\mu = \operatorname{diag}(49, 25),$$

so that the prior mean of capacity utilization is assumed to be 75% and that of the inflation rate is assumed to be 4% (the prior standard deviations of 7 and 5 are sufficient to cover the most likely values of these rates), and where

$$\boldsymbol{\gamma}_0 = (-100, -100, -100)',$$

$$\mathbf{V}_\gamma = \operatorname{diag}(2500, 2500, 2500).$$

The prior mean of $\boldsymbol{\gamma}$ is negative to imply an upward-sloping average yield curve.

- $\boldsymbol{\theta}_8$: We assume that

$$\sigma_i^{2*} \sim \mathcal{IG}\left(\frac{a_0}{2}, \frac{b_0}{2}\right), \qquad i = 1, \ldots, p,$$

where a_0 and b_0 are such as to imply an a priori mean of σ_i^{2*} of 5 and a standard deviation of 64. Because we have let $\sigma_i^{2*} = d_i\sigma_i^2$, this implies that the prior on the pricing error variance is maturity-specific, even though the prior on σ_i^{2*} is not.

To demonstrate what these assumptions imply for outcomes, we simulate the parameters 10,000 times from the prior, and for each drawing of the parameters simulate the factors and yields for each maturity and each of 250 months. The median, 2.5% and 97.5% quantile surfaces of the resulting yield curves are reproduced in Figure 4. It can be seen that the implied prior yield curves are positively sloped, but that there is reasonable a priori variation in the yield curves. Some of the support of the yield curves (as indicated by the 5% quantiles) is in the negative region; this shortcoming of Gaussian affine models is difficult to overcome.

3.3 Posterior and Markov Chain Monte Carlo Sampling

Under our assumptions, the posterior distribution of $\boldsymbol{\psi}$ is

$$\pi(\boldsymbol{\psi}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\psi})p(u_0|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{3.2}$$

where $p(\mathbf{y}|\boldsymbol{\psi})$ is as given in (3.1), $p(u_0|\boldsymbol{\theta})$ from (2.14) is

$$\mathcal{N}(0, V_u),$$

and $\pi(\boldsymbol{\theta})$ is proportional to

$$\begin{aligned}&\mathcal{N}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{g}_0, \mathbf{V}_\mathbf{g})\mathcal{N}(\boldsymbol{\theta}_3, \boldsymbol{\theta}_4|\boldsymbol{\phi}_0, \mathbf{V}_{\boldsymbol{\phi}})\mathcal{N}(\boldsymbol{\theta}_5|\mathbf{l}_0, \mathbf{V}_\mathbf{l})\\ &\quad\times\mathcal{N}(\boldsymbol{\theta}_6|\boldsymbol{\delta}_0, \mathbf{V}_{\boldsymbol{\delta}})I_\mathcal{S}\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \mathbf{V}_{\boldsymbol{\mu}})\mathcal{N}(\boldsymbol{\gamma}|\boldsymbol{\gamma}_0, \mathbf{V}_{\boldsymbol{\gamma}})\\ &\quad\times\prod_{i=1}^{p}\mathcal{IG}\left(\sigma_i^{2*}\Big|\frac{a_0}{2}, \frac{b_0}{2}\right).\end{aligned} \tag{3.3}$$

Summarizing this distribution is challenging even with MCMC methods for the reasons discussed earlier. For one, we need to deal with the high dimension of the parameter space and the

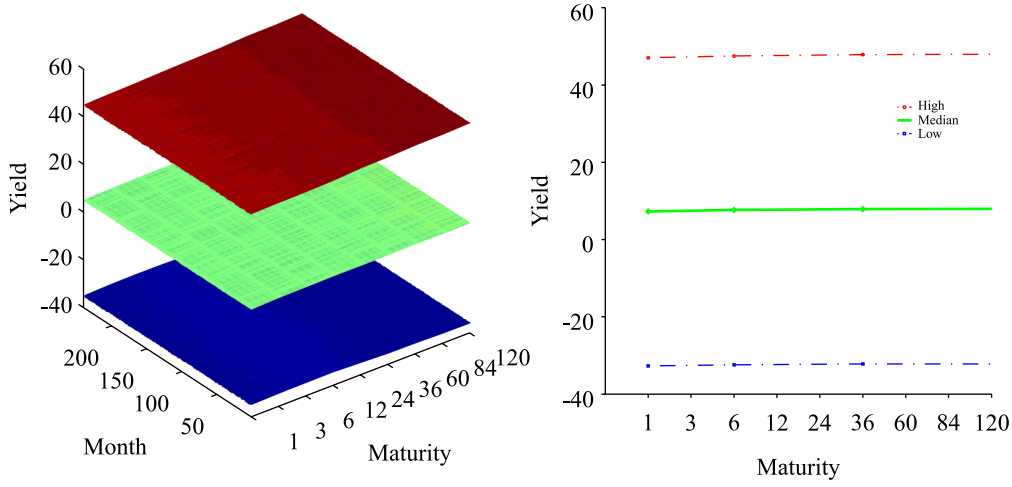

Figure 4. The implied prior yield curve dynamics. These graphs are based on 10,000 simulated draws of the parameters from the prior distribution. In the first graph, the "low," "median," and "high" surfaces correspond to the 5%, 50%, and 95% quantile surfaces of the yield curve dynamics implied by the prior distribution. In the second graph, the surfaces of the first graph are averaged over the entire period of 250 months.

facts that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are concentrated at the boundary of the parameter space—here, the stationarity region—and that inferring the market price of risk parameters is difficult. Another challenge stems from the nonlinearity of the model, arising from the recursions that produce $\bar{\mathbf{a}}$ and $\bar{\mathbf{B}}$. As a result, as shown in Figure 3, the posterior distribution is typically multimodal (but better behaved than the likelihood because of our prior). Yet another problem is that conditioning on the factors (the standard strategy for dealing with state-space models) does not help in this context, because tractable conditional posterior distributions do not emerge, except for $(u_0, \boldsymbol{\sigma})$. In fact, conditioning on the factors, as in the approach of Ang, Dong, and Piazzesi (2007), tends to worsen the mixing of the MCMC output.

