Analysis of Multi-Factor Affine Yield Curve Models

SIDDHARTHA CHIB†
Washington University in St. Louis

BAKHODIR ERMASHEV‡
The Federal Reserve Bank of Richmond

January 2008; January 2009

Abstract
In finance and economics, there is a great deal of work on the theoretical modeling and statistical estimation of the yield curve (defined as the relation between $-\frac{1}{\tau} \log p_t(\tau)$ and $\tau$, where $p_t(\tau)$ is the time $t$ price of the zero-coupon bond with payoff 1 at maturity date $t+\tau$). Of much 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 implied yields are an affine function of the factors. The model is particularly interesting from a statistical perspective because the parameters in the model of the yields are complicated non-linear functions of the underlying parameters (for example those that appear in the evolution dynamics of the factors and those that appear in the model of the factor risks). This non-linearity tends to produce a likelihood function that is multi-modal. In this paper we revisit the question of how such models should be fit. Our discussion, like that of Ang et al. (2007), is from the Bayesian MCMC viewpoint, but our implementation of this viewpoint is different. Key aspects of the inferential framework include (i) 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; (ii) 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 through 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 (iii) measures to mitigate numerical instabilities in the fitting through reparameterizations and

*The views in this paper 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. In addition, the authors thank the editor, the referees, and Kyu Ho Kang and Srikanth Ramamurthy, for their insightful and constructive comments on previous versions of the paper.

†Address for correspondence: Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130. E-mail: chib@wustl.edu.

‡Address for correspondence: Charlotte Office, FRB of Richmond, PO Box 30248, Charlotte, NC 28230. E-mail: bakhodir.ergashev@rich.frb.org
square root filtering recursions. We apply the techniques to explain the monthly yields on nine US Treasuries (with maturities ranging from 1 to 120 months) over the period January 1986 to 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.

**Keywords:** Term structure; Yield curve; No-arbitrage condition; Markov chain Monte Carlo; Simulated annealing; Square-root filter; Forecasting.

1 **Introduction**

In finance and economics, a great deal of attention is devoted to understanding the pricing of default-free zero coupon bonds (bonds such as the T-bills issued by the U.S. Treasury that have no risk of default and which provide a single payment - typically normalized to one - at a date in the future when the bond matures, and are sold prior to the maturity date at a discount from the face value of one). 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. If one lets $\tau$ 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 $z_{t\tau}$ of the bond is essentially equal to $-\frac{1}{\tau} \log p_t(\tau)$. Of crucial interest in this context is the so-called yield curve which is the set of yields that differ only in their time to maturity $\tau$. This yield curve is generally plotted with the yields to maturity $z_{t\tau}$ against the time to maturity $\tau$ 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 several different ways, 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), it has become popular in the last ten years to strike a middle ground, by building models that have a statistical orientation, and hence are flexible and have the potential of fitting the data well, and at the same time connected to economics through the enforcement of a no-arbitrage condition on bond prices. The 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.

A class of models with the foregoing features that has attracted the most attention are multi-factor affine yield curve models. This class of models was introduced in an important paper by Duffie and Kan (1996). The general modeling strategy is to try 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 $\gamma_t$ are then assumed to determine the so-called pricing kernel, or stochastic discount factor, $\kappa_{t,t+1}$. The market price of factor risks $\gamma_t$ are in turn modeled as an affine function of the factors. The no-arbitrage condition is enforced automatically by pricing the $\tau$ 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) show 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. Thus, on taking logs, and dividing by minus $\tau$, the yields become an affine function of the factors.

The Duffie and Kan framework provides a versatile approach for modeling the yield curve. Ang and Piazzesi (2003) enhance its practical value by incorporating macro-economic variables in their list of factors that drive the dynamics of the model. In particular, one of their factors is taken to be latent and two are taken to be observed macro-economic variables - we refer to this model as the L1M2 model. A version of this model is systematically examined by Ang, Dong and Piazzesi (2007) (ADP henceforth). A convenient statistical aspect of this multi-factor 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 consisting of the set of yields derived from the pricing model. What makes this model particularly interesting from a statistical perspective is that the parameters in the observation equation are highly non-linear functions of the underlying deep parameters of the model (for example the parameters that appear in the evolution dynamics of the factors and those
that appear in the model of $\gamma_t$. This non-linearity is quite severe and produces a likelihood function that can be multi-modal as we show below.

To deal with the estimation challenges, ADP (2007) adopt 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 the estimation of the parameters that are otherwise ill-determined. However, ADP (2007) in their work employ diffuse priors and therefore do not fully exploit this aspect of the Bayesian approach. Another reason for pursuing the Bayesian approach is that it focuses on summaries of the posterior distribution, such as the posterior expectations and posterior credibility intervals of parameters, which can be easier to interpret than the (local) mode of an irregular likelihood function. ADP (2007) demonstrate the value of the Bayesian approach by estimating the L1M2 model on quarterly data and yields of maturities up to 20 quarters. They employ 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, ADP (2007) in their study concentrate on the finance implications of the fitting and do not discuss how well the MCMC approach that they use actually performs in terms of metrics that are common in the Bayesian literature. For instance, they do not provide inefficiency factors and other related measures which can be useful in evaluating the efficiency of the MCMC sampling (Chib (2001), Liu (2001), Robert and Casella (2004)).

In this paper we continue the Bayesian study of the L1M2 multi-factor affine yield curve model. Our contributions deal with several inter-related 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 the estimation of the model since it tends to smooth out and diminish the importance of regions of the parameter space that are a priori uninteresting. 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 first grouped into blocks and then each block is sampled in turn within each sweep of the MCMC algorithm.
with the help of the Metropolis-Hastings algorithm whose proposal densities are constructed by tailoring to the conditional posterior distribution of that block, along the lines of Chib and Greenberg (1994). A noteworthy aspect of this tailoring is that the modal values are found by the method of simulated annealing in order to account for the potentially multi-modal nature of the posterior surface. 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 paper is organized as follows. Section 2 introduces the arbitrage-free model, the identification restrictions and the data that is used in the empirical analysis. In Section 3 we present the state space form of the model, the likelihood function and the prior distribution. We then discuss how the resulting posterior distribution is summarized by MCMC methods. In Section 4 we present results from our analysis of the L1M2 model. We summarize our conclusions in Section 5. Details, for example those related to the instability of the coefficients in the state space model to changes in the parameter values, and the square root filtering method, are presented in appendices at the end.

