CReMFi
Centre for Research in Macroeconomics and Finance

ISSN: 2058-5160

Discussion Paper 5

Bayesian Vector Autoregressions with Non-Gaussian Shocks

Ching-Wai (Jeremy) Chiu (Bank of England)

Haroon Mumtaz (Queen Mary, University of London)

Gabor Pinter (Bank of England)

JEL classification: C11, C32, C52

Keywords: Bayesian VAR, Non-Gaussian shocks, Density Forecasting

Bayesian Vector Autoregressions with Non-Gaussian Shocks*

Ching-Wai (Jeremy) Chiu† Haroon Mumtaz‡ Gabor Pinter§

July 5, 2016

Abstract

This paper proposes a Bayesian Vector Autoregression where the orthogonalised shocks are assumed to be non-Gaussian. A Gibbs sampling algorithm is provided to approximate the posterior distribution of the model parameters. An application to a model of the yield curve suggests that there is ample evidence against the assumption of normal shocks. The proposed model provides notable improvements both in terms of in-sample fit and out of sample forecasting.

JEL classification: C11, C32, C52
Keywords: Bayesian VAR, Non-Gaussian shocks, Density Forecasting

1 Introduction

The last three decades have been characterised by the evolving nature of economic shocks and changing dynamics of macroeconomic and financial variables in the OECD. The 1970s and the early 1980s were the decade of volatile shocks and high inflation. The 'Great Moderation' followed in the mid-1980s and the early 1990s with most of these economies enjoying low inflation and stable GDP growth with large adverse shocks mostly absent. This period was disrupted in 2007 when a large negative shock lead to a severe contraction and asset price volatility.

Bayesian Vector Autoregressions (BVARs) are the model of choice for many researchers when analysing and forecasting this type of data. However, BVARs incorporate shocks drawn from a Gaussian distribution. As argued forcefully in Curdia *et al.* (2013), the assumption of normal disturbances cannot account for extreme, volatile events such as the 'Great Recession' seen in the post-2007 period. Moreover, this assumption rules out the possibility of shocks that originate from a skewed distribution. The recent movement of macroeconomic and financial data has shown that this conclusion may not be true. Short term interest rates provide a key recent example – these have been largely constant over the past five years implying a distribution for conventional monetary policy shocks that is not symmetric.

In this paper we introduce a BVAR model that allows for the possibility of non-Gaussian shocks. The non-Gaussianity is introduced through a Markov mixture of normals that is applied independently to each orthogonal residual of the VAR model. The specification is general – it allows for departures from normality ranging from fat tails to skewness and excess kurtosis. The

*We are grateful for useful comments from Katerina Petrova and participants of the workshop: 'Time-Variation and Non-Linear Models in Econometrics and Macroeconomics' at the Bank of England. The views expressed in this paper are those of the authors and should not be held to represent those of the Bank of England. All errors are our own

†Bank of England, e-mail: jeremy.chiu@bankofengland.co.uk
‡Corresponding author. Queen Mary University of London, e-mail: h.mumtaz@qmul.ac.uk
§Bank of England, e-mail: gabor.pinter@bankofengland.co.uk

specification is also flexible as the degree of non-normality can differ across the residuals of the VAR model.

The paper builds on the recent contributions in Curdia *et al.* (2013), Chib and Ramamurthy (2014) and Chiu *et al.* (2014). Curdia *et al.* (2013) and Chib and Ramamurthy (2014) introduce student-T shocks in DSGE models while Chiu *et al.* (2014) apply the same extension to BVARs (see also Clark and Ravazzolo (2015) and Chan (2015)). The present paper generalises this approach by allowing for more general forms of non-normality and builds on the mixture of two normals used in the seminal paper of Sims (1993). The paper is also closely related to Kalliovirta *et al.* (2016) who introduce a frequentist VAR model where the distribution of the vector of endogenous variables is a mixture of *multivariate* normals.[1] The BVAR proposed in the current paper differs from this contribution in at least four dimensions. First, we introduce non-normality in the orthogonalised shocks of the VAR model rather than the endogenous variables directly. These VAR shocks are proxies for underlying structural disturbances and our model allows them to be drawn from a non-Gaussian distribution. Second, our proposed model allows the degree of non-normality to differ across the VAR shocks as the mixing weights are defined independently for each equation. In other words, our specification accounts for the possibility that shocks to some equations may be more or less non-Gaussian than others. In contrast, Kalliovirta *et al.* (2016) define the mixtures jointly for all variables included in the model. This assumption may be too restrictive for models that contain a mix of macroeconomic and financial variables. Thirdly, in contrast to Kalliovirta *et al.* (2016), regimes (or components of the mixture) in our model follow a Markov process and can, therefore, be persistent. As discussed below, our model can be extended to allow the transition probabilities for each latent state to be a function of relevant covariates. This feature implies that our specification also incorporates the possibility of data-driven regime switches as in Dueker *et al.* (2011) and Kalliovirta *et al.* (2016).[2] Finally, following Villani *et al.* (2009), we adopt a Bayesian approach. Within our framework, calculation of predictive densities requires a trivial extension of the estimation algorithm. Similarly, the selection of the number of components can be carried out in a coherent manner using marginal data densities.

We apply the proposed BVAR to model the dynamics of the US yield curve. The empirical analysis suggests strong evidence to support the view that shocks to the level, slope and curvature of the yield curve are drawn from a non-Gaussian distribution. A recursive forecast experiment finds that allowing for non-Gaussian shocks can lead to substantial gains in point and density forecasting of yields relative to the standard BVAR model.

The paper is organised as follows. Section 2 presents the proposed model in detail. Section 3 presents the details of the estimation algorithm and discusses selection of components. We present some Monte-Carlo evidence on the performance of the estimation algorithm in Section 4. Finally, Section 5 uses the proposed BVAR to model and forecast the US yield curve.

2 BVAR with non-normal disturbances

The proposed BVAR model is defined as follows:

$$y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + u_t \qquad t = 1, \ldots, T. \tag{1}$$

[1] Lanne and Lütkepohl (2010) introduce a structural VAR where the residuals follow a mixture of two normals. This specification is used to identify the structural shocks of the VAR model.

[2] Lanne (2006) and Bec *et al.* (2008) also introduce normal mixture models where the mixing probability depends on the level of past data.

where y_t is an $n \times 1$ vector of observed endogenous variables; B_i, $i = 1, \ldots, p$ are $n \times n$ matrices of coefficients; u_t are heteroscedastic shocks associated with the VAR equations.

The orthogonalised shocks of the model are given as:

$$e_t = Au_t \tag{2}$$

where A is a lower triangular $n \times n$ matrix. The orthogonal shock to the ith equation of the VAR is assumed to follow:

$$e_{it} = \alpha_{i,S_{it}} + \sigma_{i,S_{it}}\varepsilon_{it}, \varepsilon_{it} \sim N(0,1) \tag{3}$$

where $S_{it} = 1, 2, ...M$ denotes the unobserved components or regimes. As explained in Koop (2003) and Geweke (2005), the formulation in equation 3, describes a mixture of M distributions where each component is $N\left(\alpha_i, \sigma_i^2\right)$. The state variable S_{it} determines the component that is active at a particular point in time. The law of motion for S_{it} is chosen to be first order Markov process with transition probabilities

$$P_i\left(S_{i,t} = J | S_{i,t-1} = I\right) = p_{i,IJ} \tag{4}$$

Note that the transition probabilities can be constant or one can assume that they depend on a set of regressors z_t and evolve over time (see Filardo and Gordon (1998)):

$$P_i\left(S_{i,t} = J | S_{i,t-1} = I\right) = p_{i,IJ}\left(z_t\right)$$

This Markov formulation captures possible persistence in the regimes but allows for the possibility of rapid transitions across components and regime switches that are data-driven.

The specification in equations 3 implies that orthogonalised residuals e_t are non-Gaussian. As the number of components increase, the specification can potentially capture features of the distribution that are very different from the normal distribution. For example, if the means α_i vary across regimes then the distribution can exhibit skewness and have kurtosis less than 3, the value for the normal distribution. If the means are the same across components, the model is then a scale mixture of normals. The resulting distribution is symmetric but may have fatter tails than the normal distribution. In fact, as shown by Geweke (1993) assuming that $e_{it} = \sigma_{i,t}\varepsilon_{it}$ and adopting a Gamma prior for $\frac{1}{\sigma_{i,t}}$ of the form $p\left(\frac{1}{\sigma_{i,t}}\right) = \prod_{t=1}^{T} \Gamma\left(1, v_i\right)$ is equivalent to a specification that assumes a Student-t distribution for e_{it} with v_i degrees of freedom. We employ the specification with Student-t errors as a competing model in the forecast comparison below.

The VAR model proposed above can also be interpreted as a Markov Switching VAR model (see Hamilton (1994), Sims *et al.* (2008)). This is easily seen in a bi-variate version of the BVAR with one lag:

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{5}$$

where

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ A_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{6}$$

and

$$\begin{aligned} e_{1t} &= \alpha_{1,S_{1t}} + \sigma_{1,S_{1t}}\varepsilon_{1t} \\ e_{2t} &= \alpha_{2,S_{2t}} + \sigma_{2,S_{2t}}\varepsilon_{2t} \end{aligned} \tag{7}$$

where $var\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} = I_2$. Using equation 7, 6 and 5 one obtains:

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ A_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha_{1,S_{1t}} \\ \alpha_{2,S_{2t}} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ A_{21} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{1,S_{1t}} & 0 \\ 0 & \sigma_{2,S_{2t}} \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{8}$$

The VAR model in equation 8 has switching intercepts $\begin{pmatrix} \alpha_{1,S_{1t}} \\ \alpha_{2,S_{2t}} - A_{21}\alpha_{1,S_{1t}} \end{pmatrix}$ and reduced form residuals with a switching covariance matrix $\begin{pmatrix} \sigma^2_{1,S_{1t}} & -A_{21}\sigma^2_{1,S_{1t}} \\ -A_{21}\sigma^2_{1,S_{1t}} & \sigma^2_{1,S_t}A^2_{21} + \sigma^2_{2,S_{2t}} \end{pmatrix}$.