After careful study of various alternatives, we have arrived at a MCMC algorithm in which the parameters are sampled marginalized over the factors. A similar approach is taken by Kim, Shephard, and Chib (1998). In addition, we sample $\{\boldsymbol{\theta}_i\}_{i=1}^{8}$ in separate blocks, as was anticipated in our discussion in Section 2, and follow that by sampling u_0. Each block is sampled from the posterior distribution of that block conditioned on the most current values of the remaining blocks. We sample each of these distributions using the MH algorithm (for example, Chib and Greenberg 1995).

Algorithm (MCMC sampling).

Step 1. Fix n_0 (the burn-in) and M (the MCMC sample size).

Step 2. For $i = 1, \ldots, 8$, sample $\boldsymbol{\theta}_i$ from $\pi(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)$, where $\boldsymbol{\theta}_i$ denotes the current parameters in $\boldsymbol{\theta}$ excluding $\boldsymbol{\theta}_i$.

Step 3. Sample u_0 from $\pi(u_0|\mathbf{y}, \boldsymbol{\theta})$.

Step 4. Repeat steps 2 and 3, discard the draws from the first n_0 iterations, and save the subsequent M draws $\{\boldsymbol{\theta}^{(n_0+1)}, \ldots, \boldsymbol{\theta}^{(n_0+M)}\}$.

A key point is that the sampling in steps 2 and 3 is done through a "tailored" MH algorithm along the lines of Chib and Greenberg (1994). In brief, we build a multivariate-t proposal density for each block of parameters in which the location and dispersion are, respectively, equated to the modal value and the negative inverse of the Hessian of the posterior density of that block of parameters. This approach has proven useful in a range of problems, although a theoretical analysis of the gains that it produces remains to be done.

For illustration, consider, for instance, block $\boldsymbol{\theta}_i$ and its target density $\pi(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)$. Assume that the value of this block after the $(j-1)$st iteration is ${\boldsymbol{\theta}_i}^{(j-1)}$. Now let

$$\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}_i} \log \pi(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0) \qquad \text{and}$$

$$\mathbf{V}_{\boldsymbol{\theta}_i} = \left(-\frac{\partial^2 \log \pi(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)}{\partial \boldsymbol{\theta}_i \, \partial \boldsymbol{\theta}_i'} \right)^{-1} \Bigg|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i}$$

denote the mode and inverse of the negative Hessian at the mode, respectively. Then, the proposal density $q(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)$ is set as a multivariate-t distribution with location $\hat{\boldsymbol{\theta}}_i$, dispersion $\mathbf{V}_{\boldsymbol{\theta}_i}$, and ζ degrees of freedom (say equal to 5):

$$q(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0) = St(\boldsymbol{\theta}_i|\hat{\boldsymbol{\theta}}_i, \mathbf{V}_{\boldsymbol{\theta}_i}, \zeta).$$

Given this proposal density, the MH step for the ith block is implemented in the usual way. One samples a proposal value

$$\boldsymbol{\theta}_i^* \sim q(\boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)$$

which is accepted as the next value ${\boldsymbol{\theta}_i}^{(j)}$ with probability given by

$$\alpha\left({\boldsymbol{\theta}_i}^{(j-1)}, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0\right) = \min\left\{ \frac{\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)}{\pi({\boldsymbol{\theta}_i}^{(j-1)}|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)} \frac{St({\boldsymbol{\theta}_i}^{(j-1)}|\hat{\boldsymbol{\theta}}_i, \mathbf{V}_{\boldsymbol{\theta}_i}, \zeta)}{St(\boldsymbol{\theta}_i^*|\hat{\boldsymbol{\theta}}_i, \mathbf{V}_{\boldsymbol{\theta}_i}, \zeta)}, 1 \right\},$$

and rejected with probability $1 - \alpha({\boldsymbol{\theta}_i}^{(j-1)}, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\theta}_{-i}, u_0)$ in which case the current value is taken as the new value.

An important point is that the modal value $\hat{\boldsymbol{\theta}}_i$ generally cannot be found by a Newton or related hill-climbing method, because of a tendency of such methods to get trapped in areas corresponding to local modes. A more effective search is possible with simulated annealing (SA) (see, e.g., Kirkpatrick, Gelatt, and Vecchi 1983; Brooks and Morgan 1995; Givens and Hoeting 2005 for detailed information about this method and its many variants) which we have found to be relatively easy to tune.

In the SA method, one searches for the maximum by proposing a random modification to the current guess of the maximum, which is then accepted or rejected probabilistically in the manner of the MH algorithm. Sometimes moves that decrease the function value are accepted. The probability of accepting such downhill moves declines over iterations according to a "cooling schedule," thereby allowing the method to converge. In our implementation, the search process is divided into various stages, denoted by k, $k = 1, 2, \ldots, K$, with the length of each stage, l_k, given by $b + l_{k-1}$, where b is a positive integer. Then the initial temperature T_0 is specified; this temperature is held constant in each stage but reduced across stages by the linear schedule $T_k = aT_{k-1}$, where $0 < a < 1$ is the cooling constant. Then, starting from an initial guess of the maximum, l_k times within each stage, random-walk proposals with a Gaussian increment of variance S are generated for a randomly chosen element. Perturbations resulting in a higher function value are always accepted, whereas those resulting in a lower function evaluation are accepted with probability

$$p = \exp\{\Delta[\log \pi]/T\},$$

where $\Delta[\log \pi]$ is the change in the log of the objective function, computed as the log of the objective function at the perturbed value of the parameters minus the log of the objective function at the existing value of the parameters. We tuned the various SA parameters in some preliminary runs, taking account of the computational load and the resulting efficiency of the method. This tuning produced the values $T_0 = 2$, $a = 0.5$, $K = 4$, $l_0 = 10$, $b = 10$, and $S = 0.1$, independent of the blocks. Note that we reduce the temperature relatively quickly, because for tailoring purposes it is enough to find the approximate modal value.