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 one at (time to) maturity $\tau$ (say measured in 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 $(t + 1)$, satisfy the conditions

$$p_t(\tau) = \mathbb{E}_t[\kappa_{t+1}p_{t+1}(\tau - 1)], \quad t = 1, 2, \ldots, n, \quad \tau = 1, 2, \ldots, \tau^*, \quad (2.1)$$
where $E_t$ is the expectation conditioned on time $t$ information and $\kappa_{t+1} > 0$ is the so-called stochastic discount factor (pricing kernel). The goal is to model the yields

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

for each time $t$ and each maturity $\tau$.

Now let $u_t$ be a latent variable, $m_t = (m_1t, m_2t)'$ a 2-vector of observed macroeconomic variables, and $f_t = (u_t, m_t)'$ the stacked vector of latent and observed factors. In the affine model it is assumed that these factors follow the vector Markov process:

$$
\begin{pmatrix}
\begin{bmatrix} u_t \\ m_t \end{bmatrix} \\
\mu
\end{pmatrix} - 
\begin{pmatrix}
\begin{bmatrix} \mu_u \\ \mu_m \end{bmatrix} \\
\mu
\end{pmatrix} = 
\begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix} 
\begin{pmatrix}
\begin{bmatrix} u_{t-1} \\ m_{t-1} \end{bmatrix} \\
\mu
\end{pmatrix} + 
\begin{pmatrix}
\eta_{ut} \\
\eta_{mt}
\end{pmatrix},
$$

where $G$ is a matrix with eigenvalues less than one and

$$\eta_t | \Omega \sim iid \mathcal{N}_{k+m}(0, \Omega), \quad \text{and} \quad \Omega = 
\begin{pmatrix}
\Omega_{11} & \Omega_{12} \\
\Omega_{12}' & \Omega_{22}
\end{pmatrix}$$

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

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 SDF is given by

$$\kappa_{t,t+1} = \exp\left\{-\delta_1 - \delta_2 f_t - \frac{1}{2} \gamma_t' \gamma_t - \gamma_t' L^{-1} \eta_{t+1}\right\},$$

where $\delta_1$ and $\delta_2$ are constants, $L$ is a lower triangular matrix such that $LL' = \Omega$, and $\gamma_t$ is a vector of time-varying market prices of factor risks that is assumed to be an affine function of the factors

$$\gamma_t = \gamma + \Phi f_t.$$  

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

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

$$p_t(\tau) = \exp\left\{-a_\tau - b_\tau f_t\right\}$$

6
where $a_\tau$ and $b_\tau$ are obtained from the following set of vector difference equations

$$a_{j+1} = a_j + b_j'\{(I - G)\mu - L\gamma\} - \frac{1}{2} b_j'\Omega b_j + \delta_1,$$  \hfill (2.5)

$$b_{j+1} = (G - L\Phi)'b_j + \delta_2; \quad j = 1, 2, \ldots, \tau, \quad \tau = 1, 2, \ldots, \tau^* \hfill (2.6)$$

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

$$a_{j+1} = a_j + b_j'\{(I - G)\mu - LH^{-1}\gamma\} - \frac{1}{2} b_j'\Omega b_j/1200 + \delta_1,$$  \hfill (2.7)

$$b_{j+1} = (G - LH^{-1}\Phi)'b_j + \delta_2; \quad j = 1, 2, \ldots, \tau, \quad \tau = 1, 2, \ldots, \tau^* \hfill (2.8)$$

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

$$H = \text{diag}(100, 100, 1200)$$

and it arises from a similar conversion applied to the factors. In particular, because one of the macroeconomic factors that we specify below (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 we 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.

We underline the fact that $a_\tau$ and $b_\tau$ are highly nonlinear functions of the unknown parameters of the factor evolution and SDF specifications. It is this complicated dependence on the parameters that causes difficulties in the analysis of this model.

If we now assume that each yield is subject to measurement or pricing error, the theoretical model of the object of interest (the yield curve) for each time $t$ can be expressed as

$$z_{t\tau} = \frac{1}{\tau} a_\tau + \frac{1}{\tau} b_\tau'f_t + \varepsilon_{t\tau}, \quad t = 1, 2, \ldots, n, \quad \tau = 1, 2, \ldots, \tau^* \hfill (2.9)$$

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

$$z_{t1} = \delta_1 + \delta_2 f_t + \varepsilon_{t1} \hfill (2.10)$$

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

7
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. For identification purposes we therefore impose some restrictions on the parameters in the model. Following Dai and Singleton (2000) we assume that $G_{11}$ is positive, that the first element of $\delta_2$ (the one corresponding to the latent factor) is positive, that $\mu_u$ is zero, and that $\Omega_{11}$ is one. Although it is not strictly necessary, we further assume that $\Omega_{12}$ is the zero 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 $G$ are less than one in absolute value. This constraint is the stationarity restriction on the factor evolution process. We also impose a similar eigenvalue restriction on the matrix

$$G - LH^{-1}\Phi$$

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

2.2 Empirical state space formulation

A useful feature of affine models (for the purpose of statistical analysis) is that it can be cast in linear state space form, consisting of the measurement equations for the yields, and the evolution equations of the factors. To do this we first need to fix the maturities of interest. The model in (2.9) delivers the yield for any maturity from $\tau = 1$ to $\tau = \tau^*$. Suppose that interest centers on the maturities in the set $A = \{\tau_1, \tau_2, ..., \tau_p\}$ 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 $z_t = (z_{t1}, \ldots, z_{tp})'$ where $z_{ti} \equiv z_{t\tau_i}$ with $\tau_i \in A$, $i = 1, 2, ..., p$.