The reduced form residuals in the model are a linear combination of the non-normal orthogonal shocks: $\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} = \begin{pmatrix} e_{1t} \\ e_{2t} - A_{21}e_{1t} \end{pmatrix}$. Therefore the setup imparts a flexible specification for u_{it} which can also depart from Gaussianity in interesting ways.

Note that unlike standard MSVAR models, there are n independent Markov chains in the proposed model that govern the behaviour of each orthogonal error. As noted above, the implied reduced form intercepts and residuals are a combination of the parameters of the process for the orthogonal shocks and this implies a more complex structure than standard MSVARs where, typically, one Markov process controls the regime shifts in the system. For example in equation 8, the intercept and the error variance in the second equation depend on both S_{1t} and S_{2t}. In other words, regime switches in X_t depend on the distributional properties of *both* orthogonal shocks.

This example also makes it clear that the ordering of the variables can matter for the interpretation of the reduced form intercepts and residual variance. The main advantage of assuming a recursive structure for the A matrix lies in the fact that it greatly simplifies the estimation algorithm. In empirical applications, one can order the variables in an economically meaningful manner and check if the results of interest are robust to this choice. Such an approach is used frequently in studies that employ VARs featuring stochastic volatility in the residuals orthogonalised using a lower triangular A matrix (see for example Cogley and Sargent (2005)).

3 Estimation

We adopt a Bayesian approach to model estimation and forecasting. In this section we describe the prior distributions and the MCMC algorithm used to obtain the posterior distribution of the parameters.

3.1 Priors

Following Banbura *et al.* (2010), the prior for the VAR coefficients $B = vec([B_1, B_2, ..B_p])$ is normal and introduced via dummy observations. The prior is defined as $p(B) \sim N(B_0, S_0)$, where $B_0 = (x_d'x_d)^{-1}(x_d'y_d)$ and $S_0 = (y_D - x_D b_0)'(y_D - x_D b_0) \otimes (x_d'x_d)^{-1}$. The dummy observations y_D and x_D that are defined as:

$$y_D = \begin{bmatrix} \frac{diag(\gamma_1 s_1 ... \gamma_n s_n)}{\tau} \\ 0_{n\times(p-1)\times n} \end{bmatrix}, \qquad x_D = \begin{bmatrix} \frac{J_P \otimes diag(s_1 ... s_n)}{\tau} \end{bmatrix} \tag{9}$$

where γ_1 to γ_n denote the prior mean for the parameters on the first lag obtained by estimating individual AR(1) regressions, τ measures the tightness of the prior on the VAR coefficients and

$J_p = diag([1, 2, ..p])$. The scaling factor s_i are set using the standard deviation of the residuals from the individual AR(1) equations. In addition, we introduce priors on the sum of lagged coefficients by defining the following dummy observations:

$$y_S = \frac{diag(\gamma_1\mu_1 \dots \gamma_n\mu_n)}{\lambda}, \qquad x_S = \left[\frac{(1_{1\times p}) \otimes diag(\gamma_1\mu_1 \dots \gamma_n\mu_n)}{\lambda} \right] \tag{10}$$

where μ_1 to μ_n denote the sample means of the endogenous variables. In our applications below, the prior tightness τ is set to 0.1, the value commonly used for US data. As in Banbura *et al.* (2010) we assume that $\lambda = 10\tau$.

The prior for the non-zero and non-one elements A_k is $P(A_k) \sim N(A_{0,k}, \Sigma_{0,k})$. In our applications, $A_{0,k}$ denotes the non-zero and non-one elements of $\tilde{A}_{ols}$ where $\tilde{A}_{ols}$ is the inverse of the Cholesky decomposition of the OLS estimate of the VAR error covariance with its diagonal elements normalised to 1. We assume that the variance of the prior for each element is $\Sigma_{0,k} = abs(A_{0,k}) \times 10$. Therefore this specification allows a large range of values for these parameters apriori.

The prior for α_i is assumed to be the same across regimes and VAR equations. The prior is normal and is given by $P(\alpha_i) \sim N(\alpha_0, v_0)$ where we set $\alpha_0 = 0$ and $v_0 = 100$ in our applications below. The prior for σ_i^2 in each regime is inverse Gamma: $P(\sigma_i^2) \sim IG(\sigma_0, v_0)$ where we use the scale parameter $\sigma_0 = 0.1$ and degrees of freedom $v_0 = 5$.

In our benchmark model, the transition probabilities are assumed to be fixed. In this case, the prior for $p_{i,IJ}$ is of the following form: $P(p_{i,IJ}) = D(u_{IJ})$where $D(.)$ denotes the Dirichlet distribution. In our empirical applications, we set $u_{IJ} = 15$ if $I = J$ and $u_{IJ} = 1$ if $I \neq J$. This prior thus places some weight on regimes that are persistent and implies apriori that the process stays in the current regime with a probability of about 93%. Note, we also consider versions of the model where the transition probabilities are time-varying. This extended model is described in section 3.2.1.

3.2 Gibbs sampling algorithm

The marginal posterior distributions are approximated via a Gibbs algorithm. This algorithm draws successively from the following conditional posterior distributions:

1. $G\left(B|S_{it}, \alpha_{i,S_{it}}, \sigma^2_{i,S_{it}}, A, \tilde{y}_T\right)$: Given the data $\tilde{y}_T = [y_1, y_2, .., y_T]$, regime dependent parameters, the latent states S_{it} and the A matrix, the model can be written as a VAR with (known) time-varying intercepts and heteroscedastic disturbances. As the regime dependent parameters are assumed to be observed, the model can be easily transformed into a standard time-invariant VAR model where the conditional posterior distribution of the VAR coefficients is linear and Gaussian: $N\left(B_{T|T}, P_{T|T}\right)$. We find that the Kalman filter offers a computationally stable method to calculate $B_{T|T}$ and $P_{T|T}$ while taking into account the time-varying intercepts and heteroscedasticity. In particular, we re-write the model in State-Space form:

$$\begin{aligned} Y_t &= X_t B_t' + \mu_t + R_t^{1/2} V_t \\ B_t &= B_{t-1} \end{aligned}$$

where $Y_t = vec(y_t)$, $X_t = I_n \otimes x_t$, $x_t = [y_{t-1}, y_{t-2}, ..., y_{t-p}]$, $\mu_t = A^{-1}\alpha$, $R_t = A^{-1} diag(\sigma^2) A^{-1\prime}$, $V_t = vec(\varepsilon_{it})$. Here α denotes $n \times 1$ vector: $\alpha = [\alpha_{1,S_{1t}}, \alpha_{2,S_{2t}}, ..\alpha_{n,S_{nt}}]$ and σ^2 is the $n \times 1$ vector: $\alpha = \left[\sigma^2_{1,S_{1t}}, \sigma^2_{2,S_{2t}}, ..\sigma^2_{n,S_{nt}}\right]$. Note that, as this step is conditioned on the regime

switching parameters, the state space model is linear with Gaussian disturbances V_t. Given the switching parameters and the knowledge of the Markov states, the time-varying matrices μ_t and R_t can be calculated at each point in time. The Kalman filter is initialised at B_0 and S_0 and the recursions are given by the following equations for $t = 1, 2..T$

$$\begin{aligned} B_{t|t-1} &= B_{t-1|t-1} \\ P_{t|t-1} &= P_{t-1|t-1} \\ \eta_{t|t-1} &= Y_t - X_t B_{t|t-1} - \mu_t \\ f_{t|t-1} &= X_t P_{t|t-1} X_t' + R_t \\ K_t &= P_{t|t-1} X_t' f_{t|t-1}^{-1} \\ B_{t|t} &= B_{t|t-1} + K_t \eta_{t|t-1} \\ P_{t|t} &= P_{t|t-1} - K_t x_t P_{t|t-1} \end{aligned}$$

The final iteration of the filter delivers $B_{T|T}$ and $P_{T|T}$. Alternatively, one can carry out a GLS transformation directly when calculating the posterior mean and variance. In this case:

$$B_{T|T} = P_{T|T} \left(vec \left(\sum_{t=1}^{T} \left(x_t \left(y_t - \mu_t \right)' R_t^{-1} \right) \right) + S_0^{-1} B_0' \right)$$

$$P_{T|T} = \left(\sum_{t=1}^{T} \left(R_t^{-1} \otimes x_t x_t' \right) + S_0^{-1} \right)^{-1}$$

We find that the computation time of this approach is similar to that of the Kalman filter in our application. The VAR coefficients can then be drawn from the multivariate Normal distribution.

2. $G\left(A|B, S_{it}, \alpha_{i,S_{it}}, \sigma^2_{i,S_{it}}, \tilde{y}_T\right)$: Conditional on the VAR coefficients B, the model can be written as $e_t = Au_t$. For a three-variable VAR (that we consider in the empirical section), this system is given as

$$\begin{pmatrix} \alpha_{1,S_{1t}} + \sigma_{1,S_{1t}} \varepsilon_{1t} \\ \alpha_{2,S_{2t}} + \sigma_{2,S_{2t}} \varepsilon_{2t} \\ \alpha_{3,S_{3t}} + \sigma_{3,S_{3t}} \varepsilon_{3t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a_1 & 1 & 0 \\ a_2 & a_3 & 1 \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{pmatrix} \tag{11}$$

where $[a_1, a_2, a_3]$ represent the elements of A.The second equation in this system is thus:

$$u_{2t} - \alpha_{2,S_{2t}} = -a_1 u_{1t} + \sigma_{2,S_{2t}} \varepsilon_{2t}$$

This a linear regression with a known variance. Given the knowledge of $\sigma_{2,S_{2t}}$, a GLS transformation can be applied to the regression and the conditional posterior for a_1 is given by the standard formula for linear regression models. Letting $y_t^* = \frac{u_{2t} - \sum_{j=1}^{M} \alpha_{2,S_{2t}} \times D_{t,j}}{\sum_{j=1}^{M} \sigma_{2,S_{2t}} \times D_{t,j}}$ and $x_t^* = \frac{-u_{1t}}{\sum_{j=1}^{M} \sigma_{2,S_{2t}} \times D_{t,j}}$ where $D_{t,j}$ is a matrix where the jth column denotes a dummy variable that equals 1 at time t when regime

j is active, the conditional posterior is $N(M^*, V^*)$:

$$\begin{aligned} M^* &= V^* \left(\Sigma_{0,k}^{-1} A_{0,k} + x_t^{*\prime} y_t^* \right) \\ V^* &= \left(\Sigma_0^{-1} + x_t^{*\prime} x_t^* \right)^{-1} \end{aligned}$$

The same procedure can be applied to the remaining equations of the system.

