VAR Models with Non-Gaussian Shocks

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

February 29, 2016

Abstract

We introduce a Bayesian VAR model with non-Gaussian disturbances that are modelled with a finite mixture of normal distributions. Importantly, we allow for regime switching among the different components of the mixture of normals. Our model is highly flexible and can capture distributions that are fat-tailed, skewed and even multimodal. We show that our model can generate large out-of-sample forecast gains relative to standard forecasting models, especially during tranquil periods. Our model forecasts are also competitive with those generated by the conventional VAR model with stochastic volatility.

(*JEL:* C11, C32, C52)

*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.

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.

1 Introduction

This paper develops a VAR model with non-Gaussian shocks by modelling the disturbances with a finite mixture of normal distributions. The purpose of the model is to capture important non-linearities that characterise time-series data and can therefore affect the forecast performance of VAR models. Our model also allows for potentially abrupt switches between the components of the mixture of normal and provides a more flexible treatment of non-Gaussian shocks relative to existing approaches.

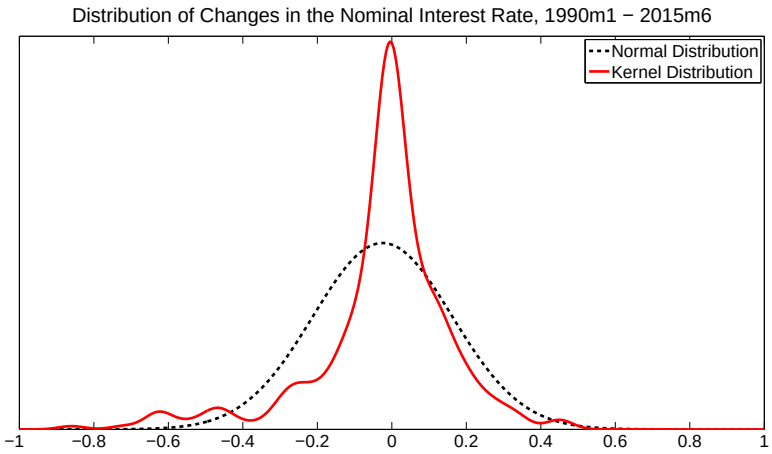
A number of approaches have already been proposed to capture non-linearities in the VAR literature. Models with stochastic volatility (Cogley and Sargent (2005); Primiceri (2005)) were developed to model time-variation in the error structure, which proved to be successful in increasing the out-of-sample forecast accuracy of VAR models (DAgostino, Gambetti, and Giannone (2013); Clark and Ravazzolo (2015)). Most of these papers use mixture of normals approximation as in Kim, Shephard, and Chib (1998) to estimate volatility with the Gibbs sampler. Nevertheless, these models assume smoothly drifting second moments. Therefore they are designed to capture persistent (low-frequency) changes in volatilities while being less capable of modelling transient (high-frequency) changes in volatility such as rare, fat-tailed events. This point was made in different contexts by Jacquier, Polson, and Rossi (2004) and Curdia, del Negro, and Greenwald (2014). Models with fat-tails relaxed the assumption of Normality by introducing Student's t-distributed shocks (Ni and Sun, 2005), and the results of Chiu, Mumtaz, and Pinter (2015) show that accounting for both stochastic volatility and fat-tails can improve the out-of-sample forecast accuracy of VAR models. Recent advances in computational algorithms allowed these features to be incorporated in larger VAR models (Carriero, Clark, and Marcellino (2016); Chan (2015))

None of these models can however capture abrupt changes in economic dynamics similar to the recent Great Recession. Markov switching VAR models (Sims and Zha (2006); Sims, Waggoner, and Zha (2008); Hubrich and Tetlow (2014)) have been designed to model such turning points, but very few applications have looked at implications for forecast accuracy. This is mainly because of the significant computational challenge that these models impose during a real-time forecasting evaluation exercise (Clark and Ravazzolo (2015)).

