**Journal of Trends in Financial and Economics**
**ISSN: 3007-696X**
**DOI: 10.61784/jtfe3012**

# AFFINE TERM STRUCTURE MODEL WITH MCMC

JingShu Liu
*Questrom School of Business, Boston University, Boston 02215, MA, US.*
*Corresponding Email: jingshul@bu.edu*

**Abstract:** This paper develops a Bayesian Markov Chain Monte Carlo (MCMC) estimation method for multi-factor affine term structure models (ATSMs). ATSMs are popular, but efficient estimation methods for them are not readily available. Using simulated price data, the MCMC algorithms developed provide good estimates with their posterior distributions converge. With real historical data, the in-sample pricing errors obtained are significantly smaller than those obtained from alternative methods. A Bayesian forecast analysis documents the superior predictive power of the MCMC approach. Finally, Bayesian model selection criteria are discussed.
**Keywords:** Affine term structure model; Markov Chain Monte Carlo; Interest rate modeling

## 1 INTRODUCTION

Term structure models are essential building blocks for valuation, hedging and risk management of interest-rate-related derivatives. The yield curve conveys important macroeconomic information and play an important role for monetary policy transition, asset management, cross-border investment, etc. Among various term structure models, ATSMs are popular mainly because of their analytical tractability and flexibility.

To balance between the economic richness and computational cost for ATSMs, Dai and Singleton (2000) classify the $n$-factor affine family into $n+1$ non-nested sub-families: $A(m,n), m = 0,\cdots,n$[1]. The order $m$ is the dimension of the restricted state variables that enter the diffusion matrix. To incorporate the investor's attitude toward factor risks, state variable dynamics are specified differently under the data-generating measure $P$ and the risk-neutral measure $Q$. Different market price of risk specifications is proposed[1-3]. In this paper, we examine two models $A(0,3)$ and $A(1,3)$, under the "extended affine" specification of Cheridito et al. (2007), which allows for more parameters and nests the "completely affine" specification of Dai and Singleton (2000), and "essentially affine" specification of Duffee (2002)[1-3]. It also imposes parameter constraints to impose no-arbitrage conditions.

Despite the analytical tractability of ATSMs, the estimation of these models is difficult. First, there is an issue of the stochastic singularity. The ATSMs are driven by low dimensional latent state vectors, whereas we observe cross-sectional yields with larger number of maturities. Measurement errors are added to the models to remove this singularity. However, the choice of the measurement error distribution is arbitrary. Many researchers assume that yields with $n$ number of maturities are observed without error, where $n$ is the dimension of the state variables[1, 3-5]. In contrast, other yields present observation errors. However, the underlying assumption is stringent. As pointed out by Piazzesi (2008), it is more plausible to assume all yields present observation errors[6]. With this alternative assumption, the state variables cannot be inverted but be estimated.

Second, for the majority of ATSMs, maximum likelihood estimation methods are difficult to apply, as there are no closed-form expressions for the transition densities of the state variables and parameters.

Finally, even in the rare cases where we do obtain closed-form expressions for the likelihood function, parameters enter such functions in a highly non-linear way. As there is no analytical expression for the maximum likelihood estimators (MLEs) in such cases, the MLEs require solving high-dimensional optimization problems by numerical methods. Some deterministic search algorithms are often applied to locate the MLEs[3]. However, choices such as termination criteria and initial parameter values can be non-trivial. Local maxima are frequent. Difficulties in estimation negatively affect the results of ATSMs.

There are several methods to estimate ATSMs. As the likelihood functions do not have closed-form expressions for most ATSMs, the first set of such methods attempt to approximate the likelihood function. One can apply the Euler discretization scheme for the stochastic differential equations (SDEs) of the state dynamics. Estimators obtained by maximizing the likelihood function based on the conditional Gaussian distribution are called quasi-maximum likelihood estimators, which is feasible for any ATSMs. However, the resulting estimators are not consistent except for the linear Gaussian dynamics. One can also numerically solve the forward Kolmogorov partial differential equations for conditional densities (Lo and MacKinlay (1988)). However, such methods face the curse of dimensionality. Yet another approach attempts to approximate the likelihood function using simulation techniques [7-9]. Such methods are also computationally costly, especially for multi-dimensional problems. Recently, Ait-Sahalia and Kimmel (2010) propose a maximum likelihood estimation method that relies on an approximation of the transition density of multi-dimensional state variables by Hermite polynomial expansions[4]. Most likelihood-based methods depend on the assumption that some yields are observed without error. There is little empirical evidence for this assumption. The presence of measurement noise in all yields requires the calculation of the filtered likelihood function, for which optimization is even more challenging. Another set of methods is based on moment matching (Hansen (1982)), using the feature that moments of affine diffusions are available in closed forms. However, one problem with this approach is that the states are never determined explicitly. Rather, it is possible to obtain state variables that lie outside of their domain [2].

The Bayesian inference analysis for ATSMs in this paper complements the previous analysis that resides in the frequentist domain. Specific to the ATSMs inference problem, the merits of Bayesian analysis can be exploited using MCMC methods. Within this framework, it is straightforward to relax the stringent assumption that an arbitrary set of yields are observed without error. This is empirically relevant. Our inference output based on market data indeed suggests that the measurement errors are of similar magnitude for yields across maturities. They also exhibit cross-sectional correlations. Now, with the relaxation of this assumption, the inversion method for state variables is not feasible.

In our MCMC framework, the latent states are inferred together with parameters. We use Gibbs samplers to alleviate the curse of dimensionality. They allow us to decompose the problem of sampling from the joint posterior distribution of parameters and latent states into a cycle of univariate sampling problems. Although the algorithm is designed to include the sampling of latent states, it can be readily applied to ATSMs with explicit states, such as moments extracted from yield data or macro factors[2,10]. Inference results on simulated yield data from different models show that our algorithm can closely replicate the observations. In the setting with explicit state variables, the algorithm generates samplers and likelihood values that converge to the true values. A common criticism against Bayesian analysis is the inclusion of prior distributions. We use diffuse prior distributions for most parameters. The performances on simulated yield data show that the algorithm is robust to disperse prior distributions.

We show that the MCMC algorithms also deliver good performances when applied to different market data sets. Data I consists of yields constructed from LIBOR and swap weekly observations for 1989.03.31 – 2007.03.02 for maturities 1, 3, 6, 9 and 12 months and 2, 3, 5, 7 and 10 years. Data II is the Fama-Bliss zero-coupon bond yields with monthly observations for 1972.01.31-2010.12.31 for maturities 1, 2, 3, 4 and 5 years. These two periods represent different economic episodes.

We document the strength of the MCMC methods in reconstructing yield data. The fitting errors, measured by the rooted mean square errors (RMSE), are smaller than that obtained by the inversion-MLE method and the model-free method in Collin-Dufresne et al. (2008)[10]. It is commonly thought that the principal component analysis (PCA) can produce small fitting errors that are hard to beat by affine models, as pointed out in Piazzesi (2008)[6]. Our results from the $A(1, 3)$ model show that this is not necessarily the case. We obtain in-sample RMSEs that are comparable to that from the PCA.

We conduct an in-sample forecast analysis to identify missing data. The resulting pricing errors are within a few basis points (One basis point is 0.0001). Some other data construction methods such as Cubic splines, Nelson-Siegel and Svensson families are used in the industry[11-12]. These fitting methods explore the data in the cross-section but fail to consistently fit them in the time-series. Furthermore, these models, including PCA, lack probabilistic interpretations and no-arbitrage (NA) free. Our MCMC algorithms satisfy NA conditions by imposing parameter constraints provided in Cheridito et al. (2007)[3].

Finally, we can construct the short rates from the output. To fit the short end of the yield curve has been challenging because of seasonality and/or microstructure noise[6]. The inferred short rates from MCMC algorithms closely resemble the 1-month yield data, which are often used as a proxy for short rates.

Equipped with these efficient sampling schemes, we forecast future yield levels. For Data I, with the $A(0,3)$ and $A(1,3)$ models, we can forecast the 12- week-ahead yield levels with out-of-sample RMSEs within 5 basis points. We run a horse-race among several prediction methods. The Bayesian forecast performance of the $A(1,3)$ model dominates the OLS prediction and frequentist type prediction for all maturities. It also dominates the random walk prediction for all maturities greater or equal to one year.

We also conduct a full-fledged Bayesian model comparison for different ATSMs. The model mis-specification analysis from previous work has been exclusively based on goodness-of-fit tests[1,3]. Ait-Sahalia and Kimmel (2010) perform a model comparison based on the likelihood ratio test for (non)-nested models[4]. However, illustrative examples in Bishop (2007) show that estimation methods that perform well within the sample do not necessarily indicate a good predictive behavior because of over-fitting[13]. Various information criteria, such as AIC and BIC, have been proposed to penalize for over-parametrization, which often tend to favor simple models. From the Bayesian perspective, the problem of over-fitting can be avoided because the effective model complexity adapts automatically to the data. We find that the ranking of the two models by the model evidence is consistent with both in-sample fitting and out-of-sample forecast performances. Data I has the features of non-normality of yield distribution and a humped shape of yield change volatility. It ranks the models as $A(1, 3)$ over $A(0, 3)$. Data II also supports the non-normality feature but has a strong demand for the correlation between state variables. The $A(0, 3)$ model is preferred by the data.