3. $G\left(\alpha_{i,S_{it}} | B, S_{it}, \sigma^2_{i,S_{it}}, A, \tilde{y}_T\right)$: As in step 2 above, the model can be written in terms of the orthogonalised residuals given $B, A : e_t = Au_t$. The ith equation of this system is

$$e_{it} = \alpha_{i,S_{it}} + \sigma_{i,S_{it}} \varepsilon_{it} \tag{12}$$

Conditional on knowing the Markov state for this equation S_{it} and the error variance $\sigma_{i,S_{it}}$, the procedure for a linear regression again applies. Following Koop (2003), we impose a labelling restriction on $\alpha_{i,S_{it}}$ in order to deal with the label switching problem inherent in Markov Switching models. In particular we impose the condition that $\alpha_{i,S_{it}=1} < \alpha_{i,S_{it}=2} < ... < \alpha_{i,S_{it}=M}$. As shown in Koop (2003), the conditional posterior is then a truncated normal $N(m, v) I\left(\alpha_{i,S_{it}=1} < \alpha_{i,S_{it}=2} < ... < \alpha_{i,S_{it}=M}\right)$ where:

$$\begin{aligned} m &= v \left[v_0^{-1} \alpha_0 + \sum_{t=1}^{T} \left\{ \sum_{j=1}^{M} D_{t,j} \times \frac{1}{\sigma_{i,S_{it}}} \right\} D_t e_{it} \right] \\ v &= \left(v_0^{-1} + \sum_{t=1}^{T} \left\{ \sum_{j=1}^{M} D_{t,j} \times \frac{1}{\sigma_{i,S_{it}}} \right\} D_t D_t' \right)^{-1} \end{aligned}$$

The same procedure is applied for each equation i.

4. $G\left(\sigma^2_{i,S_{it}} | B, S_{it}, \alpha_{i,S_{it}}, A, \tilde{y}_T\right)$: As shown above, conditional on a draw for $B, A, \alpha_{i,S_{it}}, S_{it}$ the model reduces to the set of regressions given in equation 12. As we condition on the latent states, the conditional posterior for $\sigma^2_{i,S_{it}}$ is standard and given by inverse Gamma $IG\left(\bar{\sigma}_{i,S_{it}}, \bar{T}_{i,S_{it}}\right)$. The scale parameter $\bar{\sigma}_{i,S_{it}}$ and degrees of freedom $\bar{T}_{i,S_{it}}$ are defined as:

$$\begin{aligned} \bar{\sigma}_{i,S_{it}} &= \bar{e}_{it}' \bar{e}_{it} + \sigma_0 \\ \bar{T}_{i,S_{it}} &= \dim(\bar{e}_{it}) + v_0 \end{aligned}$$

where $\bar{e}_{it}$ are the residuals from equation 12 selected for time periods when $S_{it} = j$. The same procedure is repeated for regime $j = 1...M$ and for each equation i.

5. $G(p_{i,IJ} | S_{it})$: Conditional on a draw for S_{it}, the elements of the transition probability matrix do not depend on the data. As shown in Chib (1996), the conditional posterior distribution for the elements of the transition probability matrix is Dirichlet:

$$p_{i,IJ} = D\left(u_{IJ} + \eta_{i,IJ}\right)$$

where $\eta_{i,IJ}$ denotes the number of times regime I is followed by regime J when considering the latent state for the ith orthogonal shock.

6. $G\left(\tilde{S}_{iT} | \sigma^2_{i,S_{it}}, \alpha_{i,S_{it}}, p_{i,IJ}, \tilde{e}_{iT}\right)$: Here $\tilde{S}_{iT}$ denotes the vector $[S_{i1}, S_{i2}, .., S_{iT}]$ and $\tilde{e}_{iT}$ is the

vector $[e_{i1}, e_{i2}, .., e_{iT}]$. Following Kim and Nelson (1999) we use a multi-move Gibbs step to sample from the conditional posterior of S_{it}. Kim and Nelson (1999) show that the markov property of S_{it} implies that

$$G\left(\tilde{S}_{iT}|\tilde{e}_{iT}\right) = G\left(S_{iT}|\tilde{e}_{iT}\right) \prod_{t=1}^{T-1} G\left(S_{it}|S_{it+1}, \tilde{e}_{iT}\right) \tag{13}$$

where we supress dependence on the parameters $\alpha_{i,S_{it}}, \sigma^2_{i,S_{it}}$ for notational simplicity. This density can be simulated in two steps:

(a) Calculating $G\left(S_{iT}|\tilde{e}_{iT}\right)$: The Hamilton (1989) filter provides $G\left(S_{iT}|\tilde{e}_{iT}\right)$ from which S_{iT} can be simulated. Denoting $\hat{\xi}_{i,t}$ as a vector where the jth element equals $\Pr\left(S_{it} = j\right)$, the filter iterates on the following two equations for $t = 1, ...T$:

$$\hat{\xi}_{i,t+1|t} = P_i.\hat{\xi}_{i,t|t} \tag{14}$$

$$\hat{\xi}_{i,t|t} = \frac{F\left(e_{it}|S_{it} = j\right) \odot \hat{\xi}_{i,t|t-1}}{\sum_{j=1}^{J} F\left(e_{it}|S_{it} = j\right) \odot \hat{\xi}_{i,t|t-1}} \tag{15}$$

where P_i denotes the transition probability matrix and

$$F\left(e_{it}|S_{it} = j\right) = \left(2\pi\sigma^2_{i,S_{it}}\right)^{-T/2} \exp\left(-\frac{\left(e_{it} - \alpha_{i,S_{it}}\right)'\left(e_{it} - \alpha_{i,S_{it}}\right)}{2\sigma^2_{i,S_{it}}}\right)$$

The last iteration of the filter provides the probabilities $\hat{\xi}_{i,T|T}$ which can be used to draw S_{iT}.

(b) Kim and Nelson (1999) show that:

$$G\left(S_{it}|S_{it+1}, \tilde{e}_{iT}\right) \propto G\left(S_{it+1}|S_{it}\right) G\left(S_{it}|\tilde{e}_{iT}\right) \tag{16}$$

The first term on the right hand side of this expression $G\left(S_{it+1}|S_{it}\right)$ denotes the transition probability. The second term $G\left(S_{it}|\tilde{e}_{iT}\right)$ represent the filter probabilties $\Pr\left(S_{it} = j\right)$ obtained by running the Hamilton (1989) filter in step a. Expression 16 can be used to draw S_{it}. For example, in a two regime model we proceed by calculating $\Pr\left(S_{it} = 1|S_{it+1}, \tilde{e}_{iT}\right) = \frac{G(S_{it+1}|S_{it}=1)G(S_{it}=1|\tilde{e}_{iT})}{\sum_{j=1}^{M} G(S_{it+1}|S_{it}=j)G(S_{it}=j|\tilde{e}_{iT})}$. If this probability is larger than a draw from a standard uniform, then $S_{it} = 1$, else $S_{it} = 2$.

3.2.1 Time-varying transition probabilities

The model can be easily extended to allow the transition probabilities to depend on $k \times 1$ vector of relevant covariates z_{it}. For simplicity consider a model with two components for the orthogonal shocks. The transition probabilties are then defined as:

$$P_i\left(S_{it} = J|S_{it-1} = I\right) = \begin{pmatrix} p_{i,11}\left(z_{it}\right) & 1 - p_{i,22}\left(z_{it}\right) \\ 1 - p_{i,11}\left(z_{it}\right) & p_{i,22}\left(z_{it}\right) \end{pmatrix}$$

As discussed in Filardo and Gordon (1998), the process for the transition probabilties can be represented as a Probit model:

$$\begin{aligned} S_{it} = 2 &\iff s^*_{it} \geq 0 \\ s^*_{it} &= \lambda_{i,0} + \gamma_{i,1} z_{it} + \lambda_{i,1} S_{it-1} + v_{it}, v_{it} \sim N(0,1) \end{aligned} \tag{17}$$

where s^*_{it} is a latent variable. The parameters $\lambda_{i,0}$ and $\lambda_{i,1}$ are regime-specific intercepts while the slope coefficients $\gamma_{i,1}$ drives the time-variation in the transition probabilties. The transition probabilties are then given by:

$$\begin{aligned} p_{i,11}(z_t) &= \Pr\left(v_{it} < -\left(\lambda_{i,0} + \gamma_{i,1} z_{it} + \lambda_{i,1} S_{it-1}\right)\right) \\ p_{i,22}(z_t) &= \Pr\left(v_{it} \geq -\left(\lambda_{i,0} + \gamma_{i,1} z_{it} + \lambda_{i,1} S_{it-1}\right)\right) \end{aligned}$$

which can be easily calculated using the normal CDF.

We assume a normal prior for the coefficients of equation 17: $\mathrm{P}(\gamma_i) \sim N(\gamma_0, V_\gamma)$ where $\gamma = [\lambda_{i,0}, \gamma_{i,1}, \lambda_{i,1}]$. The Gibbs algorithm presented above requires minor modifications. First, step 5 now involves a draw of the latent variable s^*_{it}. Conditional on γ_i, the latent variable can be drawn easily from a truncated normal distribution. That is

$$\begin{aligned} s^*_{it} &\sim N_{I>0}(\mu_i, \tau) \text{ if } S_t = 1 \\ s^*_{it} &\sim N_{I<0}(\mu_i, \tau) \text{ if } S_t = 2 \end{aligned} \tag{18}$$

where $I < 0$ denotes truncation below zero and $I > 0$ denotes truncation above zero. Here $\mu = \lambda_{i,0} + \gamma_{i,1} z_{it} + \lambda_{i,1} S_{it-1}$ while $\tau = 1$ for identification. Given a draw for s^*_{it}, equation 17 is simply a regression with a unit variance. The conditional posterior for $\lambda_{i,0}$ is thus normal: $N(M, V)$

$$\begin{aligned} V &= \left(V_\gamma^{-1} + \bar{z}'_{it} \bar{z}_{it}\right) \\ M &= V\left(V_\gamma^{-1} \gamma_0 + \bar{z}'_{it} s^*_{it}\right) \end{aligned}$$

where $\bar{z}_{it} = [1, z_{it}, , S_{it-1}]$ denotes the matrix of regressors. Finally note that step 6 above needs a minor modification to account for the fact that a different transition probability matrix applies at each point in time.