Our model builds on all the previous approaches by constructing a framework of finite mixtures of normals to account for non-normal shocks. In particular, a markov switching process is assumed to potentially capture any persistence in the regimes. We allow for independent Markov chains to govern the behaviour of each orthogonal shock, which allows for greater flexibility in the modelling process.

Our paper is related to Kalliovirta, Meitz, and Saikkonen (2014), who proposes a Gaussian Mixture Vector Autoregression. Our model is different from theirs in the following ways: (i) while in their set-up all coefficients in the system switch regimes at the same time, our proposed model focuses on the modelling of the regime switches in the

Figure 1: Empirical Motivation



shocks; (ii) our proposed model allows each shock to have an independent distribution; (iii) we estimate our model in a Bayesian framework and thus the predictive posterior density can be estimated.

To empirically motivate the need for our modelling strategy, Figure 1 plots the empirical distribution (using kernel estimation) of monthly changes in the 3-month T-bill rate over 1990m1-2015m6 against a fitted normal distribution. The discrepancy between the two distributions is obvious, and there are at least three things to notice.

First, the kernel distribution is much more peaked. This is because the sample is dominated by the long recent period of the zero lower bound on the nominal interest rate, which makes most of the distribution centred tightly around zero. Second, the kernel distribution is more fat-tailed and highly negatively skewed. This is because the sample features three recessions that generated strong monetary policy responses leading to large negative values of interest rate changes. In contrast, expansions are in general not as abrupt as recessions, therefore monetary policy generates large interest rate changes less frequently. Third, there are some signs of multimodality in the kernel distribution: small (0-30bp) and large (45-65bp) interest rate cuts are more frequent than medium-sized (30-45bp) cuts. This is because monetary policy either fine tunes or responds aggressively to recessions.

We argue that the model we propose can successfully capture some of these peculiarities of aggregate time-series data. We show that this can lead to sizable gains in forecast accuracy relative to other, frequently used BVAR models and models with Student's t-distributed shocks (TVAR), especially during tranquil periods. We also provide evidence that the forecasts generated by our proposed model are competitive with respect to the VAR model with stochastic volatility as investigated by Clark and Ravazzolo (2015) .

The structure of the paper is as follows. Section 2 provides a description of our baseline model. Section 3 explains the priors and the Gibbs sampling algorithm. Section 4 explains the data and the metrics used to measure forecast accuracy. Section 5 describes the estimated regimes using our model. Section 6 presents the results of pseudo-out-of-

sample forecast comparison. Section 7 concludes.

2 BVAR Model with Non-normal Disturbances

The model presented in this section is a multivariate time series model with disturbances that are allowed to be non-normal. The non-normality is introduced in the model through a finite mixture of normals.

The BVAR model is defined as follows

$$y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + u_t \qquad t = 1, \ldots, T. \tag{2.1}$$

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; and u_t are heteroscedastic shocks associated with the VAR equations. In particular, we assume that the covariance matrix of u_t is defined as

$$cov(u_t) = \Sigma = A^{-1} H A^{-1'} \tag{2.2}$$

where A is a lower triangular matrix. The orthogonalised shocks of the model are then given as

$$e_t = A u_t \tag{2.3}$$

The shock to the ith equation is assumed to follow:

$$e_{it} = \alpha_{i,S_{it}} + \sigma_{i,S_{it}} \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0,1) \tag{2.4}$$

where $S_{it} = 1, 2, ...M$ denotes the unobserved components or regimes. As explained in Koop (2003) and Geweke (2005), the formulation in equation 2.4, describes a mixture of M distributions where each component is $N(\alpha_i, \sigma_i^2)$. The state variable S 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

$$P(S_{i,t} = J \backslash S_{i,t-1} = I) = p_{i,IJ} \tag{2.5}$$

This formulation captures possible dependence in the time series data used in the paper but allows for the possibility of rapid transitions across components.