The MCMC analysis has been successfully applied to a wide range of stochastic volatility models. Jacquier et al. (1994) use MCMC methods to analyze the log stochastic volatility model, and then incorporate the leverage effect of stochastic volatilities [14-15]. Eraker et al. (2003) examine the Heston's volatility model by MCMC, and later extend the analysis to include jumps in volatility and return processes[16]. Johannes and Polson (2007) provide a review of recent developments regarding the application of MCMC in estimating various financial models[17].

In the multi-factor ATSMs literature, the MCMC analysis is relatively new. This might be due to the difficulty in evaluating the yields. In addition, state variable transition densities do not have closed-form expression for most ATSMs. Furthermore, the data-collection frequency in the literature prevents the usage of Euler discretization, which is common in other areas with wide application of MCMC. Early utilization of MCMC in ATSMs can be found in Chen and Scott (1993) and Hu (2005)[9,18]. Chen and Scott (1993) apply the extended Kalman filter for the multi-factor CIR model[9]. Hu (2005) conducts the MCMC analysis on multi-factor Vasicek and multi- factor CIR models[18]. As these models have limitations in capturing market data features, here we study more general models. Furthermore, we provide results for yield level forecasts and model comparison.

This chapter is organized as follows. In Section 2, we provide a brief introduction to MCMC methods. In Section 3, we apply the MCMC algorithms for ATSMs and investigate its performance with simulated data. Results of empirical studies on two sets of zero-coupon bond yield data are presented in Section 4. Section 5 concludes.

## 2 MCMC METHODS

In this section, we give a brief introduction to the MCMC methods applied in this paper. A thorough treatment of the theoretical foundations can be found in Robert and Casella (2004)[19]. The advantages of these methods rely mainly on two facts. First, they are useful in sampling random variables for which the conventional sampling scheme are not available. Second, they are beneficial in decomposing the problem of sampling from the high-dimensional density into a sequence of univariate sampling problems. Both features are helpful in addressing the inference problem for ATSMs.

Suppose that we observe data Y. The latent state variables X and the parameter set Φ are unknowns. The target is to infer Φ and X from the joint posterior distribution:

$$\mathrm{P}(\Phi, \mathrm{X}|\mathrm{Y}). \tag{1}$$

Depending on the specific ATSM and the sample size, there are many parameters (approximately 40-80) to estimate, most of which enter the posterior density in a highly nonlinear way. As little is known about the density's properties, the traditional inversion and/or rejection-acceptance sampling schemes are not feasible. As discussed in the next paragraph, the Metropolis-Hastings (MH) algorithms are applied to sample these parameters. In addition to the parameter set Φ, we need to infer the latent states, which contribute 1000-3000 more random variables to sample, with the size depending on data window. The decomposition of high-dimensional problems into small units is extremely useful in reducing the complexity of the problem.

### 2.1 Metropolis-Hastings Algorithms

Suppose we want to sample a Markov chain $\{\mathrm{x}^{(\mathrm{t})}, \mathrm{t} = 1, 2, 3, \cdots\}$ with the stationary distribution $\mathrm{f}(\cdot)$, the target density. Conditioning on $\mathrm{x}^{(\mathrm{t})}$, we can use the MH algorithms and choose a conditional density $\mathrm{q}(\mathrm{y}|\mathrm{x})$, the proposal or candidate distribution, to help to sample $\mathrm{x}^{(\mathrm{t}+1)}$. The choice of $\mathrm{q}(\cdot|\mathrm{x}^{(\mathrm{t})})$ is delicate and needs to be tuned to specific problems. We will provide guidance in choosing this proposal density for ATSMs in the next section.

Step 1: Generate:

$$\begin{gathered} Yt \sim q\left(y|x^{(t)}\right), \\ u \sim U(0,1), \end{gathered} \tag{2}$$

where $\mathrm{U}(0, 1)$ is the uniform distribution in $(0, 1)$.

Two special cases of MH algorithms, the independent MH algorithm and the random walk MH algorithm, are used in this paper. The former is useful when we have a good understanding of the target density and can propose efficient candidate densities. The latter explores the neighborhood of the Markov chain and gathers local information about the target stepwise. It is particularly useful for multi-dimensional densities.

### 2.2 The Independent Metropolis-Hastings Algorithm

Key: Candidate draws are independent of the previous states: $\mathrm{Y}_\mathrm{t} \sim \mathrm{q}(\mathrm{y})$

For this algorithm, a good candidate density $\mathrm{q}(\cdot)$ needs to fulfill two requirements. First, it needs to approximate the target density closely in the shape and location. With the close approximation, the acceptance probability for new samplers is high, and the sampling scheme is efficient. This requires us to be informative about the target density. Second, the candidate density should be diffusive to navigate through the entire support of the target density.

In many cases we have little information about the properties of the target density $\mathrm{f}(\cdot)$, and thus it is hard to propose a good candidate density. The random walk MH algorithm provides an alternative approach.

#### *2.2.1 The random walk Metropolis-Hastings algorithm*

Key: Candidate draws are current state with a symmetric random walk perturbation: $\mathrm{Y}_\mathrm{t} \sim \mathrm{g}(\mathrm{y} - \mathrm{x}_\mathrm{t})$.

Here $\mathrm{g}(\cdot)$ is a symmetric distribution that is independent of $\mathrm{x}_\mathrm{t}$.

The two algorithms described above are helpful for the inference problem of the ATSMs. The other difficulty that affects the inference for ATSMs is that the posterior distribution is high-dimensional. The multistage Gibbs sampler described next is suitable for tackling this problem.

#### *2.2.2 The Gibbs sampler[20]*

Key: Joint densities are decomposed into iterations of conditional densities.

Denote $(X^{(t)}, \Phi^{(t)})$ as the samplers at the t-th entry in the Markov chain.

Step 1: Generate:

$$X^{(t+1)} \sim P_1(x|\Phi^{(t)}, Y). \tag{3}$$

Step 2: Generate:

$$\Phi^{(t+1)} \sim P_2(\phi|X^{(t+1)}, Y). \tag{4}$$

Here $P_1\left(x|\Phi^{(t)}, Y\right)$ and $P_2\left(\phi|X^{(t+1)}, Y\right)$ denote the distributions of each block of parameters or states conditioning on all others. Hammersley and Clifford (1971) prove the convergence of distributions $\left\{X^{(t+1)}, \Phi^{(t+1)}\right\}$ to the stationary joint distribution $P(X, \Phi|Y)$, as the number of iterations increases[21].

In the Gibbs sampling cycle, we decompose the target posterior distribution $P(\Phi, X|Y)$ into its full conditionals $P_1(x|\Phi^{(t)}, Y)$ and $P_2(\phi|X^{(t+1)}, Y)$. The distributions $P_1(x|\Phi^{(t)}, Y)$ and $P_2(\phi|X^{(t+1)}, Y)$ can be further decomposed into finer full conditional densities until efficient sampling methods are available.

### 2.3 The Kalman Filter and FFBS

The forward filtering backward sampling (FFBS) method is designed for the linear Gaussian model and is an efficient simulation version of smoothing recursions of the Kalman filter[22-23]. For the Gibbs sampler, we iterate through sampling latent states and parameters. In the first step, the FFBS algorithm samples the latent linear Gaussian states in a block. In this section, we first introduce the Kalman filter and then the FFBS algorithm.

Suppose that the state variables $\{X_t\}_{t=0}^{T}$ have a normal prior distribution at $t = 0$:

$$X_0 \sim X_p(m_0, G_0). \tag{5}$$

The observation equation and state equation follow the linear Gaussian system:

$$\begin{aligned} Y_t &= F_t X_t + v_t, v_t \sim N(0, V_t), \\ X_t &= G_t X_{t-1} + w_t, w_t \sim N(0, W_t). \end{aligned} \tag{6}$$

First, assume that the parameter set $\Phi$ in the $(Ft, Vt, Gt, Wt)$ matrices is known. Denote the observations of $\{Y_s, s = 1 \cdots t\}$ as $Y_{1:t}$. Because of the linear Gaussian structure and normal prior, the filtered distribution $P(X_{t-1}|Y_{1:t-1}, \Phi)$ is normally distributed, hence is fully characterized by its mean and variance. The Kalman filter is a mechanism for recursively updating the mean and variance of $P(X_{t-1}|Y_{1:t-1}, \Phi)$.