3.3 Selection of the number of components

Choosing the number of components or regimes in the proposed BVAR model is a crucial specification choice. We carry out model selection by comparing the marginal likelihood across models with a different number of components. The marginal likelihood is defined as:

$$f(\tilde{y}) = \int f(\tilde{y}|\Xi)\, p(\Xi)\, d\Xi \tag{19}$$

where $\tilde{y} = [y_1, y_2, .., y_T]$, Ξ denotes the unknown parameters of the model, $f(\tilde{y}|\Xi)$ is the likelihood and $p(\Xi)$ is the proper prior distribution. As is well known, the integration problem in equation 19 is non-trivial and several numerical methods have been proposed for this calculation.

In our study we consider two approaches to estimating $f(\tilde{y})$. First we use the reciprocal importance sampling estimator proposed in Gelfand and Dey (1994). These authors show that the

reciprocal of the marginal likelihood can be defined as:

$$E\left[\frac{q\left(\Xi\right)}{f\left(\tilde{y}|\Xi\right)p\left(\Xi\right)}\right]=\frac{1}{f\left(\tilde{y}\right)} \tag{20}$$

where $q\left(\Xi\right)$ is an importance density. Given M draws of Ξ from the Gibbs algorithm described above, the expectation in equation 20 can be approximated as $\frac{1}{M}\sum_{j=1}^{M}\frac{q(\Xi_j)}{f(\tilde{y}|\Xi_j)p(\Xi_j)}$. Following Geweke (1999) we use a truncated normal distribution as the importance density. As discussed in Fruhwirth-Schnatter (2004), the performance of this estimator can be adversely affected by the tail behaviour of the importance density. Instead, Fruhwirth-Schnatter (2004) suggests using the bridge sampling estimator proposed in Meng and Wong (1996). The bridge sampling estimator is based on the following identity

$$f\left(\tilde{y}\right)=\frac{E_q\left(\alpha\left(\Xi\right)f\left(\Xi|\tilde{y}\right)\right)}{E_f\left(\alpha\left(\Xi\right)q\left(\Xi\right)\right)} \tag{21}$$

where $f\left(\Xi|\tilde{y}\right)=f\left(\tilde{y}|\Xi\right)p\left(\Xi\right)$ and E_x denotes the expectation with respect to x. Given a choice for the function $\alpha\left(\Xi\right)$, this can be approximated as $\frac{L^{-1}\sum_{l=1}^{L}\alpha\left(\Xi^l\right)f\left(\Xi^l\backslash\tilde{y}\right)}{M^{-1}\sum_{m=1}^{M}\alpha(\Xi^m)q(\Xi^m)}$ where Ξ^l denotes draws from the importance density $q\left(\Xi\right)$ while Ξ^m represent the draws from the Gibbs sampler. Meng and Wong (1996) show that an optimal choice for $\alpha\left(\Xi\right)$ is $\frac{1}{Lq(\Xi)+M\frac{f(\Xi|\tilde{y})}{f(\tilde{y})}}$. As this expression involves $f\left(\tilde{y}\right)$, Meng and Wong (1996) propose an iterative approach. The algorithm starts with an initial guess for $f\left(\tilde{y}\right)$ (set to the the reciprocal importance sampling estimator in our applications) and iterates on the following recursion until convergence:

$$\left[f\left(\tilde{y}\right)\right]^{new}=\left[f\left(\tilde{y}\right)\right]^{old}\frac{L^{-1}\sum_{l=1}^{L}\frac{f\left(\Xi^l|\tilde{y}\right)}{Lq\left(\Xi^l\right)+M\frac{f\left(\Xi^l|\tilde{y}\right)}{\left[f(\tilde{y})\right]^{old}}}}{M^{-1}\sum_{m=1}^{M}\frac{q(\Xi^m)}{Lq(\Xi^m)+M\frac{f(\Xi^m|\tilde{y})}{\left[f(\tilde{y})\right]^{old}}}} \tag{22}$$

We use a mixture distribution as our choice for the importance density. The exact components and weights are provided in Appendix A.

A simpler alternative to marginal likelihood calculation is based on model selection criteria. Koop (2003) (chapter 10) demonstrates that the Akaike (AIC) and Schwarz (SIC) information criteria perform well in selecting the number of components in a single equation mixture model. These are defined as:

$$\begin{aligned} AIC &= \ln f\left(\tilde{y}|\Xi\right)-2P \\ SIC &= \ln f\left(\tilde{y}|\Xi\right)-P\ln(T) \end{aligned}$$

where P denotes the number of parameters and following Koop (2003), the likelihood is evaluated at the posterior mean. The recent empirical literature has made heavy use of the deviance information criterion (DIC) introduced in Spiegelhalter *et al.* (2002). The DIC is defined as:

$$DIC=\bar{D}+p_D. \tag{23}$$

The first term is $\bar{D}=E\left(-2\ln f\left(\tilde{y}|\Xi^m\right)\right)\approx\frac{1}{M}\sum_m\left(-2\ln f\left(\tilde{y}|\Xi^m\right)\right)$ where $f\left(\tilde{y}|\Xi^m\right)$ is the likelihood evaluated at the draws from the Gibbs sampler. This term measures goodness of fit. The second term p_D is defined as a measure of the number of effective parameters in the model (or model

complexity). This is defined as $p_D = \bar{D} - D\left(\bar{\Xi}\right) = E\left(-2\ln f\left(\tilde{y}|\Xi^m\right)\right) - \left(-2\ln f\left(\tilde{y}|E(\Xi^m)\right)\right)$ and can be approximated as $p_D = \frac{1}{M}\sum_m\left(-2\ln f\left(\tilde{y}|\Xi^m\right)\right) - \left(-2\ln L\left(\frac{1}{M}\sum_m \Xi^m\right)\right)$ We consider the performance of these criteria along with the marginal likelihood in the Monte-Carlo experiment presented below.

Note that the calculation of marginal likelihoods and information criteria requires an estimate of the likelihood of the BVAR model. This can be calculated via the Hamilton (1989) filter. We first re-write the model as a Markov switching VAR (as in equation 8). We then define a new composite state variable $\bar{S}_t$ that accounts for the independent state variable in each equation. In the simple example considered in equation 8, $\bar{S}_t$ takes four values:

$$\begin{aligned}\bar{S}_t &= 1 \text{ if } S_{1t}=1 \text{ and } S_{2t}=1 \\ \bar{S}_t &= 2 \text{ if } S_{1t}=1 \text{ and } S_{2t}=2 \\ \bar{S}_t &= 3 \text{ if } S_{1t}=2 \text{ and } S_{2t}=1 \\ \bar{S}_t &= 4 \text{ if } S_{1t}=2 \text{ and } S_{2t}=2\end{aligned} \tag{24}$$

The transition probabilties associated with $\bar{S}_t$ are then given by $\bar{P} = P_1 \otimes P_2$. With the model written in this form, the recursions of the Hamilton (1989) filter in equations 14 and 15 provide the likelihood for each observation t : $lik_t = \sum_{j=1}^{J} F\left(e_{it}|S_{it}=j\right) \odot \hat{\xi}_{i,t\backslash t-1}$. The log likelihood for the model can be obtained as $\ln f\left(\tilde{y}|\Xi\right) = \sum_{t=1}^{T} \ln lik_t$.

4 Estimation using artificial data

In this section we present a simple Monte-Carlo experiment. The aim is to evaluate the performance of the Gibbs sampling algorithm and to assess the methods for model selection considered above.

We generate data from the following bi-variate VAR model:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{21} \\ B_{12} & B_{22} \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{25}$$

where

$$\begin{pmatrix} B_{11} & B_{21} \\ B_{12} & B_{22} \end{pmatrix} = \begin{pmatrix} 0.9 & -0.1 \\ 0.1 & 0.8 \end{pmatrix} \tag{26}$$

and

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ A_{21} & 1 \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \tag{27}$$

with $A_{21} = -0.5$. The error terms are defined as:

$$\begin{aligned} e_{1t} &= \alpha_{1,S_{1t}} + \sigma_{1,S_{1t}}\varepsilon_{1t} \\ e_{2t} &= \alpha_{2,S_{2t}} + \sigma_{2,S_{2t}}\varepsilon_{2t} \end{aligned} \tag{28}$$

with $\varepsilon_{1t}, \varepsilon_{2t} \sim N(0,1)$. The orthogonal residuals are assumed to be characterised by two compo-

nents where

$$\begin{aligned} \alpha_{i,S_{it}=1} &= -0.5, \alpha_{i,S_{it}=2} = 0.5 \\ \sigma^2_{i,S_{it}=1} &= 0.1, \sigma^2_{i,S_{it}=2} = 0.3 \end{aligned} \tag{29}$$

with transition probabilities $p_{i,11} = 0.95, p_{i,22} = 0.95$. We generate 500 samples of length $T + 100$ discarding the first 100 observations and carrying out estimation on the remaining T. We consider two sample sizes $T = 400$ and $T = 150$. For each iteration of the Monte-Carlo, we run the Gibbs sampler for 5000 iterations using the final 1000 draws for inference. In addition to estimating the model with two components we also estimate a three component model and a fixed coefficient BVAR and calculate the marginal likelihood, the AIC, SIC and the DIC for model comparison.