This completes the description of our MCMC algorithm.

3.4 Prediction

In practice, one is interested in the question of how well the affine model is able to predict the yields and macroeconomic factors out of sample. As is customary in the Bayesian context, we address this question by calculating the density of the future observations, conditioned on the sample data but marginalized over the parameters and the factors, where the marginalization

is with respect to the posterior distribution of the parameters and the factors. By the method of composition, for each drawing of the parameters from the MCMC algorithm, we draw the latent factors and the macroeconomic factors in the forecast period from the evolution equation of the factors, conditioned on $\breve{\mathbf{f}}_n$; then given the factors and the parameters, we sample the yields from the observation density for each time period in the forecast sample. This sample of yields from the predictive density can be summarized in terms of its quantiles and moments.

Algorithm (Sampling the predictive density of the macroeconomic factors and yields).

Step 1. For $j = 1, 2, \ldots, M$:
- (a) Compute $\bar{\mathbf{a}}^{(j)}$ and $\bar{\mathbf{B}}^{(j)}$ from the recursive equations (2.7)–(2.8), and the remaining matrixes of the state-space model, given $\boldsymbol{\theta}^{(j)}$ and $\breve{\mathbf{f}}_n^{(j)}$.
- (b) For $t = 1, 2, \ldots, T$:
  - (i) Compute $\breve{\mathbf{f}}_{n+t}^{(j)} = \mathbf{G}^{(j)}\breve{\mathbf{f}}_{n+t-1}^{(j)} + \boldsymbol{\eta}_{n+t}^{(j)}$, where $\boldsymbol{\eta}_{n+t}^{(j)} \sim \mathcal{N}_{k+m}(\mathbf{0}, \boldsymbol{\Omega}^{(j)})$.
  - (ii) Compute $\mathbf{z}_{n+t}^{(j)} = \bar{\mathbf{a}}^{(j)} + \bar{\mathbf{B}}^{(j)}(\breve{\mathbf{f}}_{n+t}^{(j)} + \boldsymbol{\mu}^{(j)}) + \boldsymbol{\varepsilon}_{n+t}^{(j)}$, where $\boldsymbol{\varepsilon}_{n+t}^{(j)} \sim \mathcal{N}_p(\mathbf{0}, \operatorname{diag}(\boldsymbol{\sigma}^{(j)}))$.
  - (iii) Set $\mathbf{y}_{n+t}^{(j)} = \{\mathbf{z}_{n+t}^{(j)}, \mathbf{m}_{n+t}^{(j)}\}$.
- (c) Save $\mathbf{y}_f^{(j)} = \{\mathbf{y}_{n+1}^{(j)}, \ldots, \mathbf{y}_{n+T}^{(j)}\}$.

Step 2. Return $\mathbf{y}_f = \{\mathbf{y}_f^{(1)}, \ldots, \mathbf{y}_f^{(M)}\}$.

4. RESULTS

Our results are based on $M = 25{,}000$ iterations of the MCMC algorithm beyond a burn-in of $n_0 = 5000$ iterations. We give summaries of the sampled output, and in order to show the efficiency of the sampling procedure, we also give the value of the MH acceptance rate (by block) and the average value of the inefficiency factor (by block), defined for each sampled sequence within each block as

$$1 + 2\sum_{l=1}^{N}\left(1 - \frac{l}{N}\right)\rho(l), \tag{4.1}$$

where $\rho(l)$ is the sample autocorrelation at lag l from each MCMC sequence and $N = 500$.

For contrast, we also compute the results (that we do not report, however) from a random-walk MH (RW-MH) algorithm. We adopt the same blocking structure as our tailored algorithm, and sample $\boldsymbol{\theta}$ marginalized over the factors. For each block, output from the simulated annealing algorithm is used to determine the negative of the inverse Hessian at the mode of the current posterior. This matrix is scaled downward by a factor of 0.01 or 0.001 to form the variance of the increment in the random-walk proposal densities. We find is that the results are similar but that the inefficiency factors are on average 2.4 times higher than those from our tailored MCMC algorithm. If we eliminate any of the elements just described (e.g., sampling $\boldsymbol{\theta}$ without marginalizing out the factors, or not using simulated annealing to define the covariance matrix of the increments), then the mixing of the RW-MH algorithm worsens further.

4.1 Estimates of $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\delta}$

The estimates of the $\mathbf{G}$ matrix given in Table 2 show that it is essentially diagonal and that the elements corresponding to the macroeconomic factors are close to 1.

The intercept of the short-rate equation δ_1 is significantly negative. A negative intercept is necessary to keep the mean of the short rate low when the factor loadings of all three factors (i.e., $\boldsymbol{\delta}_2$) are positive and significantly different from 0. These estimates are consistent with the Taylor rule intuition. The estimates of the mean parameters of the macroeconomic factors lie within $1/2$ standard deviation of their sample means. As shown in the last two columns of the table, the inefficiency factors are somewhat large. The important point is that these factors are much larger from an algorithm that is not as well tuned.