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

$$ z_t = \bar{a} + \bar{B}f_t + \varepsilon_t, \quad \varepsilon_t | \Sigma \sim \mathcal{N}_p(0, \Sigma), \quad t = 1, 2, \ldots, n, $$

where \( \Sigma \) is diagonal with unknown elements given by \((\sigma^2_1, \ldots, \sigma^2_p)\). It is important to bear in mind that \( \bar{a} \) and \( \bar{B} \) must be recalculated for every new value of the parameters.

Because the factors in this case contain some observed components (namely \( m_t \)), we have to ensure that these are inferred without error. An economical way to achieve this is by defining the outcome as

$$ y_t = \begin{pmatrix} z_t \\ m_t \end{pmatrix}, $$

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

$$ \begin{pmatrix} z_t \\ m_t \end{pmatrix} = \begin{pmatrix} \bar{a} \\ 0_{3 \times 1} \end{pmatrix} + \begin{pmatrix} \bar{B} \\ J_{2 \times 3} \end{pmatrix} f_t + \begin{pmatrix} I_p \\ 0_{2 \times p} \end{pmatrix} \varepsilon_t, \quad (2.11) $$

where \( J = (0_{2 \times 1}, I_2) : 2 \times 3 \). The state space model is completed by the set of evolution equations which are given in (2.2).

We conclude by noting that in practice we parameterize the factors in terms of deviations from \( \mu \) as

$$ \tilde{f}_t = (f_t - \mu), $$

in which case the model of interest becomes

$$ y_t = a + B(\tilde{f}_t + \mu) + T\varepsilon_t, \quad (2.12) $$

$$ \tilde{f}_t = G\tilde{f}_{t-1} + \eta_t, t \leq n, \quad (2.13) $$

and where, at \( t = 0 \), \( \tilde{f}_0 = (u_0, m_0 - \mu) \). The parameter \( \mu \) is thus present in \( \tilde{f}_0 \). It is natural now to assume that \( m_0 \) is known from the data and that \( u_0 \), independently of \( m_0 \), follows the stationary distribution

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

where \( V_u = (1 - G^2_{11})^{-1} \).

This is the model that we study in this paper.
2.3 Data

The term structure data that is 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 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 utilize this time span because monetary policy in this period was relatively stable.

![Figure 1: Term Structure of the US treasury interest rates and macroeconomic variables. The data covers the period between January 1986 and December 2006. The yields data consists 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. This data is presented in the top two graphs in the form of three 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.](image)

The model is estimated on data until December 2005. The last 12 months of the sample is used for prediction and validation purposes. Our proxy for the month one yield is the
Federal funds rate (FFR), as suggested by Duffee (1996) and Piazzesi (2003), among others. It should be noted that Treasury bonds of over one year pay semiannual coupon payments while Treasury bills (of maturities of one year or less) do not pay any coupons. We extract the implied zero-coupon yield curves by the interpolation method that is used by the US 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, for example, Ang and Piazzesi (2003). These data are taken from the Federal Reserve Bank of St. Louis’ FRED II database.

We provide a graphical view of our data in Figure 1. The top panel has the time series plots of the yields in three and two dimensions and the bottom panel has the time series plots of our macroeconomic factors. Table 1 contains a descriptive summary of these data.

<table>
<thead>
<tr>
<th>Macro variables</th>
<th>CU</th>
<th>Infl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample average (%)</td>
<td>80.92</td>
<td>3.04</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>(2.82)</td>
<td>(1.11)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bond Maturity (month)</th>
<th>Average yield (%)</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.96</td>
<td>(2.17)</td>
</tr>
<tr>
<td>3</td>
<td>4.71</td>
<td>(1.97)</td>
</tr>
<tr>
<td>6</td>
<td>4.89</td>
<td>(2.00)</td>
</tr>
<tr>
<td>12</td>
<td>5.06</td>
<td>(1.99)</td>
</tr>
<tr>
<td>24</td>
<td>5.39</td>
<td>(2.01)</td>
</tr>
<tr>
<td>36</td>
<td>5.61</td>
<td>(1.93)</td>
</tr>
<tr>
<td>60</td>
<td>5.96</td>
<td>(1.77)</td>
</tr>
<tr>
<td>84</td>
<td>6.22</td>
<td>(1.68)</td>
</tr>
<tr>
<td>120</td>
<td>6.41</td>
<td>(1.63)</td>
</tr>
</tbody>
</table>

Table 1: Descriptive statistics for the macro factors and the yields. This table presents the descriptive statistics for the macro factors, the short rate (approximated by the Federal funds rate) that corresponds to the yield on 1 month and eight yields on the constant maturity Treasury securities for the period of January, 1986 - December 2006. The macro factors are the Manufacturing capacity utilization (CU) and inflation (Infl). Inflation is measured by the Consumer Price Index. Source: Federal Reserve Bank of St. Louis FRED II database.