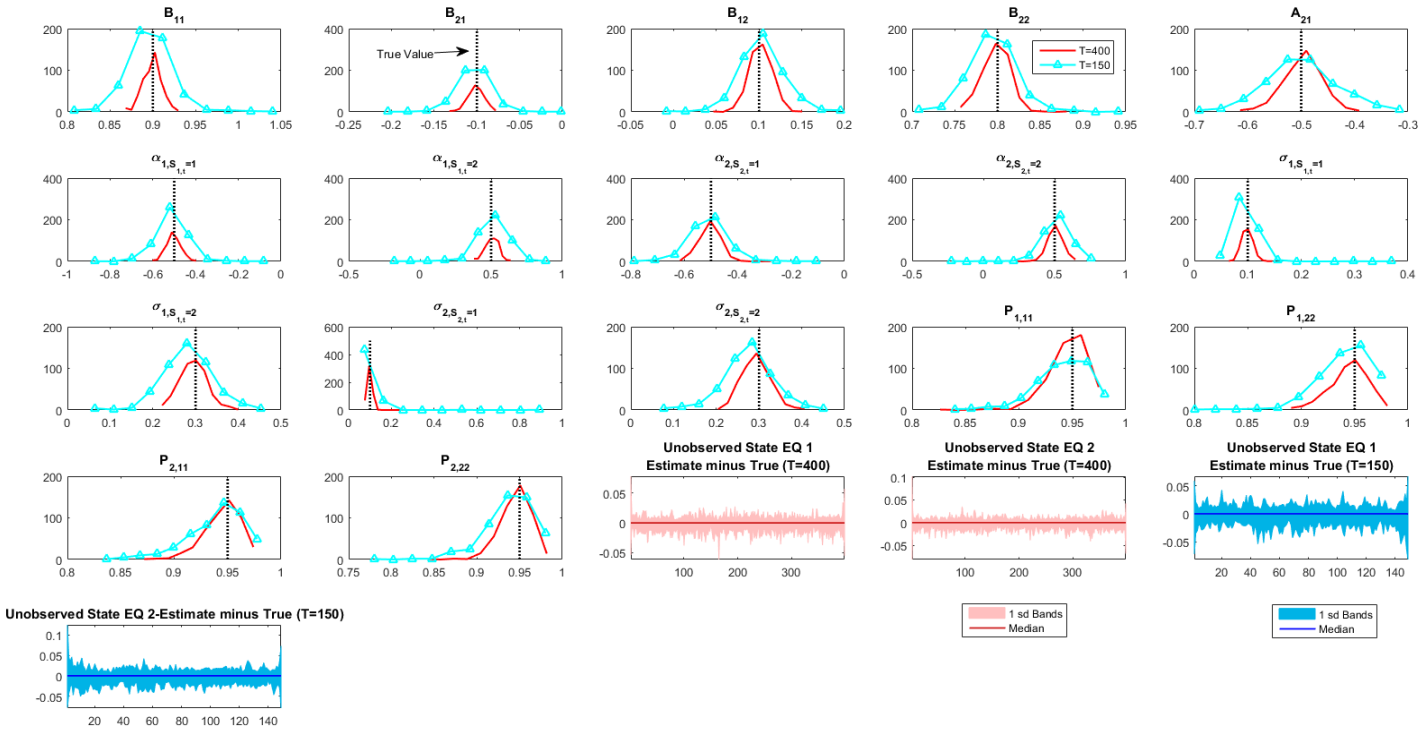

Figure 1: Results from the Monte-Carlo Experiment

Figure 1 plots the histogram of the posterior mean of the parameters across the 500 Monte-Carlo iterations. The red lines display the results when the sample size equals 400 while the blue lines display the results for the smaller sample with 150 observations. The vertical black dotted lines represent the true values of the parameters. The final four sub-plots present the distribution of the difference between the true and estimated value of S_{it}.

When the sample size equals 400, the estimated distribution of the VAR coefficients B_{ij} and A_{21} is tightly centered around the true values. Similarly, the regime switching parameters and transition probabilties are estimated with a reasonable precision. Finally, note that the bias in S_{it} is close to zero across the replications. With the smaller sample size, there is an expected increase in the variance of the estimated distribution. While the mean estimate is close to the true value for the time-invariant parameters and $\alpha_{i,S_{it}}$, there appears to be a slight downward bias in the estimates of $\sigma^2_{i,S_{it}}$ and the transition probabilties. The results therefore suggest that given a reasonable sample size, the MCMC algorithm delivers a good performance.[3]

	Probability of selecting the model	
	T=400	T=150
Marginal Likelihood (Reciprocal Importance Sampling)		
3 Components	0.002	0
2 Components (True Model)	0.998	0.854
Linear BVAR	0	0.146
Marginal Likelihood (Bridge Sampling)		
3 Components	0.002	0
2 Components (True Model)	0.998	0.862
Linear BVAR	0	0.138
AIC		
3 Components	0.002	0.006
2 Components (True Model)	0.998	0.992
Linear BVAR	0	0.002
SIC		
3 Components	0.002	0
2 Components (True Model)	0.998	0.954
Linear BVAR	0	0.046
DIC		
3 Components	0.974	0.964
2 Components (True Model)	0.026	0.036
Linear BVAR	0	0

Table 1: Performance of model selection procedures.

Table 1 considers the performance of the model selection procedures discussed in section 3.3. It reports the probability of selecting the linear BVAR or the model with two or three components using either the marginal likelihood or model selection criteria. With a sample size of $T = 400$, the two estimators of the marginal likelihood perform equally well, selecting the true model almost 100% of the time. It is interesting to note that the AIC and the SIC deliver an equally impressive performance. In contrast, the DIC selects the 3 component model 97% of the time. When $T =$

[3] In a previous version of this paper we show that similar conclusions are reached if data is generated from a three component model. See Chiu *et al.* (2016).

150, the performance of the marginal likelihood based method deteriorates slightly, with a larger probability attached to the selection of the BVAR model. In contrast, the AIC still delivers an excellent performance and selects the correct model in almost all replications. The performance of the SIC in the smaller sample is also impressive. These results suggest tentatively that both marginal likelihood comparisons and the Akaike and Schwarz criteria perform well in moderately large samples. The latter criteria appear to be particularly useful when the sample is small.

5 Empirical application: Modelling and forecasting the yield curve

Several recent studies have highlighted the fact that the yield curve and the macroeconomy are closely related (see for e.g. Diebold *et al.* (2006)). This is one of the motivations behind the large literature that focuses on forecasting the yield curve. Note, however, that some recent papers have pointed out that the dynamics of the yield curve are subject to structural shifts. For example, Mumtaz and Surico (2009) and Bianchi *et al.* (2009) show that innovations to the level and the slope of the yield curve feature heteroscedasticity. In addition, it is well known that yields at longer maturities have been trending downwards in recent years (i.e. 'Greenspan's conundrum'). This provides prima facie evidence that the shocks to the yield curve may be characterised by non-normality. In this section we investigate if allowing for non-Gaussian shocks in a simple model of the yield curve can improve model fit and out-of-sample forecasting performance.

We follow Diebold and Li (2006) and model the yield curve using the Nelson and Siegel (1987) specification. Letting $y_t(\tau)$ denote zero coupon government bond yields at maturity τ, the Nelson and Siegel (1987) model is defined as:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-e^{-\lambda_t\tau}}{\lambda_t\tau} - e^{-\lambda_t\tau}\right) \tag{30}$$

where λ_t controls the exponential decay rate. The level, slope and the curvature of the yield curve are captured by β_{1t}, β_{2t} and β_{3t}, respectively. These three factors can be easily estimated by fixing λ_t and estimating the factors via OLS for each t. Diebold and Li (2006) use $\lambda_t = 0.0609$, the value that maximises the loading on the curvature factor. The dynamics of the factors can be modelled as a VAR process which is used in Diebold and Li (2006) to produce out of sample forecasts. In this section, we compare this VAR specification with the following extended model:

$$Z_t = c + \sum_{l=1}^{L} B_l Z_{t-l} + u_t$$

where $Z_t = \{\beta_{1t}, \beta_{2t}, \beta_{3t}\}$ and $e_t = Au_t$. The orthogonal shock to the ith equation of the VAR is a Markov mixture of normals:

$$e_{it} = \alpha_{i,S_{it}} + \sigma_{i,S_{it}}\varepsilon_{it}, \varepsilon_{it} \sim N(0,1) \tag{31}$$

where we allow for the possibility of multiple regimes.

5.1 Data

We obtain monthly data on zero coupon yields from Gürkaynak *et al.* (2007) which runs from December 1971 to January 2016. The estimation of the yield curve factors follows Diebold and Li

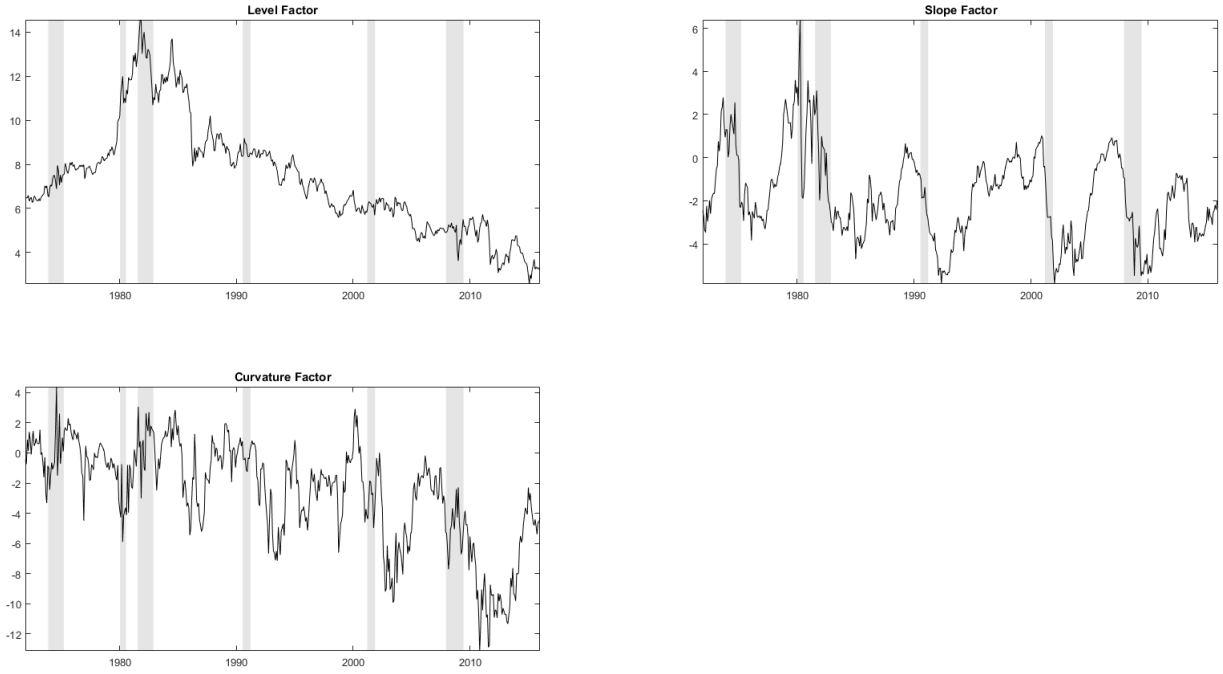

Figure 2: Estimated yield curve factors. Shaded areas represent NBER recession dates.