Figure 5 presents the prior-posterior updates of selected parameters from Table 2. These updates show that the prior and posterior densities are generally different, indicating that the data carries information, or in other words, that there is learning from the data.

4.2 Risk Premia Parameters

The factor risk parameters $\boldsymbol{\gamma}$, given in Table 3, have negative posterior support except for the first one. This finding is consistent with a yield curve that is upward-sloping on average. Moreover, the relatively large value (in absolute terms) of the constant prices of risk of the latent factor suggests that the latent factor is largely responsible for determining the level of the yield curve.

We also see that the estimate of the time-varying risk premium of inflation $\boldsymbol{\phi}_{33}$ is positive. This suggests that investors demand higher compensation for the risk of inflation rising above its average. Figure 6, however, shows that accurately estimating some of the risk premia parameters in $\boldsymbol{\Phi}$ is difficult.

4.3 Covariance Matrices

We note from Table 4 and Figure 7 that the posterior mean of $\boldsymbol{\sigma}$ is large for the short and long maturities. This is not surprising, due to the fact that we have approximated the short rate by the FFR, which is much less volatile than any other yield. An alternative approach would be to assume that the short rate is unobserved; however, we have found that in this case it becomes more difficult to infer the short rate parameters, $\boldsymbol{\delta}$. Because the parameters of the model are all scrambled together through the no-arbitrage recursions, the difficulty in inferring $\boldsymbol{\delta}$ makes it more difficult to infer other parameters of the model.

4.4 Predictive Densities

As can be seen from Figure 8 the predictive performance of the model is quite good. In the out-of-sample forecast for the 12 months of 2006, based on information from 1986–2005, the observed yield curve lies between the 2.5% ("low") and 97.5% ("high") quantile surfaces of the yield curve forecasts.

In addition, the model closely predicts the future dynamics of both macroeconomic factors. Except for 1 month in the forecast sample, the observed time series of the macroeconomic factors lies between the low and high quantiles of the forecasts.

Although the yield curve forecasts are quite good, Figure 8 indicates that some room for improvement remains. In particular, the forecasts do not adequately capture the curvature of the yield curve. This shortcoming can likely be overcome by

Table 2. Estimates of $\mathbf{G}$, $\boldsymbol{\mu}$, and $\boldsymbol{\delta}$

Parameters	Prior			Posterior			Average acc. rate	Average ineff.
$\mathbf{G}$	0.95	0.00	0.00	0.99	−0.02	0.06	39.7	161.5
	(0.33)	(1.41)	(1.41)	(0.01)	(0.01)	(0.02)		
	0.00	0.95	0.00	0.00	1.00	−0.09		
	(1.41)	(0.33)	(1.41)	(0.00)	(0.01)	(0.02)		
	0.00	0.00	0.95	0.00	0.01	0.96		
	(1.41)	(1.41)	(0.33)	(0.00)	(0.00)	(0.01)		
$\boldsymbol{\mu}$	0	75.00	4.00	0	76.25	2.75	87.3	69.7
	...	(7.00)	(5.00)	...	(4.46)	(0.78)		
δ_1	−3.000			−3.650			88.3	62.8
	(1.000)			(0.992)				
$\boldsymbol{\delta}_2$	0.200	0.100	0.700	0.233	0.107	0.163	88.3	109.7
	(0.447)	(0.333)	(0.447)	(0.011)	(0.017)	(0.053)		

NOTE: Acceptance rates (acc. rate) are given as percentages. Inefficiency factors (ineff.) are computed by (4.1). Standard deviations are given in parentheses.

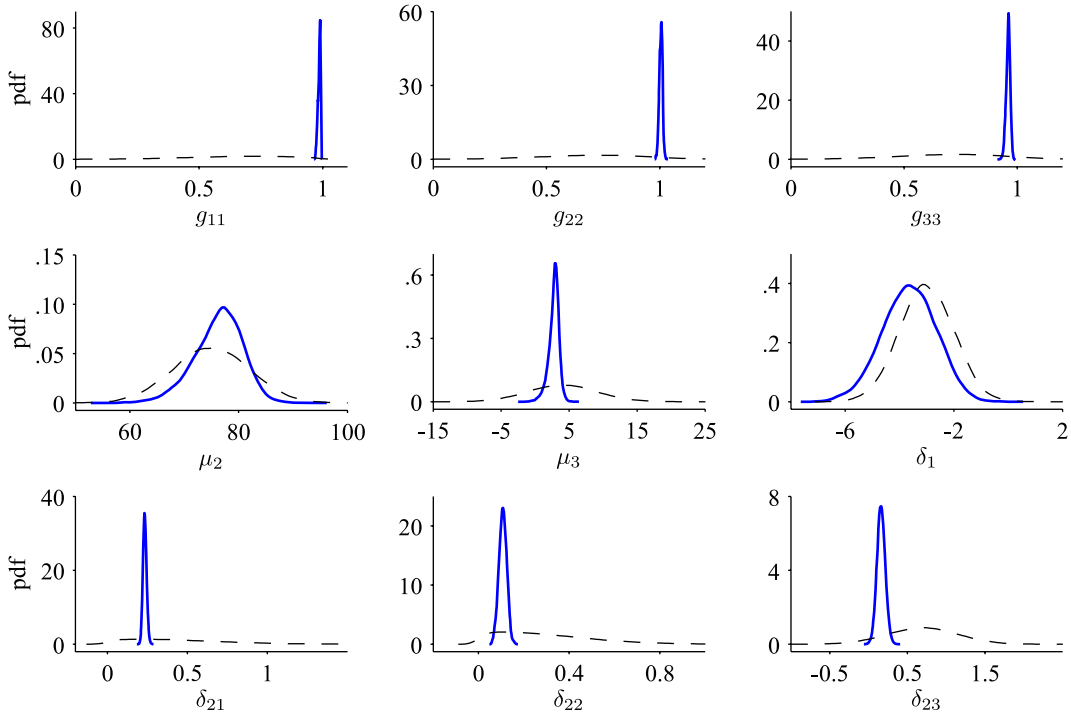

Figure 5. Prior-posterior updates of selected parameters from Table 2.