3 Prior-posterior analysis

3.1 Preliminaries

In doing inference about the unknown parameters it is helpful (both for specifying the prior distribution and for conducting the subsequent MCMC simulations) to group the unknowns
into separate blocks. To begin, we let
\[ \theta_1 = (g_{11}, g_{22}, g_{33})' \] and
\[ \theta_2 = (g_{12}, g_{13}, g_{21}, g_{31}, g_{23}, g_{32})' \]
Thus, \( \theta_1 \) consists of the diagonal elements of \( G \), since these are likely to be large, and \( \theta_2 \) the remaining elements of \( G \), since those that are likely to be smaller. We also let
\[ \theta_3 = (\phi_{11}, \phi_{22}, \phi_{23}, \phi_{32}, \phi_{33})' \] and
\[ \theta_4 = (\phi_{12}, \phi_{13}, \phi_{21}, \phi_{31})' \]
for the elements of \( \Phi \). Next we express \( \Omega \) as \( LL' \) and collect the three free elements of the lower-triangular \( L \) as
\[ \theta_5 = (l_{22}^*, l_{32}, l_{33}^*) \]
where \( l_{22} = \exp(l_{22}^*) \) and \( l_{33} = \exp(l_{33}^*) \), so that any value of \( \theta_5 \) leads to a positive definite \( \Omega \) in which \( \Omega_{12} \) is zero. Also, we let
\[ \theta_6 = \delta \] and
\[ \theta_7 = (\mu, \gamma) \]
Finally, because the elements \( \sigma_i^2 \) of the matrix \( \Sigma \) are liable to be small, and to have a U-shape with relatively larger values at the low and high maturity ends, we reparametrize the variances and let
\[ \theta_8 = (\sigma_1^{2*}, ..., \sigma_p^{2*}) \]
where \( \sigma_i^{2*} = d_i \sigma_i^2 \) and \( d_1 = d_2 = d_7 = d_8 = 10; d_3 = d_5 = d_6 = 100, \) and \( d_4 = 2000 \). The choice of these \( d_i \)'s is not particularly important. What is important is that we do inferences about \( \sigma_i^2 \) indirectly (through the much larger \( \sigma_i^{2*} \)). These transformations of the variances are introduced primarily 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 \( \psi = (\theta, u_0) \), where \( \theta = \{\theta_i\}_{i=1}^8 \). In a model with \( p = 9 \) yields, the dimension of each block in \( \psi \) is 5, 4,
5, 4, 2, 5, 5, 9, and 1, respectively. In addition, the parameters $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5,$ and $\theta_6$ are constrained to lie in the set $S = S1 \cap S2 \cap S3$ where $S1 = \{\theta_1, \theta_2 : \text{abs(eig}(G)) < 1\}$, $S2 = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5 : \text{abs(eig}(G - LH^{-1}\Phi)) < 1\}$ and $S3 = \{\theta_6 : \delta_{2u} \in R_+\}$.

Now if we let $y = (y_1, ..., y_n)$ denote the data, then the density of $y$ given $\psi$ may be written as

$$\log p(y|\psi) = -\frac{np}{2} \log(2\pi) - \sum_{t=1}^{n} \left[ \log(\det(R_{t|t-1})) + (y_t - a - B(f_{t|t-1} + \mu))' (R_{t|t-1})^{-1} (y_t - a - B(f_{t|t-1} + \mu)) \right],$$

where $f_{t|t-1} = \mathbb{E}(\hat{f}_t|Y_{t-1}, \psi)$ and $R_{t|t-1} = \mathbb{V}(y_t|Y_{t-1}, \psi)$ are the one step ahead forecast of the state and the conditional variance of $y_t$, respectively, given information $Y_{t-1} = (y_1, ..., y_{t-1})$ up to time $(t - 1)$. Generally, the latter quantities can be calculated by the Kalman filtering recursions (see for example, Harvey (1989)). In this model, however, for some parameter values, the recursions in (2.12)-(2.13) can produce values of $a_i$ and $b_i$ that are large (Appendix A exemplifies this possibility), and $R_{t|t-1}$ can become non-positive definite. In such cases, we invoke the square root filter (Grewal and Andrews (2001), Anderson and Moore (1979)). This filter tends to be more stable than the Kalman filter because the state covariance matrices are propagated in square root form. We present this filter in Appendix B in notation that corresponds to our model and with the inclusion of details that are missing in the just cited references.

Another issue is that the likelihood function can be multi-modal. We can see this problem by considering the posterior distribution under a flat prior. Sampled variates drawn from this posterior distribution can be summarized in one or two dimensions. Because the prior is flat, these distributions effectively reveal features of the underlying likelihood function. Although the technicalities are not important at this stage, we sample the latter posterior distribution by specializing the MCMC simulation procedure of the next section. Figure 2 contains the graphs of the 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. Finding the maximum of the likelihood is largely infeasible even with a stochastic optimization method such as simulated
Figure 2: Kernel smoothed likelihood surface plots for some pairs of parameters in the arbitrage-free model.

annealing. This is not surprising given the shape of the likelihood surface and the size of the parameter space.

We seek to avoid such problems from the Bayesian approach. The shift of focus to the posterior distribution, away from solely the likelihood, can be helpful provided the prior distribution is carefully formulated. If the prior distribution, for example, down weights regions of the parameter space that are not economically meaningful, the posterior distribution can be smoother and better behaved than the likelihood function. To see how this can happen we provide in Figure 3 the corresponding bivariate posterior densities from the prior we describe next. These bivariate posterior densities are considerably smoother and the effective support of the last two distributions has narrowed. This preamble to our analysis can be seen as the motivation for the Bayesian viewpoint in this problem.
Figure 3: Kernel smoothed posterior surface plots for some pairs of parameters in the arbitrage-free model.

### 3.2 Prior distribution

One useful way for developing a prior distribution on $\theta$ is to reason in terms of the yield curve that is implied by the prior on the parameters. Specifically, one can formulate a prior which implies that the yield curve is upward sloping on average. The latter is, of course, a reasonable a priori assumption to hold about the yield curve.

We arrive at such a prior as follows. We specify 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 see if the yield curve is upward sloping on average for each time period in the sample. If it is not we revise the prior distribution somewhat and repeat the process until we get an implied yield curve over time that we think is reasonable. It is important to note that this process of prior construction does not involve the observed data in any way at all.