#### *2.3.1 The Kalman filter[24]*

Suppose at $t - 1$, the distribution of $p(X_{t-1}|Y_{1:t-1}, \Phi)$ is:

$$p(X_{t-1}|Y_{1:t-1}, \Phi) \sim N\,(m_{t-1}, C_{t-1}). \tag{7}$$

The predictive distribution of $P(X_t|Y_{1:t-1}, \Phi)$ is Gaussian:

$$p(X_t|Y_{1:t-1}, \Phi) \sim N(a_t, R_t), a_t = G_t m_{t-1}, R_t = G_t C_{t-1} G_t^{'} + W_t. \tag{8}$$

The predictive distribution of $p(Y_t|Y_{1:t-1}, \Phi)$ is Gaussian:

$$p(Y_t|Y_{1:t-1}, \Phi) \sim N\,(f_t, Q_t),\ f_t = F_t a_t, Q_t = F_t R_{tF_t^{'}} + V_t. \tag{9}$$

The filtered distribution of $p(X_t|Y_{1:t}, \Phi)$ is Gaussian:

$$p(X_t|Y_{1:t}, \Phi) \sim N\,(m_t, C_t), m_t = a\_t + R_{tF_t^{'}Q_t^{-1}} e_t, C_t = R_t - R\_tF_t^{'}Q_t^{-1}F_t R_t, \tag{10}$$

where $e_t$ is the predicting error:

$$e_t = Y_t - f_t. \tag{11}$$

Now we assume the parameter set $\Phi$ is unknown. Denote the latent states $\{X_s, s = 1 \cdots t\}$ as $X_{1:t}$. To sample from $P(\Phi, X_{1:T}|Y)$, we iterate through the full conditional decompositions:

$$P_1(X_{1:T}|\Phi, Y), P_2(\Phi|X_{1:T}, Y). \tag{12}$$

The first component is sampled by the FFBS algorithm. The second component is sampled by the random walk MH, independent MH, and/or other standard sampling schemes.

The target of interest is

$$P(X_{1:T}|Y_{1:T}, \Phi). \tag{13}$$

This joint distribution can be decomposed as:

$$P(X_{1:T}|Y_{1:T}, \Phi) = \Pi_{t=0}^{T} P(X_t|X_{t+1:T}, Y_{1:T}, \Phi). \tag{14}$$

Starting from $P(X_T|Y_{1:T}, \Phi)$, the last component in the forward iteration of filtered distribution in the Kalman filter, we iterate backward and sample sequentially:

$$P(X_t|X_{t+1:T}, Y_{1:T}, \Phi) = P\,(X_t|X_{t+1}, Y_{1:t}, \Phi), t = 1, \cdots, T - 1. \tag{15}$$

This last equality comes from the Markovian structure of the model.

#### *2.3.2 The FFBS algorithm [22]*

Initialization: Set:

$$P(X_T|Y_{1:T}, \Phi) \sim N(m_T, C_T). \tag{16}$$

Backward iteration: For $t = T-1, \cdots, 0$, we have:

$$p(X_t|X_{t+1}, Y_{1:t}, \Phi) \sim N(ht, Ht), h_t = m_t + C_t G'_{t+1} R^{-1}_{t+1}(X_{t+1} - a_{t+1}), H_t = C_t - C_t G'_{t+1} R^{-1}_{t+1} G_{t+1} C_t. \tag{17}$$

The MH sampling algorithms, Gibbs iterations, and the FFBS algorithm provide the essential building blocks for our MCMC algorithms for ATSMs. In the section of designing the MCMC algorithms, we tailor these algorithms to each ATSM.

### 2.4 The Model Comparison

A full Bayesian treatment of ATSMs involves a model-comparison analysis. Denote a specific ATSM by M. The posterior distribution is uniquely determined by the conditional probability of the unknowns given the observed data and the specific model:

$$P(\Phi, X|M, Y) = \frac{P(Y|\Phi, X, M)P(\Phi, X|M)}{Z}, \tag{18}$$

where $P(\Phi, X|M)$ is the prior distribution and $P(Y|\Phi, X, M)$ is the likelihood. The marginal likelihood Z is defined as:

$$Z \equiv P(Y|M) = \int\int P(Y|M, \Phi, X)P(\Phi, X|M)d\Phi dX. \tag{19}$$

The marginal likelihood Z is interpreted as the model evidence, as it measures the support for model M given data Y. The quantity Z carries information that allows us to make a model comparison from a Bayesian perspective. Such a comparison does not require alternative models to be nested. A detailed analysis of the advantages of Bayesian model comparison can be found in Kass and Raftery (1995)[25].

Direct computation of the marginal likelihood Z is not feasible, as it involves multi-dimensional integration over the parameters and latent states. Previously, the Schwarz information criterion (Schwarz (1978)) and the Laplace approximation method have been used to approximate the model evidence factor. The latter is the Bayesian inference criterion (BIC) in the frequentist framework[26]. As a quite different approach, the reversible-jump MCMC algorithm (Green (1995)) has been developed to incorporate different models as discrete parameters in the Gibbs iterations. The MCMC algorithm switches among different models during the iterations, and the model-comparison result is generated as an output of such algorithms. For the ATSMs inference problem, as we can efficiently sample from the posterior distribution, we can approximate the quantity Z by the harmonic mean of the likelihood values[27]:

$$\hat{Z} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{P(Y|Xi, \Phi i)}\right]^{-1}, \tag{20}$$

where N is the number of simulations, and $\{X_i, \Phi_i\}$ are the i-th simulation from the joint posterior distribution $P(X, \Phi|Y)$.

## 3 MCMC ALGORITHMS FOR ATSMS

We study two classes of ATSMs: $A(0, n)$ and $A(m, n)$ with $m = 0{,}1$ and $n = 3$. We cast the analysis under the "extended affine" specification of Cheridito et al. (2007)[3]. In the $A(m, n)$ model, the dynamics of state variables:

$$dX = \left(O^p X_t + O^p_0\right)dt + \mathrm{diag}\left[(C_i X_i + 1)^{0.5}\right]dW^P_t = (OX_t + O_0)dt + \mathrm{diag}\left[(C_i X_i + 1)^{0.5}\right]dW^Q_t, \tag{21}$$

where $O^p, O \in R^{n\times n}$, $O^p_0, O_0 \in R^n$, and $C_i$ is the i-th row of the matrix C with rank m. The parameter values are within the constraints that insure the existence, stationary and no-arbitrage conditions. The term $W^p_t$ is an n-dimensional Brownian motion under the data-generating measure P, and $W^Q_t$ is an n-dimensional Brownian motion under the risk-neutral measure Q.

The short rate process is a linear combination of state variables:

$$r_t = \delta_0 + \delta' X_t, \tag{22}$$

for some $\delta_0 \in R$ and $\delta \in R^n$.The observation equation is

$$Y_t = V'_b X_t - V_0 + \epsilon. \tag{23}$$

Here $Y_t$ is a $T \times 1$ vector with the yields of T maturities at time t. $V_b$ is a $n \times T$ matrix and $V_0$ is a $M \times 1$ vector. The terms $V_b$ and $V_0$ solve Riccati-type ODEs in Duffie et al. (2000)[28]. In the $A(0, n)$ model, $V_b$ can be solved explicitly. The measurement error, denoted by $\epsilon$, has a normal distribution of $N(0, \Sigma)$, where $\Sigma$ is a $T \times T$ symmetric and positive definite matrix.

For the $A(0, n)$ model, we simulate zero coupon yield data for 43 years with monthly observations. The maturities are 1, 2, 4, 6, 7, 8 ad 10 years. These specifications are chosen to match those frequently used in the literature. For the simulated data, the parameter values for different models are the estimation results in Cheridito et al. (2007) with slight modifications[3]. The parameter values are reported in Table 1 and 2.

**Table 1** Parameter Inference Results, A(0,3) with Explicit States

| Parameter | True Value | Mean | Std. Dev. | 95% Confidence Interval |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| $\eta_0$ | 0.0529 | 0.0537 | 0.0003 | [0.0529, 0.0545] |
| $\eta_1$ | 0.0209 | 0.0194 | 0.0004 | [0.0186, 0.0202] |
| $\eta_2$ | 0.0226 | 0.0222 | 0.0004 | [0.0216, 0.0230] |
| $\eta_3$ | 0.0279 | 0.0272 | 0.0003 | [0.0268, 0.0278] |
| $OO_1$ | 0.4349 | 0.2655 | 0.1118 | [0.0598, 0.4985] |
| $OO_2$ | 0.2360 | 0.2977 | 0.0864 | [0.1226, 0.4538] |
| $OO_3$ | 0.3454 | 0.3411 | 0.0174 | [0.3056, 0.3735] |
| $O_{11}$ | -0.8600 | -0.4910 | 0.1136 | [-0.7001, -0.3078] |
| $O_{12}$ | -0.1600 | -0.2641 | 0.0814 | [-0.3956, -0.1125] |
| $O_{13}$ | -0.3800 | -0.2857 | 0.0626 | [-0.4254, -0.1858] |
| $O_{21}$ | -0.3200 | -0.4955 | 0.0962 | [-0.6477, -0.3325] |
| $O_{22}$ | -0.6000 | -0.4376 | 0.0679 | [-0.6000, -0.3439] |
| $O_{23}$ | -0.1200 | -0.1194 | 0.0502 | [-0.2026, -0.0434] |
| $O_{31}$ | -0.1600 | -0.1399 | 0.0220 | [-0.1806, -0.0896] |
| $O_{32}$ | -0.2400 | -0.2754 | 0.0167 | [-0.3018, -0.2313] |
| $O_{33}$ | -0.4000 | -0.4061 | 0.0122 | [-0.4274, -0.3787] |

*Note: This table reports the parameter inference results for the $A(0,3)$ model with explicit states. Monthly observations of zero-coupon bond yield data for 43 years with maturities 1 ,2, 4, 6, 7, 8 and 10 years are simulated with the true parameter values reported in the table.