Table 3. Estimates of the risk premia parameters

Parameters	Prior			Posterior			Average acc. rate	Average ineff.
$\boldsymbol{\gamma}$	−100	−100	−100	1.96	−114.2	−97.8	87.3	146.1
	(50)	(50)	(50)	(10.0)	(42.0)	(55.1)		
$\boldsymbol{\Phi}$	1.00	0.00	0.00	−0.66	−0.11	−0.72	68.6	139.1
	(1.41)	(1.41)	(1.41)	(0.54)	(0.14)	(1.41)		
	0.00	1.00	0.00	−3.54	−0.04	0.12		
	(1.41)	(1.41)	(1.41)	(1.04)	(0.56)	(1.32)		
	0.00	0.00	1.00	0.16	0.19	0.83		
	(1.41)	(1.41)	(1.41)	(1.42)	(1.39)	(1.51)		

NOTE: Acceptance rates (acc. rate) are given as percentages. Inefficiency factors (ineff) are computed by (4.1). Standard deviations are given in parentheses.

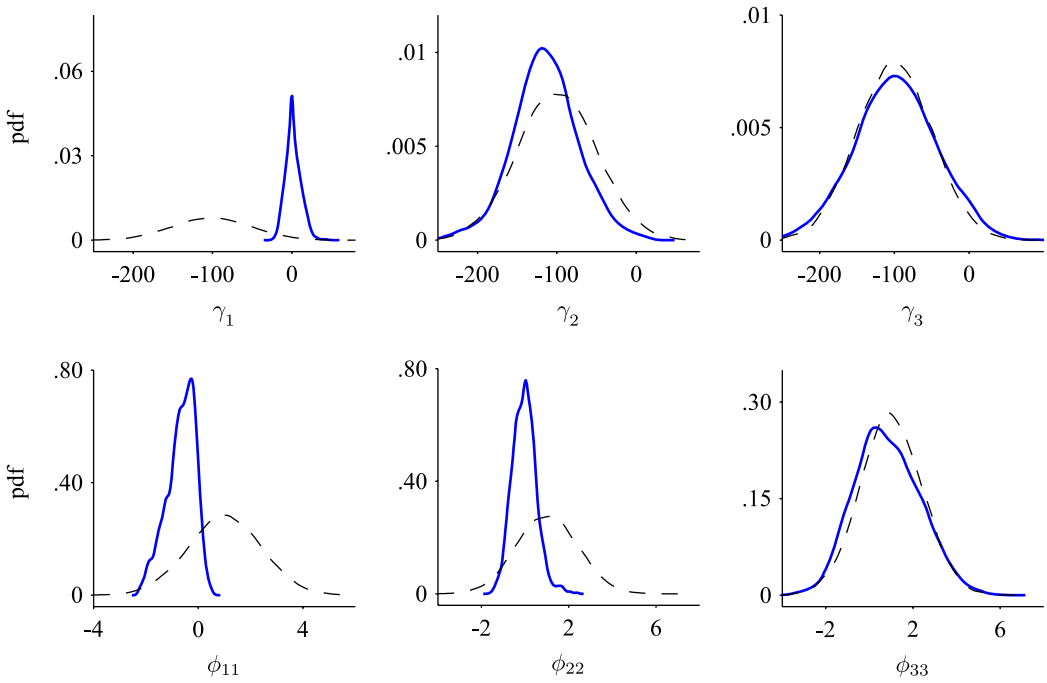

Figure 6. Prior-posterior updates of selected risk premia parameters.

including additional latent factors in the model. This extension is the subject of ongoing work.

5. CONCLUSION

We have presented a new approach for the fitting of affine yield curve models with macroeconomic factors. Although our discussion, like that of Ang, Dong, and Piazzesi (2007), is from a Bayesian viewpoint, our implementation of this viewpoint is different. We have emphasized the use of a prior on the parameters of the model which implies an upward-sloping yield curve. We believe that a prior distribution motivated and justified in this way is important for this complicated problem, because it concentrates attention on regions of the parameter space that otherwise might be missed, and also because it tends to support beliefs about which there can be consensus. Thus we feel that this sort of prior should be generally valuable.

We have also emphasized some technical developments in the simulation of the posterior distribution by tuned MCMC methods. The simulated annealing method that we have used for this purpose should have broad appeal. In addition, the square root filtering method for calculating the likelihood function, whenever the standard Kalman recursions become unstable, has relevance beyond the present problem.

In sum, our analysis demonstrates that the Bayesian viewpoint can be efficiently implemented in these models. In fact, it should be possible to apply our approach to other affine models, such as those with additional latent factors. Another interesting extension would be to affine models that allow for the possibility that the evolution equation of the factors, and the process of the market price of factor risks, can change at one or more time points. The comparison of such change point affine models with a no-change point model of the type analyzed

Table 4. Estimated covariance matrixes of the L1M2 model

Parameters	Prior			Posterior			Average acc. rate	Average ineff.
$\boldsymbol{\Omega}$	1			1			84.7	34.2
	0	0.545		0	0.195			
		(0.816)			(0.019)			
	0	0.000	0.536	0	−0.003	0.091		
		(0.396)	(0.533)		(0.009)	(0.009)		
$\sigma_1^{2*} \cdots \sigma_3^{2*}$	5.00	5.00	5.00	2.87	1.14	3.55	94	12.1
	(64.0)	(64.0)	(64.0)	(0.27)	(0.11)	(0.33)		
$\sigma_4^{2*} \cdots \sigma_6^{2*}$	5.00	5.00	5.00	1.45	3.75	2.98		
	(64.0)	(64.0)	(64.0)	(0.63)	(0.36)	(0.91)		
$\sigma_7^{2*} \cdots \sigma_9^{2*}$	5.00	5.00	5.00	2.90	4.01	5.06		
	(64.0)	(64.0)	(64.0)	(0.27)	(0.37)	(0.47)		

NOTE: Acceptance rates (acc. rate) are given as percentages. Inefficiency factors (ineff.) are computed by (4.1) according to the identification scheme $\omega_{11} = 1$. Standard deviations are given in parentheses.