(2006) with $\lambda_t = 0.0609$. The cross-section regression at each month is based on the beginning-of-month quotes for all available yields of different years of maturity. This means that the yield curve factors are estimated using Treasuries of 1-15, 1-20 and 1-30 years of maturity for the periods December 1971 to June 1981, July 1981 to November 1985 and December 1985 to January 2016, respectively. The estimated factors are shown in figure 2 and have a high correlation with those estimated by Diebold and Li (2006).

5.2 Full sample estimation

We consider models with up to four regimes. The models are estimated using 105000 Gibbs replications with last 5000 iterations used for inference. The appendix presents recursive mean plots of key parameters that provide evidence for the convergence of the algorithm. The bridge sampling estimate of the log marginal likelihood suggests that the two component model fits the data best – the marginal likelihood is estimated to be 134.77 for the two component model which is larger than the estimates for the three component model (126.62), the four component model (111.34) and the Bayesian VAR (-36.66).[4]

Table 2 and figure 3 presents the posterior estimates of the regime switching parameters and transition probabilties. It is clear from 2 that regime 1 is associated with a negative mean and high variance for the shocks to the level and curvature equations. For the slope equation, regime 2 is characterised by high variance.[5] The high variance regime was active over the late 1970s, the early

[4] The reciprocal importance sampling estimator provides the same result.

[5] In Appendix C we present results from a version of the model where the order of the factors is assumed to be: (1) curvature, (2) slope and (3) level. The results shows that the interpretation of the regimes remains as in the

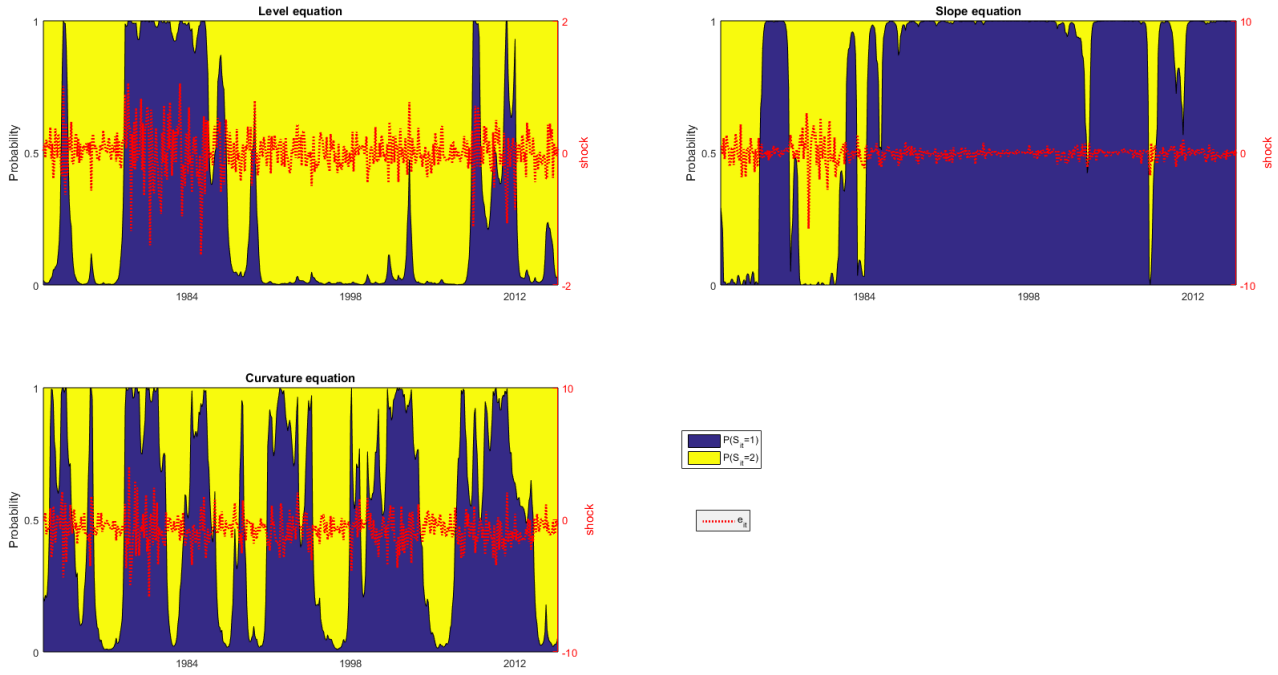

Figure 3: Regime probabilities and orthogonalised shocks. The stacked area chart plots $\Pr(S_{it} = 1)$ (blue area) and $\Pr(S_{it} = 2)$ (yellow area). The red dotted lines show e_{it} estimated using the posterior mean of the model parameters.

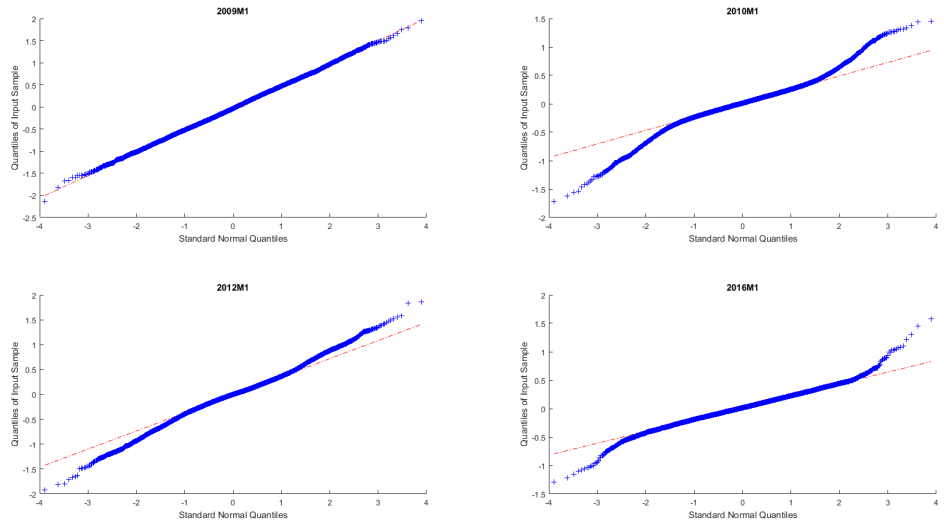

Figure 4: Quantile-quantile plots for the distribution of the orthogonalised shock for equation of the level factor.

Equation	$\alpha_{i,S_{it}=1}$	$\alpha_{i,S_{it}=2}$	$\sigma^2_{i,S_{it}=1}$	$\sigma^2_{i,S_{it}=2}$	P_i
Level	-0.024	0.017	0.241	0.041	$\begin{pmatrix} 0.946 & 0.022 \\ 0.054 & 0.978 \end{pmatrix}$
	$[-0.055, -0.006]$	$[-0.008, 0.045]$	$[0.210, 0.283]$	$[0.037, 0.046]$	
Slope	-0.029	0.064	0.105	1.340	$\begin{pmatrix} 0.983 & 0.058 \\ 0.017 & 0.942 \end{pmatrix}$
	$[-0.071, -0.007]$	$[-0.006, 0.171]$	$[0.093, 0.117]$	$[1.148, 1.573]$	
Curvature	-0.935	-0.545	1.978	0.368	$\begin{pmatrix} 0.934 & 0.069 \\ 0.066 & 0.931 \end{pmatrix}$
	$[-1.235, -0.628]$	$[-0.759, -0.326]$	$[1.733, 2.340]$	$[0.306, 0.442]$	

Table 2: Estimates of regime dependent parameters. Median and 68 percent highest posterior density interval

1980s and during the recent financial crisis for the first two equations. The shock to the curvature equation is subject to regular regime shifts. It is interesting to note that for each equation, there are periods of time when the probability of both states is non-negligible and the implied mixture distribution of the shocks may depart from normality. As an example, we draw $10,000$ random numbers from the implied distribution of the shock to the level equation in January 2009, 2010, 2012 and 2016.[6] Figure 4 plots the quantiles of the standard normal distribution against the quantiles of these random numbers. These plots strongly suggest that the distribution was characterised by non-normality in post-2010 period.[7]

5.3 Forecasting performance

We now consider how the BVAR with non-Gaussian shocks performs in terms of out-of-sample point and density forecasting. We use models with two and three regimes in the forecast experiment. The model comparison for the full sample suggests that the fit deteriorates substantially if more regimes are considered and therefore we use relatively parsimonious models. The competing models are assumed to be (1) a linear BVAR (2) a BVAR where the orthogonal shocks are assumed to be drawn from a T-distribution and (3) a BVAR where the variance of the orthogonal residuals follows a stochastic volatility (SVOL) process. The linear BVAR offers a simple benchmark that incorporates the assumptions of normality. The remaining two models feature high and low frequency fluctuations in the volatility of the errors, respectively. This can induce fat tails in the error distribution but leaves the density symmetric.

The BVAR with T-distributed shocks is defined as:

$$Z_t = c + \sum_{l=1}^{L} B_l Z_{t-l} + A e_t \tag{32}$$

where A is lower triangular the ith residual e_{it} is assumed to be distributed as $T\left(0, \sigma_i^2, df_i\right)$ where σ_i^2 is the variance while df_i denotes the degrees of freedom of the T-density. Chiu *et al.* (2014) present details on the MCMC algorithm used to estimate this model. The BVAR with stochastic volatility has the same form as equation 32 above but the orthogonal residuals are now defined as $e_{it} \sim N(0, \sigma_{it}^2)$ where $\ln \sigma_{it}^2 = \ln \sigma_{it-1}^2 + q^{1/2}\eta_{it}$. Estimation of this model is standard and also discussed in Chiu *et al.* (2014). Note that we employ the prior described in section 3.1 for

benchmark specification suggesting that ordering is not important for this application.