**Table 2** Parameter Inference Results, A(1,3) with Explicit States

| Parameter | True Value | Mean | Std. Dev. | 95% Confidence Interval |
|---|---|---|---|---|
| $\eta_0$ | 0.0529 | 0.0531 | 0.0003 | [0.0526, 0.0536] |
| $\eta_1$ | 0.0209 | 0.0209 | 0.0002 | [0.0205, 0.0213] |
| $\eta_2$ | 0.0226 | 0.0226 | 0.0001 | [0.0224, 0.0227] |
| $\eta_3$ | 0.0279 | 0.0279 | 0.0000 | [0.0278, 0.0280] |
| $OO_d$ | 1.2349 | 1.1259 | 0.0052 | [1.0232, 1.2003] |
| $OO_1$ | 0.2360 | 0.3176 | 0.0022 | [0.2836, 0.3715] |
| $OO_2$ | 0.3454 | 0.3350 | 0.0050 | [0.3255, 0.3441] |
| $O_{dd}$ | -0.8600 | -0.8901 | 0.0427 | [-0.9477, -0.7878] |
| $O_{d1}$ | -0.3200 | -0.2963 | 0.0357 | [-0.3483, -0.2257] |
| $O_{d2}$ | -0.1600 | -0.1600 | 0.0106 | [-0.1801, -0.1372] |
| $O_{11}$ | -0.6000 | -0.6053 | 0.0141 | [-0.6345, -0.5738] |
| $O_{12}$ | -0.1200 | -0.1297 | 0.0074 | [-0.1433, -0.1106] |
| $O_{21}$ | -0.2400 | -0.2379 | 0.0056 | [-0.2505, -0.2260] |
| $O_{22}$ | -0.4000 | -0.3954 | 0.0031 | [-0.4040, -0.3895] |
| $C_1$ | 0.3000 | 0.2532 | 0.0765 | [0.4078, 0.7520] |
| $C_2$ | 0.3000 | 0.5229 | 0.0940 | [0.1159, 0.3640] |

*Note: This table contains the parameter inference results for the $A(1,3)$ model with explicit states. Monthly observations of zero-coupon bond yield data for 43 years with maturities 1, 2, 4, 6, 7, 8, and 10 years are simulated with the true parameter values reported in the table.

We denote the observation of yield data as $\mathrm{Y} \equiv \{\mathrm{Y}_t^\tau\}_{t=1}^{T}$ where $\tau$ is the maturity of zero-coupon bond. The state variables are denoted as $\mathrm{X} \equiv \{\mathrm{X}_t^\tau\}_{t=1}^{T}$. The parameter space is denoted as $\Phi \equiv \{\Phi^P, \Phi^Q\}$. Here $\Phi^P$ is the parameters governing the state dynamics under the data-generating measure $\mathrm{P}$, and $\Phi^Q$ are the parameters under the risk-neutral measure $\mathrm{Q}$. We sample the parameters and latent states from the posterior distribution:

$$\mathrm{P}(\Phi^Q, \Phi^P, \mathrm{X}|\mathrm{Y}). \tag{24}$$

We can apply the Gibbs sampling to iterate through the full conditionals decomposition:

$$\mathrm{P}(\Phi^P|\Phi^Q, \mathrm{X}, \mathrm{Y}),\ \mathrm{P}(\Phi^Q|\Phi^P, \mathrm{X}, \mathrm{Y}),\ \mathrm{P}(\mathrm{X}|\Phi^P, \Phi^Q, \mathrm{Y}). \tag{25}$$

In the sampling process for both $\Phi^P$ and $\Phi^Q$, parameter constraints are imposed. The summary of parameter constraints for existence, stationarity, and boundary non-attainable conditions can be found in Ait-Sahalia and Kimmel (2010)[4]. The boundary non-attainable condition prevents arbitrage opportunities for the "extended affine" specification. Samplers falling out of such constraints are discarded.

In what follows, we first discuss sampling algorithms for posterior distributions $\mathrm{P}(\Phi^P|\Phi^Q, \mathrm{X}, \mathrm{Y})$ and $\mathrm{P}(\Phi^Q|\Phi^P, \mathrm{X}, \mathrm{Y})$. Then we introduce the sampling of latent states under different models.

### 3.1 Sampling $\Phi^P$

As only $\Phi^Q$ enters the pricing formula, the yield data do not contain information directly for $\Phi^P$ . The data provide information for $\Phi^P$ indirectly through the channel of state variables X. To sample $\mathrm{P}(\Phi^P\ |\Phi^Q, \mathrm{X}, \mathrm{Y})$, we need to sample from the posterior distribution:

$$\mathrm{P}(\Phi^P|\mathrm{X}) = \mathrm{P}(\mathrm{X}|\Phi^P) \times \mathrm{P}_0(\Phi^P), \tag{26}$$

where $\mathrm{P}_0(\Phi^{\mathrm{P}})$ denotes the prior distribution for $\Phi^{\mathrm{P}}$. The posterior distribution of $\Phi^{\mathrm{P}}$ cannot be sampled using conventional schemes. Therefore, the random walk MH algorithm is applied.

In the rest of the sections, we focus on the sampling of $\Phi^{\mathrm{Q}}$, as they are important for evaluating the performances of in-sample fitting, out-of-sample forecast, and model comparison.

### 3.2 Sampling $\Phi^{\mathrm{Q}}$

To sample $P(\Phi^Q|\Phi^P, X, Y)$, we explore information in both the state variables $X$ and yield data $Y$. To sample the total parameter set, we further break $P(\Phi^Q|\Phi^P, X, Y)$ down to the full conditional decomposition of each parameter. Here we abbreviate $P(\theta|\cdot)$ as the full conditional decomposition for the parameter $\theta$. Only few parameters, $\delta_0$, $O_b^0$ and $\Sigma$, have conjugate posterior distributions that can use conventional sampling algorithms. The two distributions, $P(\delta_0|\cdot)$ and $P(O_b^0|\cdot)$, are sampled from the normal distributions. The distribution $P(\Sigma|\cdot)$ is sampled from the Wishart distribution. Further information about the choice of conjugate distributions can be found in Bishop (2007)[13]. As there are no standard posterior distributions for the rest of the parameters, they are sampled from the random walk MH algorithms.

In the first simulation exercise, we set state variables $X$ as known and only sample parameters. We refer the models where the state variables are known as models with explicit states. The inferred parameters can be used to reconstruct the yield data. We compare the (simulated) true yield data and the yield data reconstructed with the inferred parameters. In Table 3, the “Exp” column reports the in-sample pricing errors for the two models measured by RMSE in bps. As we can see, the in-sample errors are indeed small, which indicates that our algorithm can generate parameter samplers that closely reconstruct the yield data. This table also reports the simulated and true log-likelihood values, which approach the true maximal value.

**Table 3** In-Sample Fitting Performances with Simulated Data

| | $A(0{,}3)$ | | $A(1{,}3)$ | |
|---|---|---|---|---|
| | Exp | Latent | Exp | Latent |
| | | RMSE (bps) | | |
| 1-Year | 0.72 | 29.45 | 0.27 | 36.38 |
| 2-Year | 1.67 | 20.67 | 0.39 | 24.42 |
| 4-Year | 1.94 | 13.89 | 0.64 | 5.97 |
| 6-Year | 1.54 | 11.30 | 0.75 | 6.21 |
| 7-Year | 1.04 | 10.50 | 0.87 | 7.43 |
| 8-Year | 1.13 | 9.86 | 1.40 | 8.85 |
| 10-Year | 3.02 | 8.89 | 2.00 | 10.71 |
| | | Log Likelihood | | |
| True | 23267.4 | 23265.5 | 23262.2 | 23298.1 |
| Simulated | 21822.0 | 20900.0 | 23254.0 | 22599.0 |

*Note: This table reports the in-sample pricing errors and log-likelihood values of the inference output from the simulated data. Monthly observations of zero-coupon bond yield data for 43 years are simulated with the true parameter values and maturities reported in Table 1-2.