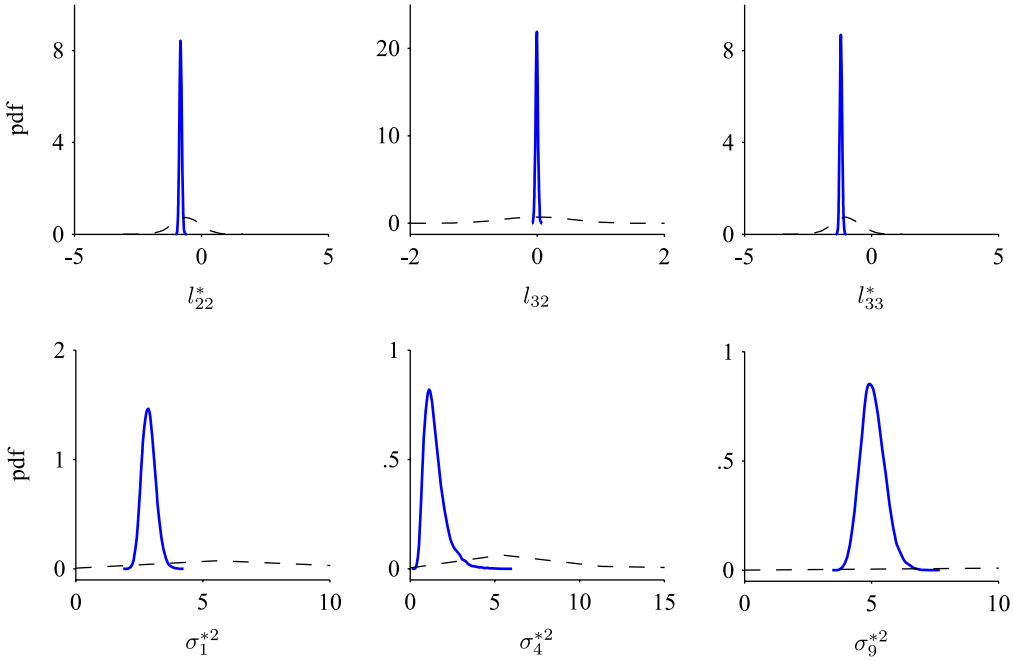

Figure 7. Prior-posterior updates of selected parameters from Table 4.

here, through marginal likelihoods and Bayes factors and out-of-sample predictive performance, obviously is of considerable interest. We have started work on these models, using MCMC estimation algorithms adapted from the present work and using the method of Chib (1995), and its MH variant in Chib and Jeliazkov (2001), to calculate the marginal likelihoods. We will report on the details elsewhere.

APPENDIX A: AN EXAMPLE DEMONSTRATING THE POSSIBILITY OF SUDDEN LARGE CHANGES IN THE LIKELIHOOD

Consider the arbitrage-free model with the following parameters:

$$\mathbf{G} = \text{diag}(0.93, 0.93, 0.93),$$

$$\boldsymbol{\mu} = (0, 75, 4),$$

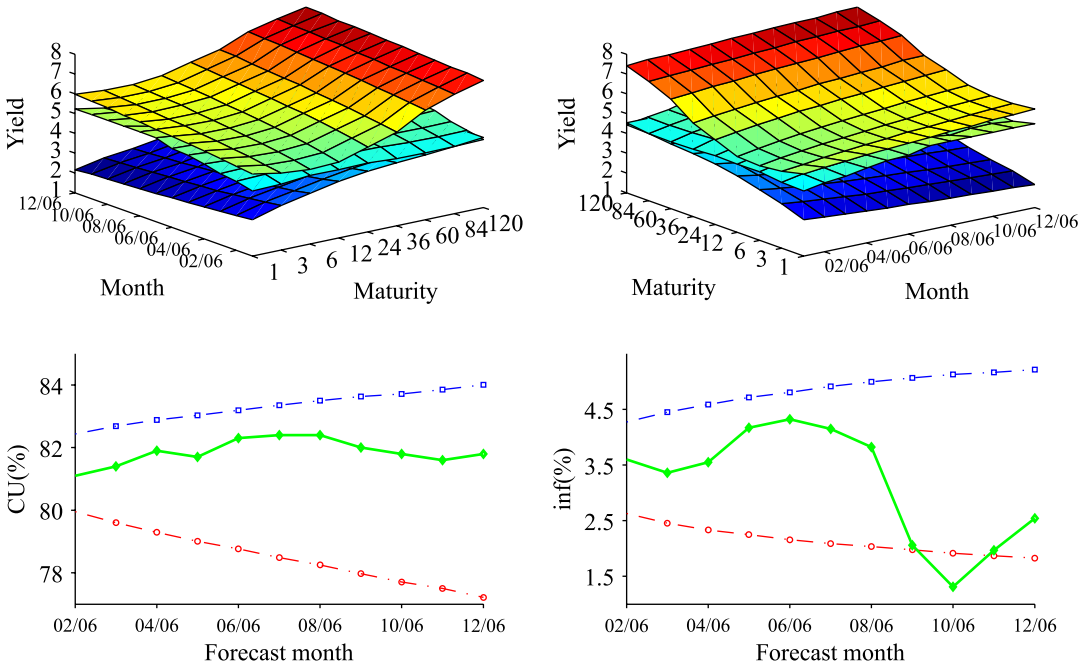

Figure 8. Out-of-sample (January–December 2006) forecasts of the yield curve and macroeconomic factors by the L1M2 model. The figure presents 12-month-ahead forecasts of the yields on the Treasury securities (three-dimensional graphs) and the macro factors (two-dimensional graphs). In each case, 5% and 95% quantile surfaces (curves), labeled "Low" and "High" are based on 25,000 draws. The observed surface and curves are labeled "Real." The top two graphs present two different views of the same yield forecasts.

$$\boldsymbol{\delta} = (-3, 0.2, 0.1, 0.5),$$

$$\boldsymbol{\gamma} = (-100, -100, -100),$$

$$\boldsymbol{\Phi} = \operatorname{diag}(1, 1, 1),$$

$$\boldsymbol{\Omega} = \operatorname{diag}(1, 0.30, 0.13),$$

$$\boldsymbol{\Sigma}^{*} = \operatorname{diag}(5, 5, 5, 5, 5, 5, 5, 5, 5, 5).$$

A simulation exercise shows that these parameter values generate plausible dynamics of the yields and macroeconomic variables. From the no-arbitrage condition (2.7)–(2.8) we find that the average of the highest maturity annual percentage yield, $\bar{a}_{120} + \bar{b}_{120} \times \mu$, equals 8.45. This value is comparable with historically observed average yields. The logarithm of the likelihood at the foregoing point is about -2.6×10^4. Now consider the following change in the value of δ_{22} from 0.2 to 0 with all other parameter values as before. Under this solitary change in the 39-dimensional parameter space, $\bar{a}_{120} + \bar{b}_{120} \times \mu$ is now 0.79. This large change in the factor loadings produces a similarly large change in the likelihood value. The new value of the logarithm of the likelihood is about -1.0×10^5. If we also change the parameter $\boldsymbol{\Phi}$ so that $(\mathbf{G} - \mathbf{L}\mathbf{H}^{-1}\boldsymbol{\Phi})$ does not imply stationarity, then the change is even larger. To see this, suppose that $\Phi_{11} = -12$ with all other parameters as before. In that case, $\bar{a}_{120} + \bar{b}_{120} \times \mu = 168.9$. Now the logarithm of the likelihood is about -1.7×10^5.

APPENDIX B: SQUARE ROOT FILTERING

If $\mathbf{M}$ is a nonnegative definite symmetric matrix, then a square root of $\mathbf{M}$ is a matrix $\mathbf{N}$ such that $\mathbf{M} = \mathbf{N}\mathbf{N}'$. Following the convention, we also use the notation $\mathbf{M}^{1/2}$ to denote an arbitrary square root of $\mathbf{M}$. Let $\mathbf{S}_{t|t}$ and $\mathbf{S}_{t|t-1}$ denote square roots of $\mathbf{R}_{t|t} = \mathbb{V}(\mathbf{f}_t|\mathbf{Y}_t, \boldsymbol{\psi})$ and $\mathbf{R}_{t|t-1} = \mathbb{V}(\mathbf{f}_t|\mathbf{Y}_{t-1}, \boldsymbol{\psi})$. In a square root filter, the update equations are expressed in terms of $\mathbf{S}_{t|t}$ and $\mathbf{S}_{t|t-1}$. Square root filters have at least two important advantages. First, both $\mathbf{R}_{t|t}$ and $\mathbf{R}_{t|t-1}$ are always nonnegative definite. Second, the numerical conditioning of $\mathbf{S}_{t|t}$ ($\mathbf{S}_{t|t-1}$) is much better then that of $\mathbf{R}_{t|t}$ ($\mathbf{R}_{t|t-1}$), because the condition number of the latter is the square of the condition number of the former. Now we turn to the description of the square root covariance filter used in this article.

The time update of the square root covariance matrix from $\mathbf{S}_{t-1|t-1}$ to $\mathbf{S}_{t|t-1}$ is based on the following matrix equation:

$$\begin{pmatrix} \mathbf{S}'_{t|t-1} \\ \mathbf{0}_{(k+m)\times(k+m)} \end{pmatrix} = \mathbf{Q} \begin{pmatrix} \mathbf{S}'_{t-1|t-1}\mathbf{G}' \\ \boldsymbol{\Omega}^{1/2\prime} \end{pmatrix}, \tag{B.1}$$

where $\mathbf{Q}$ is an orthogonal matrix that makes $\mathbf{S}_{t|t-1}$ upper triangular. Equation (B.1) shows how one can compute $\mathbf{S}_{t|t-1}$ given $\mathbf{S}_{t-1|t-1}$, $\mathbf{G}$ and $\boldsymbol{\Omega}$. This is a standard procedure that can be done by, for example, the Householder or Givens transformation. For instance, from the Householder transformation, one can create a simple function, $[\mathbf{D}, \mathbf{Q}] = H(\mathbf{C})$, that takes

$$\mathbf{C} = \begin{pmatrix} \mathbf{S}'_{t-1|t-1}\mathbf{G}' \\ \boldsymbol{\Omega}^{1/2\prime} \end{pmatrix}$$

as its input and returns $\mathbf{Q}$ and

$$\mathbf{D} = \begin{pmatrix} \mathbf{S}'_{t|t-1} \\ \mathbf{0} \end{pmatrix}.$$