[6] This is defined as $\Pr(S_{1t} = 1) \times R_1 + \Pr(S_{1t} = 2) \times R_2$ where $R_j \tilde{} N(\alpha_{1,S_{1t}=j}, \sigma^2_{1,S_{1t}=j})$.

[7] We also estimated a version of the model where the transition probabilties are assumed to depend on lagged output growth and inflation. The results from this model are similar to benchmark case and available on request.

the coefficients of all the VARs used in the forecasting experiment.

We conduct a pseudo real-time forecasting exercise. In particular, we estimate the forecasting models using data from December 1971 to December 1979. Then each model is estimated recursively adding one month of data at a time until January 2015. At each recursion, we produce a 12 month density forecast for the three factors.[8] The k step ahead forecast density from the proposed model is defined as:

$$P\left(\hat{Z}_{t+k}|Z_t\right) = \sum_{m=1}^{M} P\left(\hat{Z}_{t+k}|S_{it+k}=o, Z_t, \Xi\right) P_i\left(S_{it+k}=o|S_{it}=m\right) P\left(S_{it}=m|Z_t,\Xi\right) \tag{33}$$

where Ξ represents the model parameters. The first term in equation 33 denotes the Gaussian forecast density conditioned on the parameters in regime o. The second term is denotes the transition probability and the final term is the filter probability obtained from the Hamilton (1989) filter. The point forecast is obtained as the mean of the forecast density.

We assess the point forecasts using root mean squared errors (RMSE) and the density forecasts using the continuous rank probability score (CRPS). Our preference to use CRPS instead of using log scores is related to the relative advantages of CRPS: it is better at rewarding values from the predictive density that are close to but not equal to the outcome, and it is less sensitive to outlier outcomes (see Gneiting and Raftery (2007)).

	RMSE				CRPS			
	1M	3M	6M	12M	1M	3M	6M	12M
Level Factor								
VAR-T	0.990	0.984	0.988	0.981	0.986	0.987	0.994	0.987
M2-VAR	0.981	0.977	0.975	0.947	0.964	0.953	0.953	0.902
M3-VAR	0.996	1.006	1.006	0.972	0.990	1.003	0.994	0.923
SVOL-VAR	0.979	0.977	0.979	0.958	0.960	0.959	0.966	0.924
Slope Factor								
VAR-T	0.965	0.943	0.948	0.977	0.938	0.932	0.957	1.017
M2-VAR	0.991	0.977	0.982	0.985	0.930	0.952	0.981	0.991
M3-VAR	0.982	0.971	0.970	0.988	0.924	0.947	0.991	1.050
SVOL-VAR	0.949	0.920	0.918	0.940	0.900	0.881	0.928	1.011
Curvature Factor								
VAR-T	0.972	0.981	0.993	1.009	0.973	0.982	1.000	1.017
M2-VAR	0.984	0.987	0.991	0.995	0.978	0.989	1.011	1.032
M3-VAR	0.986	0.991	0.996	1.002	0.986	0.999	1.030	1.046
SVOL-VAR	0.991	1.006	1.023	1.052	0.985	1.014	1.051	1.091

Table 3: RMSE and CRSP relative to a BVAR. Average over the period 1980M1-2015M1. VAR-T is the VAR with fat tailed shocks, SVOL-VAR is the VAR with stochastic volatility. M2-VAR is the proposed model with 2 components. M3-VAR is the proposed model with 3 components

Tables 3 and 4 present the results of the forecast evaluation. The first four columns of each table shows the RMSE of the forecasting model relative to the RMSE obtained from the BVAR. The last four columns display the relative CRPS. In 3 these relative RMSE's and CRPS's are averaged

[8] As noted by Diebold and Li (2006), forecasting the factors in this model is equivalent to forecasting the underlying yields.

	RMSE				CRPS			
	1M	3M	6M	12M	1M	3M	6M	12M
				Level Factor				
VAR-T	0.787	0.774	0.780	0.762	0.775	0.763	0.788	0.769
M2-VAR	0.779	0.772	0.773	0.730	0.765	0.748	0.759	0.682
M3-VAR	0.792	0.797	0.807	0.761	0.780	0.781	0.801	0.708
SVOL-VAR	0.786	0.776	0.783	0.758	0.767	0.751	0.774	0.737
				Slope Factor				
VAR-T	0.778	0.765	0.807	0.900	0.750	0.766	0.868	0.999
M2-VAR	0.777	0.761	0.790	0.858	0.729	0.730	0.816	0.918
M3-VAR	0.784	0.762	0.789	0.866	0.730	0.729	0.835	0.982
SVOL-VAR	0.760	0.735	0.765	0.856	0.717	0.705	0.805	0.967
				Curvature Factor				
VAR-T	0.934	0.968	0.996	1.036	0.928	0.973	1.016	1.075
M2-VAR	0.944	0.981	1.016	1.049	0.937	1.003	1.070	1.137
M3-VAR	0.943	0.978	1.016	1.053	0.941	1.006	1.089	1.153
SVOL-VAR	0.954	1.005	1.053	1.112	0.942	1.022	1.100	1.178

Table 4: RMSE and CRSP relative to a BVAR. Average over the period 1990M1-2015M1. VAR-T is the VAR with fat tailed shocks, SVOL-VAR is the VAR with stochastic volatility. M2-VAR is the proposed model with 2 components. M3-VAR is the proposed model with 3 components

over the period January 1980 to January 2015. Table 4 presents the average, post-1990 a period associated with a slow, steady decline in the level of yields (see figure 2). Note that a number less than 1 in the tables implies that the forecasting model improves upon the standard BVAR.

The relative RMSE's in table 3 show that all models offer a modest improvement in point forecasting over the BVAR. For the level factor, the model with two regimes (M2-VAR) has the smallest relative RMSE at the six and twelve month horizons, albeit with the VAR with stochastic volatility (SVOL-VAR) a very close second. The SVOL-VAR provides the best forecasting performance for the slope factor at the one year horizon, while all models are roughly comparable to the BVAR in forecasting the curvature factor twelve months ahead. The M2-VAR delivers the best density forecasts for the level factor at all horizons offering an improvement of about 10% over the BVAR at the one year horizon. However, as in the case of the point forecasts, the SVOL-VAR provides the best density forecasts for the slope factor at the three and six month horizons with the VAR with T-distributed shocks (VAR-T) a close competitor. Both the VAR-T and the M2-VAR provide modest gains in the density forecast of the curvature factor at short horizons.

Over the post-1990 forecast sample, the improvement in forecasting performance over the BVAR is substantially larger. For example, the M2-VAR and the SVOL-VAR deliver RMSEs in forecasting the level factor which are more than 20% lower than those obtained from the BVAR model. Similarly, the performance of the M2-VAR in forecasting the slope factor is very close to the SVOL-VAR which is again the best performing model. At short horizons, the VAR-T and the M2-VAR provide the lowest relative RMSEs for the curvature factor. At the one year horizon, the density forecast for the level factor from the M2-VAR is 32% more accurate than that obtained from the BVAR model. Note that this provides a substantial improvement over both the VAR-T and the SVOL-VAR. Similarly, the M2-VAR improves upon the SVOL-VAR in terms of the density forecast of the slope factor at the one year horizon. Note also that the accuracy of the M2-VAR density forecast for the slope factor is very close to that delivered by the SVOL-VAR at shorter horizons.

The forecast experiment suggests the following conclusions: (1) The proposed model with two components delivers point and density forecasts for the level of the yield curve that are more accurate than those obtained from the competing models. The M2-VAR's forecasting performance is highly competitive with the alternatives when considering the slope and the curvature factors. (2) The proposed models perform best over a period where the slope of the yield curve shows a slow and steady decline and displays little volatility. This suggests that these models may be especially useful over periods when the data is relatively stable. In summary, there is some evidence that the BVAR with non-Gaussian shocks can be useful for point and density forecasting.

6 Conclusions

This paper proposes a BVAR model where the orthogonal shocks have a non-normal distribution. The non-normality is introduced via a Markov mixture of normals. We provide a Gibbs sampling algorithm to approximate the posterior distribution of the parameters and discuss methods to select the number of components in the mixture. An application to a yield curve model suggests that the proposed model fits the data better than a BVAR with Gaussian shocks. Moreover, the BVAR with non-normal shocks provides point and density forecasts of the level and slope of the yield curve that are more accurate than those obtained from the standard BVAR. The relative improvement in forecasting performance is especially large over the post-1990 period when the level of the yield curve displayed little volatility. In future work it would be interesting to check if evidence for non-normality of shocks can be found in BVAR models that include asset prices such as exchange rates and oil prices. Given the recent political and economic uncertainty, it has become crucial for policy makers to obtain accurate forecasts for these variables. It would therefore be useful to investigate if the proposed model can be helpful for this task.

References

Banbura, Marta, Domenico Giannone and Lucrezia Reichlin, 2010, Large Bayesian vector auto regressions, *Journal of Applied Econometrics* **25**(1), 71–92.

Bec, Frédérique, Anders Rahbek and Neil Shephard, 2008, The ACR Model: A Multivariate Dynamic Mixture Autoregression, *Oxford Bulletin of Economics and Statistics* **70**(5), 583–618.

Bianchi, Francesco, Haroon Mumtaz and Paolo Surico, 2009, The great moderation of the term structure of UK interest rates, *Journal of Monetary Economics* **56**(6), 856 – 871.

Chan, Joshua C.C., 2015, Large Bayesian VARs: A flexible Kronecker error covariance structure, *CAMA Working Papers 2015-41*, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

Chib, Siddhartha, 1996, Calculating posterior distributions and modal estimates in Markov mixture models, *Journal of Econometrics* **75**(1), 79–97.

Chib, Siddhartha and Srikanth Ramamurthy, 2014, DSGE Models with Student-t Errors, *Econometric Reviews* **33**(1-4), 152–171.