The specification in equations 2.4 implies that orthogonalised residuals e_t are non-Gaussian. As the number of components increases, 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, which is a special case of our more general model. 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 $p\left(\frac{1}{\sigma_{i,t}}\right) = \prod_{t=1}^{T} \Gamma(1, v_i)$ 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)). Using equation 2.4, 2.3 and 2.1 one obtains:

$$y_t = B_1 y_{t-1} + \cdots + B_p y_{t-p} + A^{-1}\left(\alpha_{S_t} + \sigma_{S_t}\varepsilon_t\right) \tag{2.6}$$

where α_{S_t} and σ_{S_t} are respectively vectors of $\alpha_{i,S_{it}}$ and $\sigma_{i,S_{it}}$. The VAR model in equation 2.6 has switching intercepts $A^{-1}\alpha_{S_t}$, and reduced form residuals with switching variances. Note that unlike standard MSVAR models, there are n independent Markov chains in the proposed model that govern the behaviour of each orthogonal error. The implied reduced form intercepts and residuals are a linear combination of these and thus implies a more complex structure than MSVARs with switching intercepts and variance where, typically, one Markov process controls the regime shifts in the system.

Equation 2.6 also shows that the reduced form residuals are a linear combination of non-normal orthogonal shocks. Therefore the setup imparts a flexible specification for u_t which can also depart from Gaussianity in interesting ways. It implies, however, that the ordering of the variables in y_t can affect the forecast from the model. In the analysis below, we order the variables in an economically meaningful manner and check the robustness of the results to this choice.

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

To define priors for the VAR dynamic coefficients, we follow the dummy observation approach of Banbura, Giannone, and Reichlin (2010). We assume Normal priors, $p(B) \sim N(B_0, S_0)$, where $B = vec([B_1, B_2, ..B_p])$, $B_0 = (x_d' x_d)^{-1}(x_d' y_d)$ and $S_0 = (Y_D - X_D b_0)'(Y_D - X_D b_0) \otimes (x_d' x_d)^{-1}$. The priors are implemented by the dummy observations y_D and x_D that are defined as:

$$y_D = \begin{bmatrix} \frac{diag(\gamma_1 s_1 \ldots \gamma_n s_n)}{\tau} \\ 0_{n\times(p-1)\times n} \\ \ldots\ldots\ldots\ldots \\ diag\left(s_1 \ldots s_n\right) \\ \ldots\ldots\ldots\ldots \\ 0_{1\times n} \end{bmatrix}, \qquad x_D = \begin{bmatrix} \frac{J_P \otimes diag(s_1 \ldots s_n)}{\tau} & 0_{np\times 1} \\ 0_{n\times np} & 0_{n\times 1} \\ \ldots\ldots\ldots\ldots & \\ 0_{1\times np} & c \end{bmatrix} \tag{3.1}$$

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 c is the tightness of the prior on the constant term. We use set $\tau = 0.1$ for all models. The scaling factor s_i are set using the standard deviation of the residuals from the individual AR(1) equations. We set $c = 1/1000$, implying a relatively flat prior on the constant. In addition, we introduce priors on the sum of lagged coefficients by defining the following dummy observations:

$$y_S = \frac{diag\left(\gamma_1 \mu_1 \ldots \gamma_n \mu_n\right)}{\lambda}, \qquad x_S = \left[\frac{(1_{1\times p}) \otimes diag\left(\gamma_1 \mu_1 \ldots \gamma_n \mu_n\right)}{\lambda} \quad 0_{n\times 1}\right] \tag{3.2}$$

where μ_1 to μ_n denote the sample means of the endogenous variables using a training sample, and the tightness of period on this sum of coefficients is set to $\lambda = 10\tau$.

The prior for the non-zero and non-one elements A_k is $P(A_k) \sim N\left(A_0, \Sigma_0\right)$ where $A_0 = A_{ols}$ from the Cholesky decomposition of the OLS estimate of the VAR error covariance matrix and $\Sigma_0 = 10$.