### 3.3 Sampling X in $\mathrm{A}(\mathbf{0}, \mathbf{n})$

In this section, we describe the sampling scheme for latent states in the $A(0, n)$ model. Specifically, the state variables have the dynamics:

$$dX_t = O_p X_t dt + dW_t^P = (O_{bb}Xt + O_b^0)dt + dW_t^Q. \tag{27}$$

Denote the time interval between the time indices $t$ and $t+1$ as $\Delta$. In the $A(0, n)$ model, conditioning on $X_{t-1}$, $X_t$ has a Gaussian transition density:

$$X_t \sim N\left((e^{O_{bb}\Delta}X_{t-1} + \int_{t-1}^{t} e^{O_{bb}(t-s)}O_b^0 ds\,, \int_{t-1}^{t} e^{O_{bb}(t-s)}\, e^{O_{bb}'(t-s)} ds\right). \tag{28}$$

As the state variables follow a linear Gaussian dynamic and enter observation equation linearly, the FFBS algorithm can be applied to simulate $P(X|Y, \Phi^Q)$[22,23,29].

Closely related to the FFBS algorithm is the Kalman filter. The Kalman filter gives a mechanism to recursively evaluate the Gaussian predictive, filtering and smoothing distributions when the parameters are known. In particular, the Kalman smoother provides a backward recursion mechanism to compute the conditional distribution of $P(X_t|Y, \Phi^Q)$ for $t = T, \cdots 1$. Instead of recursively sampling $P(X_t|Y, \Phi^Q)$, $t = T \cdots 1$ backwards, the FFBS algorithm simulates $X$ from $P(X|Y, \Phi^Q)$ in one block.

In the simulation exercise, we apply the MCMC algorithms described for the inference of parameters and latent states for the $A(0, 3)$ model. Table 3 reports RMSEs and log-likelihood values. Based on the results, we can see that the algorithm produces inferred parameters and latent states that replicate the in-sample data closely and
log-likelihood value that approaches the true maximal value.

### 3.4 Sampling X in $\mathrm{A}(\mathbf{1},\mathbf{n})$

In the $A(1,n)$ model, the state variables in the first dimension are restricted to be positive. This state variables drive conditional volatilities and thus compensate for the counterfactual assumption of constant conditional variances in the $A(0,n)$ model. Denote the restricted state variables as $X_0 \equiv \{X_t^0\}_{t=1}^T$ and the unrestricted state variables as $X1 \equiv \{\mathrm{X}_t^1\}_{t=1}^T$. The dynamics of the restricted state variables are

$$dX_t^0 = (O_{dd}\, X_t^0 + O_0)dt + \sqrt{X_t^0}dW_{0,t}^Q. \tag{29}$$

For $n = 3$, the unrestricted state variables are two-dimensional with the dynamics:

$$dX_t^1 = \left(O_{bd}\, X_t^0 + O_{bb}X_t^1 + O_b^0\right)dt + \begin{pmatrix} \sqrt{c_1 X_t^0 + 1} & 0 \\ 0 & \sqrt{c_2 X_t^0 + 1} \end{pmatrix} dW_{1,t}^Q. \tag{30}$$

The posterior distribution we want to sample is $P(\Phi^Q, \Phi^P, X^0, X^1|Y)$.

In the Gibbs sampling process, we iterate through the cycle of the full conditional decomposition of the posterior distribution:

$$\mathrm{p}(\Phi^\mathrm{Q}|\mathrm{X}^0, \mathrm{X}^1, \Phi^\mathrm{P}, \mathrm{Y}),\, \mathrm{p}(\Phi^\mathrm{P}|\mathrm{X}^0, \mathrm{X}^1, \Phi^\mathrm{Q}, \mathrm{Y}),\, \mathrm{p}(\mathrm{X}^0|\mathrm{X}^1, \Phi^\mathrm{Q}, \Phi^\mathrm{P}, \mathrm{Y}),\, \mathrm{p}(\mathrm{X}^1|\mathrm{X}^0, \Phi^\mathrm{Q}, \Phi^\mathrm{P}, \mathrm{Y}). \tag{31}$$

As we have discussed about sampling parameters $\Phi^\mathrm{P}$ and $\Phi^\mathrm{Q}$, we focus on sampling of the latent states in this section.

#### *3.4.1 Sampling $X^0$*

The restricted state variables $\{X_t^0\}_{t=1}^T$ enter the pricing formula linearly but have a non-linear dynamic. We can apply the Gibbs samplers to further decompose the vector $X^0$ into the cycle:

$$P\left(X_t^0 \middle| X_{t-1}^0, X_{t+1}^0, Y_t, X_{t-1}^1, X_t^1, X_{t+1}^1, \Phi^Q\right), t = 1, \cdots T \tag{32}$$

Even with this decomposition, we still cannot draw the univariate state variable $X_t^0$ directly. However, we can propose an efficient candidate density in the independent MH algorithm by exploiting the linearity of $X_t^0$ in the observation equation and its dynamics. For the independent MH algorithm, an efficient candidate density needs to closely resemble the shape and location of the target density in equation (3) for each $t$. In the following analysis, the parameter set $\Phi^Q$ is omitted for notational clarity. The target posterior density in (3) can be further decomposed:

$$P\left(X_t^0 \middle| X_{t-1}^0, X_{t+1}^0, Y_t, X_t^1, X_{t+1}^1, X_{t+1}^1\right) \propto P(Y_t|X_t^0, X_t^1) \times P(X_{t-1}^0, X_t^0, X_{t+1}^0, X_{t-1}^1, X_t^1, X_{t+1}^1\ ). \tag{33}$$

The component $P\left(X_{t-1}^0, X_t^0, X_{t+1}^0, X_{t-1}^1, X_t^1, X_{t+1}^1\right)$ can be decomposed as:

$$P\left(X_{t-1}^0, X_t^0, X_{t+1}^0, X_{t-1}^1\ , X_t^1, X_{t+1}^1\ \right) \propto \mathrm{P}(\mathrm{X}_{t+1}^0|\mathrm{X}_\mathrm{t}^0) \times \mathrm{p}(\mathrm{X}_\mathrm{t}^0|\mathrm{X}_{\mathrm{t}-1}^0) \times \mathrm{P}(\mathrm{X}_{t+1}^1|\mathrm{X}_\mathrm{t}^0, \mathrm{X}_{\mathrm{t}+1}^0, \mathrm{X}_\mathrm{t}^1) \times \mathrm{P}\left(\mathrm{X}_\mathrm{t}^1 \middle| \mathrm{X}_{\mathrm{t}-1}^0, \mathrm{X}_\mathrm{t}^0, \mathrm{X}_{t-1}^1\right) \tag{34}$$

With the decomposition above, the posterior distribution in equation (3) has the expression:

$$\begin{gathered} P\left(X_\mathrm{t}^0 \middle| X_{\mathrm{t}-1}^0, X_{\mathrm{t}+1}^0, Y_t, X_t^1, X_{\mathrm{t}-1}^1, X_{\mathrm{t}+1}^1\right) \propto \mathrm{e}^{-0.5\left(\mathrm{Y_t}+V_0-X_t^0V_d-X_t^1V_b\right)'\Sigma^{-1}\left(\mathrm{Y_t}+V_0-X_t^0V_d-X_t^1V_b\right)} \times P_X\left(\Delta, X_{t+1}^0 \middle| X_t^0, O_{dd}, O_d^0\right) \times \\ P_X\left(\Delta, X_t^0 \middle| X_{t-1}^0, O_{dd}, O_d^0\right) \times \frac{e^{-\frac{\left(X_t^1-e^{O_{bb}\Delta}X_t^1-m_1\right)^2}{2\,\sigma_1^2}}}{\sqrt{\sigma_1^2}} \times \frac{e^{-\frac{\left(X_t^1-e^{O_{bb}\Delta}X_{t-1}^1-m_0\right)^2}{2\,\sigma_{01}^2}}}{\sqrt{\sigma_0^2}}. \end{gathered} \tag{35}$$

Here $P_X(\Delta, x|x_0, O_{dd}, O_d^0)$ denotes the transition density of the restricted state variable:

$$P_X\left(\Delta, x \middle| x_0, O_{dd}, O_d^0\right) = c \times e^{-v-u}\left(\frac{v}{u}\right)^{\frac{q}{2}} I_q\left(2(u \times v)^{0.5}\right), \tag{36}$$

with $q = 2O_d^0 - 1, c = \frac{-2O_{dd}}{1-e^{O_{dd}\Delta}}, u = cx_0e^{O_{dd}\Delta}$ and $v = cx$. $I_q$ is the modified Bessel function of the first kind of order $q$. The terms $(m_0, \sigma_0^2)$ and $(m_1, \sigma_1^2)$ have the following expressions:

$$\begin{aligned} m_0 &= \left[O_{bd}X_t^0 + O_0^b + e^{O_{bb}\Delta}(O_{bd}X_{t-1}^0 + O_0^b)\right]\frac{\Delta}{2}, \\ \sigma_0^2 &= \left[C_{bd}X_t^0 + O_0^b + e^{2\times O_{bb}\Delta}(O_{bd}X_{t-1}^0 + O_0^b)\right]\frac{\Delta}{2}, \\ m_1 &= \left[O_{bd}X_{t+1}^0 + O_0^b + e^{O_{bb}\Delta}(O_{bd}X_t^0 + O_0^b)\right]\frac{\Delta}{2}, \\ \sigma_1^2 &= \left[C_{bd}X_{t+1}^0 + O_0^b + e^{2\times O_{bb}\Delta}(O_{bd}X_t^0 + O_0^b)\right]\frac{\Delta}{2}. \end{aligned} \tag{37}$$