- $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \text{and } \theta_6)$: We suppose that the joint distribution of these parameters
is proportional to

\[ \mathcal{N}(\theta_1, \theta_2|g_0, V_g)\mathcal{N}(\theta_3, \theta_4|\phi_0, V_\phi)\mathcal{N}(\theta_5|l_0, V_l)\mathcal{N}(\theta_6|\delta_0, V_\delta)I_S \]

For the hyperparameters, we let

\[ g_0 = (0.95, 0.95, 0.95, 0, 0, 0, 0, 0, 0) \quad \text{and} \quad V_g = \text{diag}(0.4, 0.4, 0.4, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20) \]

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 0.63 and the off-diagonal elements at zero with a standard deviation of 0.45. Given that \( G \) 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

\[ \phi_0 = (1, 1, 0, 0, 1, 0, 0, 0, 0) \quad \text{and} \quad V_\phi = 2I_9 \]

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

\[ l_0 = (-0.6, 0, -1) \quad \text{and} \quad V_l = 0.25 \times I_2 \]

as the mean and covariance of \( \theta_5 \), respectively. The standard deviation of each element is thus 0.5 which implies a relatively diffuse prior assumption on these parameters. Finally, based on the Taylor rule intuition that high values of capacity utilization and inflation should be associated with high values of the short rate, we let

\[ \delta_0 = (-3, 0.20, 0.10, 0.70)' \quad \text{and} \quad V_\delta = \text{diag}(1, 0.2, 0.1, 0.2). \]

- \( \theta_7 \): We suppose that the joint distribution of these parameters is given by

\[ \mathcal{N}(\mu|\mu_0, V_\mu)\mathcal{N}(\gamma|\gamma_0, V_\gamma) \]
where
\[ \mu_0 = (75, 4)' \quad \text{and} \quad V_{\mu} = \text{diag}(49, 25) \]
so that the prior mean of capacity utilization is assumed to be 75% and that of the inflation rate to be 4% (the prior standard deviations of 7 and 5 are sufficient to cover the most likely values of these rates) and where
\[ \gamma_0 = (-100, -100, -100)', \quad V_{\gamma} = \text{diag}(100, 100, 100). \]
The prior mean of \( \gamma \) is negative in order to imply an upward sloping average yield curve.

- \( \theta_8 \): We assume that
\[ \sigma_i^{2*} \sim IG\left(\frac{a_0}{2}, \frac{b_0}{2}\right), \quad i = 1, \ldots, p \]
where \( a_0 \) and \( b_0 \) are such as to imply that the a priori mean of \( \sigma_i^{2*} \) is 5 and the standard deviation is 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 \( \sigma_i^{2*} \) is not.

To show what these assumptions imply for the outcomes, we simulate the parameters 10,000 times from the prior, and for each drawing of the parameters, we 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 considerable 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). From our perspective, however, this is a necessary consequence of a reasonably well dispersed prior distribution on the parameters.

### 3.3 Posterior and MCMC sampling

Under our assumptions, the posterior distribution of \( \psi \) is
\[ \pi(\psi|y) \propto p(y|\psi)p(u_0|\theta)\pi(\theta) \]  \hspace{1cm} (3.2)
where $p(y|\psi)$ is given in (3.1), $p(u_0|\theta)$ from (2.14) is

$$N(0, V_u)$$

and $\pi(\theta)$ is proportional to

$$N(\theta_1, \theta_2|g_0, V_g)N(\theta_3, \theta_4|\phi_0, V_\phi)N(\theta_5|l_0, V_l)N(\theta_6|\delta_0, V_\delta)I_S \times N(\mu|\mu_0, V_\mu)N(\gamma|\gamma_0, V_\gamma) \prod_{i=1}^p IG(\sigma_i^2|\frac{a_0}{2}, \frac{b_0}{2})$$

This distribution is challenging to summarize even with MCMC methods because of the facts we have documented in the foregoing discussion. For one, we have to deal with the high dimension of the parameter space and the fact that $\theta_1$ and $\theta_2$ are concentrated at the boundary of the parameter space - here the stationarity region - and the fact that the market price of risk parameters are difficult to infer. Another is the nonlinearity of the model arising from the recursions that produce $\bar{a}$ and $\bar{B}$. As a result, as shown in Figures 2 and 3, the posterior distribution is typically multi-modal (but better behaved than the likelihood on
account 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 \((u_0, \sigma)\). In fact, conditioning on the factors, as in the approach of ADP (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. This is similar to the approach taken in Kim, Shephard, and Chib (1998), and Chib, Nardari, and Shephard (2006). In addition, we sample \(\{\theta_i\}_{i=1}^8\) 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 by the Metropolis-Hastings algorithm.

**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 \(\theta_i\) from \(\pi(\theta_i|y, \theta_{-i}, u_0)\), where \(\theta_i\) denotes the current parameters in \(\theta\) excluding \(\theta_i\)

**Step 3** Sample \(u_0\) from \(\pi(u_0|y, \theta)\)

**Step 4** Repeat Steps 2-3, discard the draws from the first \(n_0\) iterations and save the subsequent \(M\) draws \(\{\theta^{(n_0+1)}, \ldots, \theta^{(n_0+M)}\}\)

A key point is that the sampling in Steps 2 and 3 is done by a “tailored” M-H algorithm along the lines of Chib and Greenberg (1994) and Chib and Greenberg (1995). In brief, the idea is to build a proposal density that is similar to the target posterior density at the modal value. This is done by first finding the modal value of the current target density and the inverse of the negative Hessian of this density at the modal value. The proposal density is then based on these two quantities. This idea has proved useful in a range of problems. Its value from a theoretical perspective, however, still needs to be formalized.
For illustration, consider for instance block $\theta_i$ and its target density $\pi(\theta_i|y, \theta_{-i}, u_0)$. Suppose that the value of this block after the $(j-1)$st iteration is $\theta_i^{(j-1)}$. Now let

$$\hat{\theta}_i = \arg \max_{\theta_i} \log \pi(\theta_i|y, \theta_{-i}, u_0)$$

and

$$V_{\theta_i} = \left( -\frac{\partial^2 \log \pi(\theta_i|y, \theta_{-i}, u_0)}{\partial \theta_i \partial \theta'_i} \right)^{-1} \bigg|_{\theta_i = \hat{\theta}_i}$$

the mode and inverse of the negative Hessian at the mode, and let the proposal density $q(\theta_i|y, \theta_{-i}, u_0)$ be a multivariate-t distribution with location $\hat{\theta}_i$, dispersion $V_{\theta_i}$ and (say) 5 degrees of freedom:

$$q(\theta_i|y, \theta_{-i}, u_0) = St(\theta_i|\hat{\theta}_i, V_{\theta_i}, 5)$$