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

The prior for the non zero elements of the transition probability matrix $p_{i,IJ}$ is of the following form: $P(p_{i,IJ}) = D\left(u_{IJ}\right)$ where $D(.)$ denotes the Dirichlet distribution and $u_{IJ} = 20$ 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 95%.

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. $G\left(B \backslash S_{it}, \alpha_{i,S_{it}}, \sigma^2_{i,S_{it}}, A, p_{i,IJ}\right)$: The conditional posterior distribution of the VAR coefficients conditional on the remaining parameters is linear and Gaussian: $N\left(B_{T\backslash T}, P_{T\backslash T}\right)$.

We use the Kalman filter to calculate $B_{T\backslash T}$ and $P_{T\backslash T}$. 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 $Y_t = vec(y_t)$, $X_t = I_n \otimes [y_{t-1}, y_{t-2}, ..., y_{t-p}]$, $\mu_t = A^{-1}\alpha_{i,S_{it}}$, $R_t = A^{-1} diag(\sigma_{i,S_{it}})$, $V_t = \varepsilon_{it}$. 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 $t = 1, 2..T$

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

The final iteration of the filter delivers $B_{T\backslash T}$ and $P_{T\backslash T}$. The VAR coefficients can then be drawn from the multivariate Normal distribution.

2. $G\left(A\backslash B, S_{it}, \alpha_{i,S_{it}}, \sigma^2_{i,S_{it}}, p_{i,IJ}\right)$: Conditional on the VAR coefficients B, the model can be written as $e_t = Au_t$. For a four variable VAR, this system is given as

$$\begin{pmatrix} \alpha_{1,S_{1t}} + \sigma_{1,S_{1t}}\varepsilon_{1t} \\ \alpha_{2,S_{2t}} + \sigma_{2,S_{2t}}\varepsilon_{2t} \\ \alpha_{3,S_{3t}} + \sigma_{3,S_{3t}}\varepsilon_{3t} \\ \alpha_{4,S_{4t}} + \sigma_{4,S_{4t}}\varepsilon_{4t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a_1 & 1 & 0 & 0 \\ a_2 & a_3 & 1 & 0 \\ a_4 & a_5 & a_6 & 1 \end{pmatrix} \begin{pmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ u_{4t} \end{pmatrix} \tag{3.3}$$

where $[a_1, ...a_6]$ 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 is a linear regression with a known variance. Given 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 $y_t^* = \frac{u_{2t} - \alpha_{2,S_{2t}}}{\sum_{j=1}^{M} \sigma_{2,S_{2t}} \times D_{t,j}}$ 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 $N(M^*, V^*)$

$$\begin{aligned} M^* &= \left(\Sigma_0^{-1} + x_t^{*\prime} x_t^*\right)^{-1} \left(\Sigma_0^{-1} A_0 + x_t^{*\prime} y_t^*\right) \\ V^* &= \left(\Sigma_0^{-1} + x_t^{*\prime} x_t^*\right)^{-1} \end{aligned}$$

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

3. $G\left(\alpha_{i,S_{it}} \backslash B, S_{it}, \sigma^2_{i,S_{it}}, A, p_{i,IJ}\right)$: 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

$$e_{it} = \alpha_{i,S_{it}} + \sigma_{i,S_{it}} \varepsilon_{it} \tag{3.4}$$

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 to each equation i.

4. $G\left(\sigma^2_{i,S_{it}} \backslash B, S_{it}, \alpha_{i,S_{it}}, A, p_{i,IJ}\right)$: Conditional on a draw for B, α_i, S_{it} the conditional posterior for $\sigma_{i,S_{it}}$ is inverse Gamma $IG\left(\bar{\sigma}, \tilde{T}\right)$

$$\begin{aligned} \bar{\sigma} &= \bar{e}_{it}' \bar{e}_{it} + \sigma_0 \\ \bar{T} &= \tilde{T} + v_0 \end{aligned}$$

where $\bar{e}_{it}$ are the residuals from equation 3.4 for $S_{it} = j$. The same procedure is repeated for regime $j = 1...M$ and each equation i.