Matrix $\mathbf{Q}$ is not important for our purposes. What is important is $\mathbf{D}$, which contains $\mathbf{S}_{t|t-1}$. The measurement update from $\mathbf{S}_{t|t-1}$ to $\mathbf{S}_{t|t}$ is based on the following equation:

$$\begin{pmatrix} (\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2\prime} & \bar{\mathbf{K}}'_t \\ \mathbf{0}_{(k+m)\times(p+m)} & \mathbf{S}_{t|t} \end{pmatrix} = \bar{\mathbf{Q}} \begin{pmatrix} (\mathbf{T}\boldsymbol{\Sigma}\mathbf{T})^{1/2\prime} & \mathbf{0}_{(p+m)\times(k+m)} \\ \mathbf{S}'_{t|t-1}\mathbf{B}' & \mathbf{S}'_{t|t-1} \end{pmatrix}, \tag{B.2}$$

where the matrix $\bar{\mathbf{Q}}$ is orthogonal. Denote the left side of (B.2) by $\mathbf{F}$. Given $\boldsymbol{\Sigma}, \mathbf{B}$ and $\mathbf{S}_{t|t-1}$, the three nonzero submatrixes of $\mathbf{F}$ can be found similarly to $\mathbf{S}_{t|t-1}$ from (B.1). For example, one can create another simple function, $[\mathbf{F}, \bar{\mathbf{Q}}] = G(\mathbf{E})$, that takes the second matrix on the right side of (B.2) as an input and, using the Givens transformation, returns $\mathbf{F}$ and $\bar{\mathbf{Q}}$. What is important for us in this transformation is the three nonzero submatrixes of $\mathbf{F}$, namely $(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2}$, $\bar{\mathbf{K}}_t$, and $\mathbf{S}_{t|t}$, which are used in the measurement update. Note that in (B.2) notation $(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2}$ is used only to emphasize the fact that the upper left corner of $\mathbf{F}$ equals this expression. Now we are ready to present the square root covariance filter.

Algorithm. Calculating the likelihood via the square root filter:

Step 1. Set $t = 1$ and initialize $\mathbf{f}_{0|0}, \mathbf{S}_{0|0}$.
Step 2. Calculate while $t \le n$:
  (a) Time update:
    (i) Update factor forecast mean $\mathbf{f}_{t|t-1} = \mathbf{G}\mathbf{f}_{t-1|t-1}$.
    (ii) Given $\mathbf{S}_{t-1|t-1}$, $\mathbf{G}$ and, $\boldsymbol{\Omega}$, compute $\mathbf{S}_{t|t-1}$ from (B.1).
  (b) Measurement update:
    (i) Given $\boldsymbol{\Sigma}, \mathbf{B}$ and $\mathbf{S}_{t|t-1}$, find $(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2}$, $\bar{\mathbf{K}}_t$ and $\mathbf{S}_{t|t}$ from (B.2).
    (ii) Compute $\mathbf{R}_{t|t-1} = (\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2}(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}' + \mathbf{B}\mathbf{R}_{t|t-1}\mathbf{B}')^{1/2\prime}$.
  (c) Compute the tth summand and the cumulative sum on the right side of (3.1).
Step 3. Increment t to $t + 1$ and go back to step 2.
Step 4. Return $\log p(\mathbf{y}|\boldsymbol{\psi})$.

[Received January 2008. Revised January 2009.]

REFERENCES

Anderson, B. D. O., and Moore, J. B. (1979), *Optimal Filtering*, Englewood Cliffs: Prentice-Hall.

Ang, A., and Piazzesi, M. (2003), "A No-Arbitrage Vector Autoregression of Term Structure Dynamics With Macroeconomic and Latent Variables," *Journal of Monetary Economics*, 50, 745–787.

Ang, A., Dong, S., and Piazzesi, M. (2007), "No-Arbitrage Taylor Rules," working paper, Columbia University.

Brooks, S., and Morgan, B. (1995), "Optimization Using Simulated Annealing," *Statistician*, 44, 241–257.

Chib, S. (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

——— (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Vol. 5, eds. J. J. Heckman and E. Leamer, Amsterdam: North-Holland, pp. 3569–3649.

Chib, S., and Greenberg, E. (1994), "Bayes Inference in Regression Models With ARMA (p, q) Errors," *Journal of Econometrics*, 64, 183–206.

——— (1995), "Understanding the Metropolis–Hastings Algorithm," *The American Statistician*, 49, 327–335.

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood From the Metropolis–Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Chib, S., Nardari, F., and Shephard, N. (2006), "Analysis of High Dimensional Multivariate Stochastic Volatility Models," *Journal of Econometrics*, 134, 341–371.

Dai, Q., and Singleton, K. (2000), "Specification Analysis of Affine Term Structure Models," *Journal of Finance*, 55, 1943–1978.

——— (2003), "Term Structure Dynamics in Theory and Reality," *Review of Financial Studies*, 16, 631–678.

Duffee, G. R. (1996), "Idiosyncratic Variation of Treasury Bill Yields," *Journal of Finance*, 51, 527–552.

Duffie, D., and Kan, R. (1996), "A Yield-Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Givens, G., and Hoeting, J. (2005), *Computational Statistics*, New York: Wiley.

Grewal, M. S., and Andrews, A. P. (2001), *Kalman Filtering: Theory and Practice Using Matlab* (2nd ed.), Hoboken, NJ: Wiley.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

Kim, S., Shephard, N., and Chib, S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison With ARCH Models," *Review of Economic Studies*, 65, 361–393.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. (1983), "Optimization by Simulated Annealing," *Science*, 220 (4598), 671–680.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

Piazzesi, M. (2003), "Affine Term Structure Models," working paper, Anderson School, UCLA.

Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer-Verlag.