Now draw a proposal value

$$\theta_i^* \sim q(\theta_i|y, \theta_{-i}, u_0)$$

and set $\theta_i^{(j)} = \theta_i^{(j-1)}$ if the proposal does not satisfy the constraint $S$; otherwise, accept $\theta_i^*$ as the next value $\theta_i^{(j)}$ with probability given by

$$\alpha(\theta_i^{(j-1)}, \theta_i^*|y, \theta_{-i}, u_0) = \min \left\{ \frac{\pi(\theta_i^*|y, \theta_{-i}, u_0)}{\pi(\theta_i^{(j-1)}|y, \theta_{-i}, u_0)} \frac{St(\theta_i^{(j-1)}|\hat{\theta}_i, V_{\theta_i}, 15)}{St(\theta_i^*|\hat{\theta}_i, V_{\theta_i}, 15)}, 1 \right\},$$

or take $\theta_i^{(j)} = \theta_i^{(j-1)}$ with probability $1-\alpha(\theta_i^{(j-1)}, \theta_i^*|y, \theta_{-i}, u_0)$.

One point is that the modal value $\hat{\theta}_i$ cannot in general be found by a Newton or related hill-climbing method because of a tendency of these methods to get trapped in areas corresponding to local modes. A more effective search can be conducted with simulated annealing (SA) (for example, see Kirkpatrick et al. (1983), Brooks and Morgan (1995) or Givens and Hoeting (2005) for detailed information about this method and its many variants). We have found this method to be quite useful for our purposes and 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. Moves that lower the function value can be sometimes accepted. The probability of accepting such downhill moves declines over iterations according to an “cooling schedule,” thus allowing the method to converge. In our implementation, we first divide the search process into various stages, denoted by $k$, $k = 1, 2, \ldots, K$, with the length of each stage $l_k$ given by...
where $b$ is a positive integer. We then specify the initial temperature $T_0$ which is held constant in each stage but reduced across stages according to the linear cooling schedule $T_k = aT_{k-1}$, where $0 < a < 1$ is the cooling constant. Then, starting from an initial guess for the maximum, within each stage and across stages, repeated proposals are generated for a randomly chosen element from a random walk process with a Gaussian increment of variance $S$. 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 parameters in some preliminary runs striking a balance between the computational burden and the efficiency of the method. For our application, this tuning led to the choices $T_0 = 2$, $a = 0.5$, $K = 4$, $l_0 = 10$, $b = 10$ and $S = 0.1$. A point to note is that it was not necessary to tune the SA algorithm separately for each block. Another point is that the temperature parameter is reduced relatively quickly since it is enough in this context to locate 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 does in predicting the yields and macroeconomic factors out of sample. As is customary in the Bayesian context, we address this question by calculating the Bayesian predictive density. This is 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. The natural approach for summarizing this density is by the method of composition. For each drawing of the parameters from the MCMC algorithm, one draws the latent factors and the macroeconomic factors in the forecast period.
from the evolution equation of the factors, conditioned on \( \tilde{f}_n \); then given the factors and the parameters, one samples the yields from the observation density for each time period in the forecast sample. This sample of yields is a sample from the predictive density which can be summarized in the usual ways.

**Algorithm: Sampling the predictive density of the macroeconomic factors and yields**

**Step 1** For \( j = 1, 2, \ldots, M \)

(a) Compute \( \tilde{a}^{(j)} \) and \( \tilde{B}^{(j)} \) from the recursive equations (2.7)-(2.8), and the remaining matrices of the state-space model, given \( \theta^{(j)} \) and \( \tilde{f}_n \)

(b) For \( t = 1, 2, \ldots, T \)

(i) Compute \( \tilde{z}^{(j)}_{n+t} = G^{(j)} \tilde{z}^{(j)}_{n+t-1} + \eta^{(j)}_{n+t} \) where \( \eta^{(j)}_{n+t} \sim N_{k+m}(0, \Omega^{(j)}) \)

(ii) Compute \( z^{(j)}_{n+t} = \tilde{a}^{(j)} + \tilde{B}^{(j)}(\tilde{x}^{(j)}_{n+t} + \mu^{(j)}) + \varepsilon^{(j)}_{n+t} \), where \( \varepsilon^{(j)}_{n+t} \sim N_p(0, \text{diag}(\sigma^{(j)})) \)

(iii) Set \( y^{(j)}_{n+t} = \{ z^{(j)}_{n+t}, m^{(j)}_{n+t} \} \)

(c) Save \( y^{(j)}_f = \{ y^{(j)}_{n+1}, \ldots, y^{(j)}_{n+T} \} \).

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

The resulting collection of macroeconomic factors and yields, is a sample from the Bayesian predictive density. We summarize it in terms of its quantiles and moments.

**4 Results**

In this section we summarize our results. The results are based on \( M = 25000 \) iterations of our algorithm beyond a burn-in of \( n_0 = 5000 \) iterations. In addition to summaries of the posterior distribution we also report on the efficiency of our MCMC algorithm. For each of the M-H steps, we report the average values of the M-H acceptance rates and the corresponding inefficiency factors

\[
1 + 2 \sum_{k=1}^{N} \left( 1 - \frac{k}{N} \right) \rho(k)
\]  

(4.1)
where $\rho(k)$ is the autocorrelation at lag $k$ of the MCMC draws of that parameter and $N = 500$.