5. $G\left(p_{i,IJ} \backslash B, S_{it}, \alpha_{i,S_{it}}, A, \sigma^2_{i,S_{it}}\right)$: 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 for the ith orthogonal error.

6. $G\left(S_{it}\backslash\sigma^2_{i,S_{it}}, B, \alpha_{i,S_{it}}, A, p_{i,IJ}\right)$:

   Following Kim and Nelson (1999) we use a multi-move Gibbs step to sample from the conditional posterior of S_{it}. The Markov property of S_{it} implies that

$$f\left(S_{it}|y_t\right) = f\left(S_{iT}|e_{iT}\right)\prod_{t=1}^{T-1} f\left(S_{it}|S_{it+1}, e_{it}\right) \tag{3.5}$$

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

   - Calculating $f\left(S_{iT}|e_{iT}\right)$: The Hamilton (1989) filter provides $f\left(S_{it}|e_{it}\right), t = 1, ...T$. Denoting $\hat{\xi}_{i,t}$ as a vector where the jth element equals $\Pr\left(S_t = j\right)$, the filter iterates on the following two equations:

$$\hat{\xi}_{i,t\backslash t-1} = P_i\hat{\xi}_{i,t-1\backslash t-1}$$

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

     where P_i denotes the transition probability matrix and:

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

     The last iteration of the filter delivers $f\left(S_{iT}|e_{it}\right)$.

   - Calculating $f\left(S_{it}|S_{it+1}, e_{it}\right)$: Kim and Nelson (1999) show that

$$f\left(S_{it}|S_{it+1}, e_{it}\right) \propto f\left(S_{it+1}|S_{it}\right) f\left(S_{it}|e_{it}\right) \tag{3.6}$$

     where $f\left(S_{it+1}|S_{it}\right)$ is the transition probability matrix and $f\left(S_{it}|e_{it}\right)$ is obtained via the Hamilton (1989) filter in the previous step. Kim and Nelson (1999) (pp 214) show how to sample S_t from (3.6).

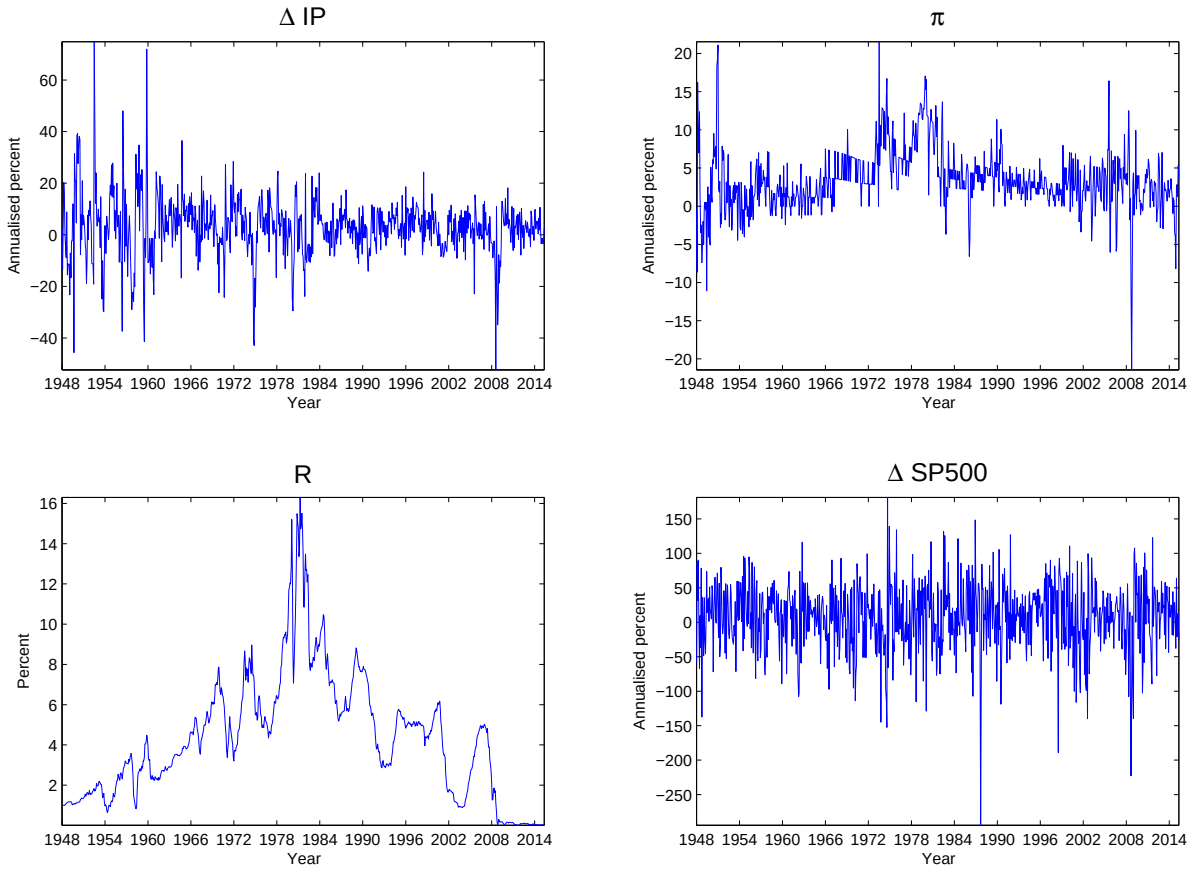
In the technical appendix, we present a simple Monte Carlo experiment that shows that this algorithm displays a satisfactory performance in approximating the posterior moments. The appendix also presents the Gibbs algorithm for the model that features orthogonalised errors that have a student-T distribution.