Now we are ready to choose the candidate density to sample the posterior distribution of equation (3). First, conditioning on $Y_t$, $X_t^0$ follows the distribution:

$$X_t^0|Y_t \propto (U_t, V_t), \tag{38}$$

with

$$U_t = V_t^{-1}\left(V_d^{'}\Sigma^{-1}(Y_t + V_0 - X_t^0V_d - X_t^1V_b)\right), V_t^{-1} = V_d^{'}\Sigma^{-1}V_d. \tag{39}$$

Conditioning on $X_{t-1}^0$, the distribution $P(X_t^0|X_{t-1}^0)$ is approximately normally distributed as $N(U_t^1, V_t^1)$ with:

$$U_t^1 = X_{t-1}^0 + \left(O_{dd}X_{t-1}^0 + O_d^0\right)\Delta, V_t^1 = X_{t-1}^0\Delta. \tag{40}$$

Conditioning on $X_{t+1}^0$, the $X_t^0$ can be approximated to be normal $N(U_t^2, V_t^2)$, with

(41)

$$U_t^2 = \frac{X_{t+1}^0 - O_d^0\Delta}{1 + O_{dd}\Delta}, V_t^2 = \frac{X_{t+1}^0\Delta}{(1 + O_{dd}\Delta)^2}.$$

Combining the three normal distributions, we can propose the candidate density that approximates the target density in equation (3) closely. Denote $q1(\cdot)$ and $q2(\cdot)$ as the candidate densities with the following definitions:

$$\tilde{q} \sim N\left(\frac{V_t^1V_t^2U_t + V_tV_t^2U_1 + V_tV_t^1U_2}{V_tV_t^1 + V_tV_t^2 + V_t^1V_t^2}, \frac{V_tV_t^1V_t^2}{V_tV_t^1 + V_tV_t^2 + V_t^1V_t^2}\right). \tag{42}$$

In the experiment where we set the parameters and the unrestricted state variables as known, and sample only the restricted states, the acceptance rate for the independent MH algorithm is around 99%, which indicates that the candidate density $\tilde{q}(\cdot)$ is indeed efficient.

### *3.4.2 Sampling $X^1$*

The last step in the Gibbs cycle is to sample:

$$P(X^1|X^0, \Phi^Q, \Phi^P, Y). \tag{43}$$

Without loss of generality, we consider the $A(1, 2)$ model. For $A(1, 3)$, the algorithm applies, and we only need to adjust the matrix calculations accordingly. Conditioning on $X1$, the state variable $X1$ can be solved by:

$$X_t^1 = e^{O_{bb}\Delta}X_{t-1}^1 + \int_{t-1}^{t} e^{O_{bb}(t-s)}\left[\left(O_{bd}X_s^0 + O_b^0\right)ds + \sqrt{c_1X_s^0 + 1}dW_s^Q\right]. \tag{44}$$

Conditioning on $X^0$, the integrals can be approximated as:

$$\begin{gathered} \int_{t-1}^{t} e^{O_{bb}(t-s)}\left(O_{bd}X_s^0 + O_b^0\right)ds = \left(O_{bd}X_t^0 + O_b^0 + e^{O_{bb}\Delta}(O_{bd}X_{t-1}^0 + O_b^0)\right)\frac{\Delta}{2}, \\ \int_{t-1}^{t} e^{O_{bb}(t-s)}\sqrt{C_{bd}X_s^0 + 1}\, dW_s^Q \sim \left(0, \left(C_{bd}X_t^0 + 1 + e^{O_{bd}\Delta}\left(c_1X_{t-1}^0 + 1\right)e^{O'_{bb}\Delta}\right)\frac{\Delta}{2}\right). \end{gathered} \tag{45}$$

Based on this approximation, $X_t^1$ resumes the linear Gaussian structure:

$$\begin{gathered} y_t^1 = X_t^1V_b + N(0, \Sigma), \\ X_t^1 = e^{O_{bb}\Delta}X_{t-1}^1 + D_t + N(0, V_t), \end{gathered} \tag{46}$$

with

$$\begin{gathered} y_t^1 = Y_t - X_t^0V_d + V_0, \\ D_t = \left(O_{bd}X_t^0 + O_b^0 + e^{O_{bb}\Delta}(O_{bd}X_{t-1}^0 + O_b^0)\right)\frac{\Delta}{2}, \\ V_t = \left(c_1X_t^0 + 1 + e^{O_{bb}\Delta}\left(c_1X_{t-1}^0 + 1\right)e^{O'_{bb}\Delta}\right)\frac{\Delta}{2}. \end{gathered} \tag{47}$$

The FFBS algorithm can be applied to sample $X^1$ in a block.

Table 3 reports the performance of the algorithm for $A(1,3)$ based on simulated data. The pricing errors, being bigger than the models with explicit states, are within good precision ranges. The log-likelihood value approaches the true maximal value. The increase in pricing errors comes from the fact that, in addition to the parameters, we have 1548 state variables for $A(1, 3)$ to sample. To access the efficiency of the candidate density for the restricted state variables, the acceptance rate is 60%-70% for all restricted states. This indicates that the candidate density we propose is efficient.

## 4 THE EMPIRICAL ANALYSIS

In this section, we apply the MCMC algorithms on two market data sets of different economic episodes and show that it can replicate the yield data and forecast future yield levels. Then we perform a model comparison analysis for each of the two data sets.

### 4.1 The Data

We use two sets of zero-coupon bond yield data. Data I consists of weekly observations of zero-coupon bond yield with a sample period 1989.03.31-2007.03.02. The data are like that used in Collin-Dufresne et al. (2008)[10]. The zero-coupon bond prices are constructed by bootstrapping the LIBOR rates with maturities 1, 3, 6, 9 and 12 months and the swap rates with maturities 2, 3, 5, 7 and 10 years. The 9-month LIBOR rates are not available for the period of 1989.03.31-1991.05.31. The linear interpolation method is applied to obtain the missing data.

Data II consists of zero-coupon bond yield data with maturities 1, 2, 3, 4, and 5 years from the CRSP monthly treasury file. The sample period is 1972.01.31-2010.12.31. We trace the data sets used in Ait-Sahalia and Kimmel (2010) and Cheridito et al. (2007) as closely as possible[3-4]. We can only obtain data with five maturities from CRSP. To have similar sample size, we use a longer sample period of 39 years.

The two yield data sets cover different economic episodes and exhibit different empirical features. During the sample period of Data I, Piazzesi (2008) suggests that there is evidence of "a return to normality" and constant conditional second moments seem to be enough to describe the state dynamics[6]. Data II covers the two regimes of extreme volatility movements: the oil price shock in 1974 and monetary experiment in 1979-1982. The conditional second moments of yields exhibit peaks corresponding to these two regimes. We exam these two empirical facts from a Bayesian perspective. For Data I, $A(1,3)$ dominates the other model. For Data II, $A(0,3)$ model outperforms better.

### 4.2 The In-Sample Fitting Performance

We study the in-sample fitting and forecasting performances of the MCMC algorithms applied on Data I. Parameters and latent states inferred from the MCMC algorithms can be used to replicate the in-sample yield data closely. Table 4 reports the in-sample fitting performance for Data I with $A(0,3)$ and $A(1,3)$ models.

**Table 4** In-Sample Fitting Performances-Data I

| | $A(0,3)$ | | $A(1,3)$ | |
|---|---|---|---|---|
| | Mean | RMSE | Mean | RMSE |
| 1-Month | 0.42 | 5.83 | -0.10 | 5.99 |
| 3-Month | 1.13 | 4.98 | 0.69 | 4.52 |
| 6-Month | -0.08 | 4.54 | 0.18 | 4.40 |
| 9-Month | -1.28 | 4.86 | -0.36 | 4.60 |
| 12-Month | -2.19 | 6.01 | -0.88 | 4.92 |
| 2-Year | 1.29 | 7.32 | 2.33 | 6.01 |
| 3-Year | 1.64 | 6.64 | 1.62 | 6.03 |
| 4-Year | 0.18 | 6.06 | -0.44 | 5.56 |
| 5-Year | 0.69 | 5.95 | 0.04 | 5.78 |
| 7-Year | -0.50 | 6.68 | 0.26 | 6.01 |
| 10-Year | -0.56 | 8.29 | 0.86 | 6.92 |

*Note: This table reports the in-sample fitting performances of the $A(0,3)$ and $A(1,3)$ models for Data I: weekly observations of zero-coupon bond yield data of maturities 1, 3, 6, 9 and 12 months, 2, 3, 4, 5, 7, and 10 years from 1989.03.31 – 2007.03.02.

The in-sample fitting performance is evaluated by two measures: the average pricing errors and RMSEs for each maturity. To compare the RMSEs, we use the results reported in Table III and V in Collin-Dufresne et al. (2008) as a benchmark[10]. Data I is similar to the data used in their paper. It turns out that for both $A(0,3)$ and $A(1,3)$ models, our in-sample pricing errors are smaller than those from their model-free estimation method for all maturities.