For the sake of contrast, we also compute the results (that we, however, do not report) from a random-walk Metropolis-Hastings (RW-MH) algorithm that uses the same blocking structure as our tailored algorithm, sampling $\theta$ marginalized over the factors, and utilizing the output of our simulated annealing algorithm to find the negative of the inverse Hessian at the mode of the current posterior of each block. The latter is scaled downwards by a multiplier of .01 or .001 and is used as the variance of the increment in the random walk proposal densities. What we find is that the results are similar but 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, for instance, sampling $\theta$ without marginalizing out the factors, or not using simulated annealing to define the covariance matrix of the increments, the performance of the RW-MH algorithm in terms of mixing worsens further.

A. Estimates of $G$, $\mu$ and $\delta$

The estimates of the $G$ matrix in Table 2 show that the matrix is essentially diagonal and that the diagonal elements corresponding to the macroeconomic factors are close to one.

The intercept of the short rate equation $\delta_1$ 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., $\delta_2$) are positive and significantly different from zero. These estimates are consistent with the Taylor rule intuition. The estimates of the mean parameters of the macroeconomic factors lie within half a standard deviation from their sample means. It can also be seen from the last two columns of this table that the inefficiency factors are somewhat large. The important point is that these factors would be much larger from an algorithm that is not as well tuned as ours.

In Figure 5 we report the prior-posterior updates of selected parameters from Table 2. These updates show that the prior and posterior densities are generally different, which indicates that the data carries information or, in other words, that there is significant learning from the data.
<table>
<thead>
<tr>
<th>param.</th>
<th>prior</th>
<th>posterior</th>
<th>average acc. rate</th>
<th>average ineff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>0.95</td>
<td>0.00</td>
<td>0.99</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.33)</td>
<td>(1.41)</td>
<td>(1.41)</td>
<td>(0.01)</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.95</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(0.33)</td>
<td>(1.41)</td>
<td>(0.00)</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.95</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.41)</td>
<td>(0.33)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0</td>
<td>75.00</td>
<td>0</td>
<td>76.25</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>(7.00)</td>
<td>...</td>
<td>(4.46)</td>
</tr>
<tr>
<td>$\delta_1$</td>
<td>-3.00</td>
<td>-3.650</td>
<td>88.3</td>
<td>62.8</td>
</tr>
<tr>
<td></td>
<td>(1.00)</td>
<td>(0.992)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta_2$</td>
<td>0.200</td>
<td>0.100</td>
<td>0.233</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>(0.447)</td>
<td>(0.333)</td>
<td>(0.447)</td>
<td>(0.011)</td>
</tr>
<tr>
<td></td>
<td>0.700</td>
<td>0.80</td>
<td>0.163</td>
<td>0.163</td>
</tr>
<tr>
<td></td>
<td>(0.447)</td>
<td>(0.447)</td>
<td>(0.011)</td>
<td>(0.017)</td>
</tr>
</tbody>
</table>

| $\gamma_1$ | 0.5  | 0.8      | 0.5               | 1              |
| $\gamma_2$ | -6   | -2       | 2                 | -15            |
| $\gamma_3$ | -5   | 5        | 15                | 25             |
| $\gamma_4$ | 80   | 80       | 100               |                 |

**Table 2:** Estimates of $G$, $\mu$ and $\delta$. Acceptance rates (acc.rate) are in percentages. Inefficiency factors (ineff.) are computed by (4.1). Standard deviations are in parentheses.

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

**B. Risk premia parameters**

The constant prices of risk of all factors, $\gamma$ are all negative and significant except for the
first. This 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 primarily responsible for determining the level of the yield curve. Moreover, we find that the estimate of the time-varying risk premium of inflation 

<table>
<thead>
<tr>
<th>Param.</th>
<th>Prior</th>
<th>Posterior</th>
<th>Average acc. rate</th>
<th>Average ineff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ</td>
<td>-100</td>
<td>-100</td>
<td>1.96</td>
<td>-114.2</td>
</tr>
<tr>
<td></td>
<td>(50)</td>
<td>(50)</td>
<td>(10.0)</td>
<td>(42.0)</td>
</tr>
<tr>
<td>Φ</td>
<td>1.00</td>
<td>0.00</td>
<td>-0.66</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.41)</td>
<td>(0.54)</td>
<td>(0.14)</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>1.00</td>
<td>0.16</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.41)</td>
<td>(1.04)</td>
<td>(0.56)</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.00</td>
<td>0.19</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>(1.41)</td>
<td>(1.41)</td>
<td>(1.42)</td>
<td>(1.39)</td>
</tr>
</tbody>
</table>

Table 3: Estimates of the risk premia parameters. Acceptance rates (acc.rate) are in percentages. Inefficiency factors (ineff) are computed by (4.1). Standard deviations are in parentheses.

\( \phi_{33} \) is positive. Our result suggests that investors demand higher compensation for the risk of inflation rising above its average level. However, Figure 6 show that it is difficult to accurately estimate some of the risk premia parameters in \( \Phi \).

C. Covariance matrices

We note that the estimated standard deviations of the residuals, \( \sigma \) of the measure equation (2.12) are large for the short and long maturities. This is not surprising on account of the fact that we have approximated the short rate by the Federal Funds Rate, which is much less volatile than any other yield. An alternative approach would be 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, \( \delta \). Because the parameters of the model are all scrambled together through the no-arbitrage recursions, the difficulty in inferring \( \delta \) makes it then more difficult to infer other parameters of the model.