Chiu, Ching-Wai (Jeremy), Haroon Mumtaz and Gabor Pinter, 2014, Fat-tails in VAR Models, *Working Papers 714*, Queen Mary University of London, School of Economics and Finance.

Chiu, Ching-Wai (Jeremy), Haroon Mumtaz and Gabor Pinter, 2016, VAR Models with Non-Gaussian Shocks, *Discussion Papers 1609*, Centre for Macroeconomics (CFM).

Clark, Todd E. and Francesco Ravazzolo, 2015, Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility, *Journal of Applied Econometrics* **30**(4), 551–575.

Cogley, Timothy and Thomas J. Sargent, 2005, Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S, *Review of Economic Dynamics* **8**(2), 262–302.

Curdia, Vasco, Marco Del Negro and Daniel L. Greenwald, 2013, Rare shocks, Great Recessions, *Technical report*.

Diebold, Francis X. and Canlin Li, 2006, Forecasting the term structure of government bond yields, *Journal of Econometrics* **130**(2), 337–364.

Diebold, Francis X., Glenn D. Rudebusch and S. BoragËĞan Aruoba, 2006, The macroeconomy and the yield curve: a dynamic latent factor approach, *Journal of Econometrics* **131**(1âĂŞ2), 309 – 338.

Dueker, Michael, Zacharias Psaradakis, Martin Sola and Fabio Spagnolo, 2011, Multivariate contemporaneous-threshold autoregressive models, *Journal of Econometrics* **160**(2), 311–325.

Filardo, Andrew J. and Stephen F. Gordon, 1998, Business cycle durations, *Journal of Econometrics* **85**(1), 99–123.

Fruhwirth-Schnatter, Sylvia, 2004, Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques, *Econometrics Journal* **7**(1), 143–167.

Gelfand, A. E. and D. K. Dey, 1994, Bayesian Model Choice: Asymptotics and Exact Calculations, *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(3), pp. 501–514.

Geweke, John, 1993, Bayesian Treatment of the Independent Student- t Linear Model, *Journal of Applied Econometrics* **8**(S), S19–40.

Geweke, John, 1999, Using simulation methods for bayesian econometric models: inference, development,and communication, *Econometric Reviews* **18**(1), 1–73.

Geweke, John, 2005, *Contemporary Bayesian Econometrics and Statistics*, Wiley.

Gneiting, Tilmann and Adrian E. Raftery, 2007, Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association* **102**, 359–378.

Gürkaynak, Refet S., Brian Sack and Jonathan H. Wright, 2007, The U.S. Treasury yield curve: 1961 to the present, *Journal of Monetary Economics* **54**(8), 2291 – 2304.

Hamilton, James D, 1989, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica* **57**(2), 357–84.

Hamilton, James D., 1994, *Time Series Analysis*, 1 edition, Princeton University Press.

Kalliovirta, Leena, Mika Meitz and Pentti Saikkonen, 2016, Gaussian mixture vector autoregression, *Journal of Econometrics* **192**(2), 485 – 498. Innovations in Multiple Time Series Analysis.

Kim, Chang Jin and Charles R Nelson, 1999, *State-Space Models with Regime Switching*, MIT Press.

Koop, Gary, 2003, *Bayesian Econometrics*, Wiley.

Lanne, Markku, 2006, Nonlinear dynamics of interest rate and inflation, *Journal of Applied Econometrics* **21**(8), 1157–1168.

Lanne, Markku and Helmut Lütkepohl, 2010, Structural Vector Autoregressions With Nonnormal Residuals, *Journal of Business & Economic Statistics* **28**(1), 159–168.

Meng, Xiao Li and Wing Hung Wong, 1996, Simulating ratios of normalizing constants via a simple identity: A theoretical exploration, *Statistica Sinica* pp. 831–860.

Mumtaz, Haroon and Paolo Surico, 2009, Time-varying yield curve dynamics and monetary policy, *Journal of Applied Econometrics* **24**(6), 895–913.

Nelson, C R. and A F Siegel, 1987, Parsimonious modeling of yield curves, *The Journal of Business* **60**(4), 473–489.

Sims, Christopher A., 1993, A Nine-Variable Probabilistic Macroeconomic Forecasting Model, *Business Cycles, Indicators and Forecasting*, NBER Chapters, National Bureau of Economic Research, Inc, pp. 179–212.

Sims, Christopher A., Daniel F. Waggoner and Tao Zha, 2008, Methods for inference in large multiple-equation Markov-switching models, *Journal of Econometrics* **146**(2), 255–274.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin and Angelika Van Der Linde, 2002, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.

Villani, M., R. Kohn and P. Giordani, 2009, Regression density estimation using smooth adaptive Gaussian mixtures, *Journal of Econometrics* **153**(2), 155–173. cited By 29.

A Appendix A: Importance density for calculating the bridge sampling estimate of the marginal likelihood

We assume that the importance density can be written as $p\left(B,\alpha_{i,S_{it}},A\right)\times p\left(\sigma_{i,S_{it}}^{2}\right)\times p\left(P_{i}\right)$. For the first set of parameters $B,\alpha_{i,S_{it}},A$ we use a mixture of three normals with means: $[\mu,\bar{u},\mu]$ and variances: $[\Omega,5\Omega,10\Omega]$ with weights: $[0.5,0.4,0.1]$. Here μ denotes the posterior mean, $\bar{\mu}$ is the draw consistent with the maximum value of the posterior over the MCMC draws and Ω is the posterior covariance matrix of the draws. $p\left(\sigma_{i,S_{it}}^{2}\right)$ is assumed to be a mixture of four normals with means: $[\mu,2\mu,1.5\mu,0.5\mu]$ and variances: $[0.1\Omega,\Omega,\Omega,\Omega]$ with weights: $[0.9,0.05,0.025,0.025]$. Finally $p\left(P_{i}\right)$ is assumed to be a Dirichlet distribution $D\left(u_{IJ}+\eta_{i,IJ}\right)$ where the number of transitions $\eta_{i,IJ}$ are calculated using the posterior median of $S_{i,t}$.

B Appendix B: Convergence

Figure 5 below shows the recursive means of the parameters of the two component model calculated every 20 retained draws. The means show little fluctuation providing evidence for convergence.

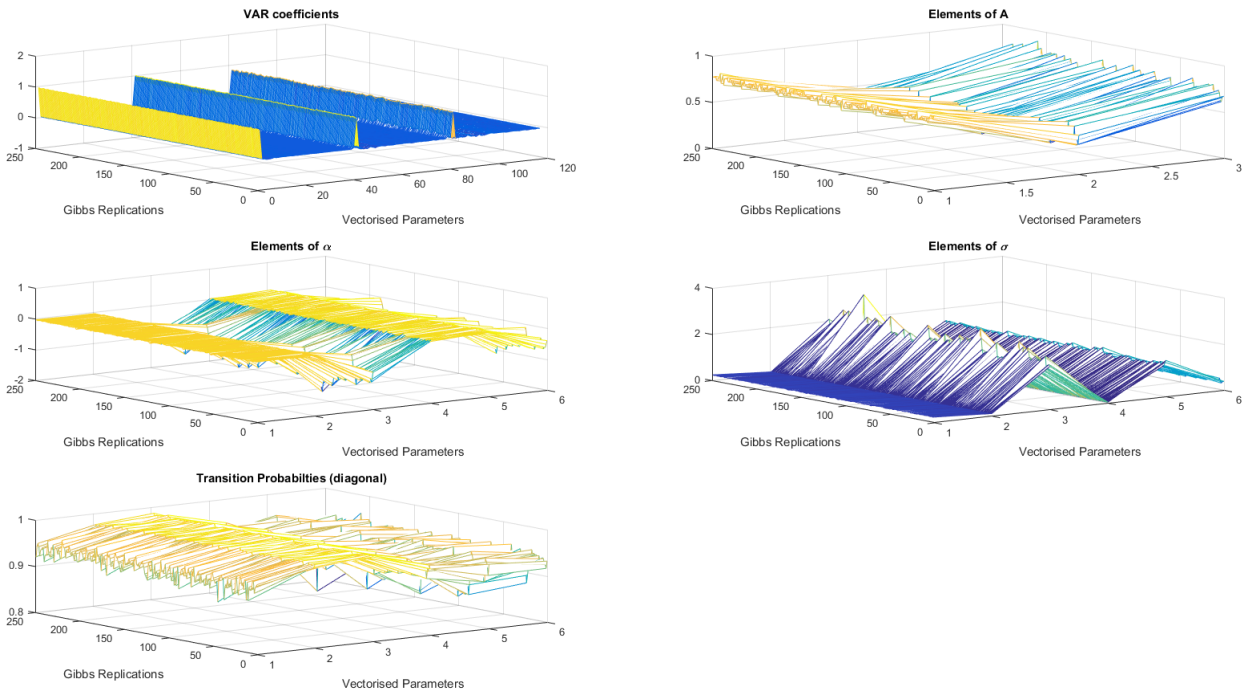

Figure 5: Recursive means of retained draws

C Appendix C: Ordering of yield curve factors

Table 5 provides the posterior medians of the switching parameters for a version of benchmark model in section 5.2 where the order of the factors is reversed. The results show that the pattern of regime switches remains very similar to the benchmark case. For the level and curvature equations, regime 1 is the high variance regime while regime 2 is the high variance state for the level factor.

Equation	$\alpha_{i,S_{it}=1}$	$\alpha_{i,S_{it}=2}$	$\sigma^2_{i,S_{it}=1}$	$\sigma^2_{i,S_{it}=2}$	P_i
Level	-0.0218	0.0262	0.3054	0.0337	$\begin{pmatrix} 0.9508 & 0.0174 \\ 0.0492 & 0.9826 \end{pmatrix}$
Slope	-0.0615	0.0055	0.1034	1.0899	$\begin{pmatrix} 0.9719 & 0.0860 \\ 0.0281 & 0.9140 \end{pmatrix}$
Curvature	-1.2419	-0.7963	2.0415	0.3374	$\begin{pmatrix} 0.9521 & 0.0672 \\ 0.0479 & 0.9328 \end{pmatrix}$

Table 5: Estimate regime dependent parameters