For the in-sample fitting performance, the PCA is a benchmark hard to beat. We provide a comparison between MCMC with the PCA decomposition. As suggested by the fitting performance of the models in Data I, we choose $A(1,3)$ model for this comparison. We also assume that the measurement error is uncorrelated across maturities. We could gain more precision in in-sample fitting. As shown in Table 5, the in-sample errors are marginally bigger than but comparable to that implied by the PCA.

**Table 5** In-Sample Pricing Errors, $A(1,3) -$ Data I

| | PCA | $A(1,3)$ | | |
|---|---|---|---|---|
| | In-sample fit | In-sample fit | Out-sample forecast | In-sample forecast |
| | | | RMSE | |
| 1-Month | 2.96 | 8.71 | 2.07 | 7.32 |
| 3-Month | 2.96 | 4.61 | 3.34 | 6.19 |
| 6-Month | 2.01 | 0.63 | 5.10 | 5.28 |
| 9-Month | 2.12 | 2.53 | 5.53 | 4.96 |
| 1-Year | 2.00 | 3.93 | 5.11 | 5.18 |
| 2-Year | 1.83 | 1.92 | 6.62 | 6.77 |
| 3-Year | 2.01 | 1.49 | 6.63 | 6.50 |
| 4-Year | 1.44 | 1.06 | 8.00 | 5.78 |
| 5-Year | 1.31 | 1.22 | 7.71 | 5.73 |
| 7-Year | 1.11 | 1.25 | 6.88 | 5.82 |
| 10-Year | 1.28 | 3.94 | 4.99 | 8.25 |

*Note: This table reports the in-sample fitting and forecast performances of the $A(1,3)$ model for Data I.

We also conduct an in-sample forecast analysis. We apply the MCMC algorithms on yield data with nine maturities from Data I, while leaving out yields of maturities 2 and 10 years. With the inferred parameters and latent states, we reconstruct the yield data. In addition, we "interpolate" the yield of 2 years, and "extrapolate" the yield of 10 years. Results are reported in Table 4. As we use 9 maturities compared with the previously 11 maturities, the smaller sample size results in bigger pricing errors. The pricing errors for yields of 2 and 10 years are marginally larger compared with the RMSEs of other maturities. Other yield construction methods, such as the Cubic spline and the Nelson-Siegel families, are often used in the industry. These methods, together with PCA, fit the yield data in the cross-section but fail to consistently provide fitting in the time-series. Furthermore, these methods do not rule out arbitrage opportunities. By contrast, the MCMC method consistently fits yield data in both time-series and cross-section and satisfies the no-arbitrage condition. In addition, the probabilistic characterization of parameters and latent states in MCMC allows us to make a Bayesian forecast about future yield levels.

At last, we show that our algorithm can construct the short rates closely. The $A(0, n)$ and $A(1, n)$ models have the problem that the implied short rates can be negative. Furthermore, the fitting errors for short-end yields from ATSMs are usually big for the reasons of seasonality and microstructure noises [10]. As reported in Collin-Dufresne et al. (2008), the RMSEs for 1-month yield are around 14 basis points for both $A(0, 3)$ and $A(1, 3)$ models, even though the 3-month yield data are used for inversion[10]. We use the inferred parameters and latent states to construct the short rates, and compare it with the 1-month yield, which is often used as a proxy for short rates. The inferred short rates are positive and closely resemble the 1-month yield data with a correlation of 99.75% and the in-sample RMSE of 5.99 basis points.

In conclusion, the MCMC algorithms deliver good in-sample fitting and in-sample forecast performances. They also construct the short rates that closely resemble the short-end yield data. In this next section, we explore the out-of-sample forecast performance of the MCMC algorithms.

### 4.3 The Out-of-Sample Forecast Performance

In this section, we exam the forecast performance of each model for Data I. Denote $Y_t$ as the yield data observations up to time t. As current yield data reflect the market's expectation of future, we want to forecast the future yield level given current information of $Y_t$. We accomplish this task by characterizing the conditional expectation of $E(Y_T|Y_t)$ with marginalization of parameters and latent states:

$$E(Y_T|Y_t) = \int\int Y_T P(\Phi^Q, X|Y_t) dX d\Phi^Q . \tag{48}$$

We sample $\{\Phi^Q, X\}$ from the posterior distribution $P(\Phi^Q, X|Yt)$, simulate $Y_T$, and apply to Monte Carlo integration to obtain $E(Y_T|Y_t)$. Here we see the power of the MCMC algorithms. They allow us to sample the high-dimensional posterior density $P(\Phi^Q, X|Y_t)$ efficiently.

Table 6 reports the forecast performances of $A(0, 3)$ and $A(1, 3)$ models for Data I for a prediction period of 12 weeks. The $A(1, 3)$ model has the best forecast performance. The maximal mean error is 4.28 basis points and the maximal RMSE is 4.98 basis points.

**Table 6** Forecast Performances and the Model Evidence for Data I

| | $A(0,3)$ | | $A(1,3)$ | |
|---|---|---|---|---|
| | Forecast (bps) | | | |
| | Mean | RMSE | Mean | RMSE |
| 1-Month | 0.67 | 0.72 | -2.27 | 2.61 |
| 3-Month | -0.21 | 0.75 | -4.18 | 4.42 |
| 6-Month | -0.26 | 1.97 | -4.28 | 4.75 |
| 9-Month | -0.02 | 3.05 | -3.04 | 4.12 |
| 1-Year | 0.45 | 3.82 | -1.04 | 3.47 |
| 2-Year | -4.82 | 6.65 | 0.27 | 4.09 |
| 3-Year | -6.98 | 8.33 | 1.51 | 4.57 |
| 4-Year | -8.68 | 9.69 | -1.02 | 4.29 |
| 5-Year | -7.16 | 8.29 | -0.10 | 4.30 |
| 7-Year | -3.52 | 5.28 | -1.56 | 4.45 |
| 10-Year | 1.01 | 3.86 | -2.92 | 4.98 |
| | Model Comparison | | | |
| Model Evidence | 68177.4 | | 68922.0 | |
| AIC | 117698.0 | | 132501.0 | |
| BIC | 93173.0 | | 111580.0 | |

*Note: The first panel reports the out-of-sample forecast performances of the $A(0, 3)$ and $A(1, 3)$ models for Data I. The second panel reports the logarithm of the Bayesian model evidence and information criteria: AIC and BIC.

We also compare the Bayesian forecast performance with three alternative methods: the RW method, the OLS method, and the frequentist method. The first two methods are often used as benchmarks in evaluating the forecast performances[2,30]. The RW method uses last observations of in-sample data as the forecast. The OLS uses the linear regression to forecast. The dependent variable is the difference between the future and current yield data, and the regressor is the difference between the 5-year and 3-month yield data. Note the choice of the 5-year and 3-month is merely for being the same with Duffee (2002)[2]. Yield forecasts are computed with the simulated future states and parameters.

Table 7 compares the forecast performances of these methods. The random walk method has the best performance for yields with short-term maturities ($\tau < 1$), whereas the Bayesian forecast with $A(1, 3)$ renders the best performance for longer maturities ($\tau \geq 1$). The gain of the Bayesian forecast with $A(1, 3)$ is substantial. The RMSEs is around 60% to those of the random walk approach, which is the second-best method. This improvement is larger compared with the improvement of the arbitrage-free Nelson-Siegel forecast over the random walk method (Table 5 in Christensen et al. (2011))[30]. It is also larger compared with that of the "essentially affine" model forecast over the random walk method (Table VIII in Duffee (2002))[2]. The reason that the random walk performs best in the short end is that the short-end yields are extremely flat in the prediction period.

**Table 7** Forecast Performances-Data I

| | Forecast (bps) | | | | | |
|---|---|---|---|---|---|---|
| | RW | OLS | Freq. $A(0,3)$ | Freq. $A(1,3)$ | Bay. $A(0,3)$ | Bay. $A(1,3)$ |
| 1-Month | 0.00 | 4.53 | 4.81 | 8.80 | 7.17 | 2.61 |
| 3-Month | 0.40 | 5.59 | 3.09 | 5.90 | 7.52 | 4.42 |
| 6-Month | 1.56 | 6.86 | 5.62 | 4.61 | 1.97 | 4.75 |
| 9-Month | 2.84 | 7.78 | 7.80 | 5.02 | 3.05 | 4.12 |
| 1-Year | 3.49 | 7.98 | 9.74 | 6.25 | 3.82 | 3.47 |
| 2-Year | 7.06 | 9.16 | 21.68 | 6.42 | 6.65 | 4.09 |
| 3-Year | 7.35 | 8.18 | 29.78 | 7.29 | 8.34 | 4.57 |
| 4-Year | 7.91 | 8.16 | 36.49 | 6.17 | 9.70 | 4.29 |
| 5-Year | 7.80 | 7.41 | 40.02 | 6.29 | 8.29 | 4.30 |
| 7-Year | 8.00 | 7.31 | 43.66 | 5.92 | 5.28 | 4.45 |
| 10-Year | 8.39 | 7.48 | 44.50 | 5.67 | 3.86 | 4.98 |

*Note: This table reports the out-of-sample forecast performances of several methods for Data I.