D. Predictive densities

25
As one can see 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 predicts well the future dynamics of the both macroeconomic factors. Except for one 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 there is some room for improvement. In particular, the forecasts do not adequately capture the curvature of the yield curve. This shortcoming can likely be overcome by including additional latent factors in the model. This extension is the subject of ongoing work.
<table>
<thead>
<tr>
<th>Param.</th>
<th>Prior</th>
<th>Posterior</th>
<th>Average acc. rate</th>
<th>Average ineff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Omega$</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0.545</td>
<td>0</td>
<td>0.195</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>(0.816)</td>
<td>(0.019)</td>
<td></td>
<td>34.2</td>
</tr>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.536</td>
<td>0</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.396)</td>
<td>(0.533)</td>
<td></td>
<td>0.091</td>
</tr>
<tr>
<td>$\sigma_1^{2*}$ ... $\sigma_3^{2*}$</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(0.27)</td>
</tr>
<tr>
<td>$\sigma_4^{2*}$ ... $\sigma_6^{2*}$</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(0.63)</td>
</tr>
<tr>
<td>$\sigma_7^{2*}$ ... $\sigma_9^{2*}$</td>
<td>5.00</td>
<td>5.00</td>
<td>5.00</td>
<td>2.90</td>
</tr>
<tr>
<td></td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(64.0)</td>
<td>(0.27)</td>
</tr>
</tbody>
</table>

Table 4: Estimates of the covariance matrices of the L1M2 model. Acceptance rates (acc.rate) are in percentages. Inefficiency factors (ineff.) are computed by (4.1). According to the identification scheme $\omega_{11} = 1$. Standard deviations are in parentheses.

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

5 Conclusion

We have provided 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
Figure 8: Out of sample (January 2006 - December 2006) forecasts of the yield curve and macroeconomic factors by L1M2 model. The figure presents twelve months 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” respectively, are based on 25,000 draws. The observed surface and curves are labeled “Real.” Top two graphs represent two different views of the same yield forecasts.

In the 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 in this complicated problem because it concentrates attention on regions of the parameter space that might otherwise be missed, and 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 employed 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, is of relevance beyond our problem.
In sum, our analysis shows 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, for instance those with additional latent factors. We are studying such models and will report on them elsewhere.

References


A An example that demonstrates the possibility of sudden large changes in the likelihood

Consider the arbitrage-free model with the following parameters

\[ G = \text{diag}(0.93, 0.93, 0.93), \quad \mu = (0, 75, 4), \]
\[ \delta = (-3, 0.2, 0.1, 0.5), \quad \gamma = (-100, -100, -100), \]
\[ \Phi = \text{diag}(1, 1, 1), \quad \Omega = \text{diag}(1, 0.30, 0.13), \]
\[ \Sigma^* = \text{diag}(5, 5, 5, 5, 5, 5, 5, 5, 5). \]
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 number is comparable with historically observed average yields. The logarithm of the likelihood at the above given point is about \(-2.6 \times 10^4\). Now, consider the following change in the value of \( \delta_{22} \) 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 \(-1.0 \times 10^5\). If we also change the parameter \( \Phi \) so that \((G - LH^{-1}\Phi)\) 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 \(-1.7 \times 10^5\).

**B Square root filtering**

If \( M \) is a nonnegative definite symmetric matrix, a square root of \( M \) is a matrix \( N \) such that \( M = NN' \). Following the convention we also use the notation \( M^{1/2} \) to denote an arbitrary square root of \( M \). Let \( S_{t|t} \) and \( S_{t|t-1} \) denote square roots of \( R_{t|t} = \nabla(f_t|Y_t, \psi) \) and \( R_{t|t-1} = \nabla(f_t|Y_{t-1}, \psi) \), respectively. In a square-root filter the update equations are expressed in terms of \( S_{t|t} \) and \( S_{t|t-1} \). There are at least two important advantages of square-root filters. First, both \( R_{t|t} \) and \( R_{t|t-1} \) are always nonnegative definite. Second, the numerical conditioning of \( S_{t|t} \) (\( S_{t|t-1} \)) is much better than that of \( R_{t|t} \) (\( R_{t|t-1} \)) 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 that is used in this paper.

The time update of the square-root covariance matrix from \( S_{t-1|t-1} \) to \( S_{t|t-1} \) is based on the following matrix equation:

\[
\begin{pmatrix}
S'_{t|t-1} \\
0_{(k+m) \times (k+m)}
\end{pmatrix} = Q 
\begin{pmatrix}
S'_{t-1|t-1} \\
G'
\end{pmatrix},
\]

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

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

as its input and returns $Q$ and

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

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

$$\begin{pmatrix} (T\Sigma T' + BR_{t|t-1} B')^{1/2} & \bar{K}_t \\ 0_{(k+m)\times(p+m)} & S_{t|t} \end{pmatrix} = \bar{Q} \begin{pmatrix} (T\Sigma T)^{1/2} & 0_{(p+m)\times(k+m)} \\ S_{t|t-1} B' & S_{t|t-1} \end{pmatrix},$$

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

**Algorithm:** Calculation of the likelihood via the square root filter

**Step 1** Set $t = 1$ and initialize $f_{0|0}$, $S_{0|0}$

**Step 2** Calculate while $t \leq n$

(a) Time update:

(i) update factor forecast mean $f_{t|t-1} = G f_{t-1|t-1}$

(ii) given $S_{t-1|t-1}$, $G$ and $\Omega$ compute $S_{t|t-1}$ from (B.1)

(b) Measurement update:

(i) given $\Sigma$, $B$ and $S_{t|t-1}$, find $(T\Sigma T' + BR_{t|t-1} B')^{1/2}$, $\bar{K}_t$ and $S_{t|t}$ from (B.2)

(ii) compute $R_{t|t-1} = (T\Sigma T' + BR_{t|t-1} B')^{1/2}(T\Sigma T' + BR_{t|t-1} B')^{1/2}'$

(c) Compute $t$-th summand and the cumulative sum on the RHS of (3.1)
Step 3  Increment $t$ to $t + 1$ and go to Step 2

Step 4  Return $\log p(y|\psi)$