The comparison between the frequentist and Bayesian forecast methods reveals the advantage of the latter. As reported in Table 7, the frequentist forecast with $A(0, 3)$ performs worst, whereas the frequentist forecast with $A(1, 3)$ dominates the Bayesian forecast with $A(0,3)$ for some maturities. However, this pattern is not persistent. With different simulated trajectories of future states, the forecast performance by the frequentist forecast method changes. The Bayesian forecast method is advantageous in that it averages the future yield level over simulated state paths and parameters. We find that $A(1, 3)$ model has a better out-of-sample forecast performance compared with the $A(0, 3)$ model. This is in contrast with Duffee (2002) who finds that $A(0, 3)$ is better[2]. A question arises as how to choose the best model that delivers good in-sample fitting and out-of-sample forecast performances. In the next section, we exam the model comparison from a Bayesian perspective.

### 4.4 The Model Comparison

A full treatment of Bayesian inference includes a model-comparison analysis. We evaluate the relative performance of each model using the measure of the model evidence $Z = P(Y|M)$. This quantity is approximated by the harmonic mean of likelihood values[25]:

$$Z \approx \left( \frac{1}{N} \Sigma_{i=1}^{\mathrm{N}} \frac{1}{P\left(Y \middle| X_i, \Phi_i^Q\right)} \right)^{-1}, \tag{49}$$

where $N$ is the number of simulations and the pair $\{X_i, \Phi_i^Q\}$ is the i-th simulated parameters and latent states from the posterior distribution. The efficient MCMC algorithms for posterior densities make this feasible.

Table 6 reports the logarithm of the model evidence for each of the two models examined for Data I. The ranking is coherent with both the in- sample fitting and out-of-sample forecast performances of each model.

For Data I, there is a dominance of $A(1, n)$ over $A(0, n)$. The restricted state variables introduce stochastic volatilities and induce fat tails to the yield distribution. For Data II, model ranking is different from that of Data I[31-32]. As shown in Table 8, the model evidence suggests a ranking of $A(0, 3), A(1, 3)$ in descending order.

**Table 8** Fitting & Forecast Performances and the Model Evidence for Data II

| | $A(0,3)$ | | $A(1,3)$ | |
|---|---|---|---|---|
| | Fit | Forecast (3m) | Fit | Forecast (3m) |
| | RMSE (bps) | | | |
| 1-Year | 20.75 | 12.46 | 20.80 | 21.27 |
| 2-Year | 16.23 | 7.62 | 19.11 | 20.44 |
| 3-Year | 15.90 | 13.92 | 17.44 | 9.52 |
| 4-Year | 16.29 | 21.57 | 17.70 | 22.25 |
| 5-Year | 15.49 | 26.04 | 19.50 | 45.75 |
| | Model Comparison | | | |
| Model Evidence | 13837.0 | | 13688.0 | |
| AIC | 11225.1 | | 10966.0 | |

| | | |
|---|---|---|
| BIC | 3002.8 | 2701.0 |

*Note: The first panel reports the in-sample fitting and out-of-sample forecast performances or Data II. The second panel reports the logarithm of the Bayesian model evidence and information criteria: AIC and BIC.

The reason could be, for the $A(1,3)$ model, the two unrestricted state variables can support large negative correlations and the unrestricted state variables can support the non-normality property. However, $A(0,3)$ cannot support the non-normality property.

## 5 THE CONLUSION

We develop MCMC algorithms to conduct a Bayesian inference analysis for multi- factor term structure models. We apply the algorithms on two market data sets with different regimes. The in-sample pricing errors are smaller than those in the literature with the similar sample. We also conduct a Bayesian forecast analysis on future yield levels. With the $A(1,3)$ model, the Bayesian forecast performance dominates the OLS forecast and frequentist forecast approaches for all maturities. It also dominates the random walk forecast for maturities greater or equal to one year. We study the Bayesian model comparison for the two market data sets. The model evidence delivers a ranking consistent with the in-sample fitting and out-of-sample forecast performances for each model. Data I supports the non-normality of the yield change distribution and a humped shape of yield volatility. The model evidence ranks the models as $A(1,3), A(0,3)$ in descending order. Data II supports the non-normality feature but demands a strong correlation between state variables. As a result, the model ranking is $A(0,3)$ and $A(1,3)$ in descending order.

Future research can explore the MCMC algorithms of the $A(m,n)$ model for $m \geq 2$. Also, it is interesting to analyze the model comparison across different market price of risk specifications.

## COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

## REFERENCES

[1] Q Dai, KJ Singleton. Specification analysis of affine term structure models. Journal of Finance, 2000, 55(5): 1943–1978.

[2] GR Duffee. Term premia and interest rate forecasts in affine models. Journal of Finance, 2002, 57(1): 405–443, .

[3] P Cheridito, D Filipovic, RL Kimmel. Market price of risk specifications for affine models: Theory and evidence. Journal of Financial Economics, 2007, 83(1): 123–170.

[4] Y Ait-Sahalia, RL Kimmel. Estimating affine multi factor term structure model using closed-form likelihood expansions. Journal of Financial Economics, 2010, 98(1): 113–144.

[5] JD Hamilton, JC Wu. Identification and estimation of Gaussian affine term structure models. Journal of Econometrics, 2012, 168(2): 315–331.

[6] M Piazzesi. Affine term structure models. In Handbook of Financial Econometrics. Elsevier, 2008.

[7] AR Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scandinavian Journal of Statistics, 1995, 22(1): 55–71.

[8] MW Brandt, P Santa-Clara. Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. Journal of Financial Economics, 2002, 63(2): 161–210.

[9] RR Chen, L Scott. Multi-factor Cox-Ingersoll-Ross models of the term structure: Estimates and tests from a Kalman filter model. Journal of Fixed Income, 1993, 3 (3): 14–31.

[10] P Collin-Dufresne, RS Goldstein, CS Jones. Identification of maximal affine term structure models. Journal of Finance, 2008, 63(2)743–795.

[11] Nelson-Siegel term structure models. Journal of Econometrics, 2011, 164(1): 4–20.

[12] LEO Svensson. Estimating and interpreting forward interest rates: Sweden 1992-1994. NBER Working Papers 4871, National Bureau of Economic Research, Inc., 1994.

[13] CM Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 2nd edition, 2007.

[14] E Jacquier, NG Polson, PE Rossi. Bayesian analysis of stochastic volatility models. Journal of Business and Economic Statistics, 1994, 12(4): 371–389.

[15] E Jacquier, NG Polson, PE Rossi. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. Journal of Econometrics, 2004, 122(1): 185– 212.

[16] B Eraker, M Johannes, N Polson. The impact of jumps in volatility and returns. Journal of Finance, 2003, 58(3): 1269–1300.

[17] M Johannes, N Polson. MCMC methods for continuous-time financial econometrics. In Handbook of Financial Econometrics. Elsevier, 2007.

[18] H Hu. Markov chain Monte Carlo estimation of multi-factor affine term-structure models. Unpublished doctoral dissertation, University of California, Los Angeles, 2005.

[19] CP Robert, G Casella. Monte Carlo Statistical Methods. Springer, 2nd edition, 2004.

[20] S Geman, D Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Ma- chine Intelligence, 1984, 6(6): 721–741.

[21] JM Hammersley, PE Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

[22] CK Carter, R Kohn. On Gibbs sampling for state space models. Biometrika, 1994, 81(3): 541–553.

[23] S Fruwirth-Schnatter. Data augmentation and dynamic linear models. Journal of Time Series Analysis, 1994, 15(2): 183–202.

[24] RE Kalman. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering, 1960, 82(1): 35–45.

[25] RE Kass, AE Raftery. Bayes factors. Journal of the American Statistical Association, 1995, 90(430): 773–795.

[26] L Tierney, JB Kadane. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 1986, 81(393): 82–86.

[27] MA Newton, AE Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society. Series B (Methodological), 1994, 56(1): 3–48.

[28] D Duffie, J Pan, KJ Singleton. Transform analysis and asset pricing for affine jump-diffusions. Econometrica, 2000, 68(6): 1343–1376.

[29] N Shephard, S Kim. Bayesian analysis of stochastic volatility models: Comment. Journal of Business and Economic Statistics, 1994, 12(4): 406–410.

[30] JHE Christensen, FX Diebold, GD Rudebusch. The affine arbitrage-free class of C.R. Nelson and A.F. Siegel. Parsimonious modeling of yield curves. Journal of Business, 1987, 60(4): 473–489.

[31] CP Robert, G Casella. Introducing Monte Carlo Methods with R. Springer Verlag, 2009.

[32] D Duffie, R Kan. A yield-factor model of interest rates. Mathematical Finance, 1996, 6(4): 379–406.