4 Data and Forecasting Methodology

4.1 Data and Forecasts

We employ monthly data on the following US variables: (1) annualised growth rate of Industrial Production, (2) annualised CPI inflation rate, (3) three-month Treasury bill

Figure 2: Plots of the US data



rate and (4) annualised S&P500 stock returns. The data spans the sample January 1947 to June 2015. The first three variables are obtained from the FRED database, while the stock market index is downloaded from Global Financial Database. The four series are plotted in Figure 2.

The forecasting exercise is carried out recursively. The forecasting models are estimated over the initial sample January 1947 to December 1960. They are then re-estimated 654 times adding one month of data at each iteration until June 2014. At each iteration, we construct the forecast density for the models:

$$P\left(\hat{y}_{t+k} \backslash y_t\right) = \int P(\hat{y}_{t+k} \backslash y_t, \Psi_{t+k}) P\left(\Psi_{t+k} \backslash \Psi_t, y_t\right) P\left(\Psi_t \backslash y_t\right) d\Psi \tag{4.1}$$

where $k = 1, 2, ..12$ and Ψ denotes the model parameters. The last term in equation 4.1 represents the posterior density of the parameters that is obtained via the MCMC simulation. The preceding two terms denote the forecast of the (time-varying) parameters and the data that can be obtained by simulation. The point forecast is obtained as the mean of the forecast density.

Our measure of point forecast performance is the Root Mean Squared Error (RMSE), while the Continuous Ranked Probability Score (CRPS), as in Hersbach (2010); Jolliffee

and Stephenson (2003) and implemented by Shrestha (2014), is our measure of density forecast accuracy. 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 (Gneiting and Raftery (2007); Clark and Ravazzolo (2015); Smith and Vahey (2015)).

4.2 Forecasting Models and Forecast Evaluation

We consider the forecasting performance of VAR models with non-Gaussian errors to a standard Bayesian VAR. In particular, we consider two versions of the benchmark model shown in equation 2.1, allowing for the possibility of two (M2-VAR) and three (M3-VAR) components or regimes in the model for the orthogonal shocks (equation 2.4).[1] We consider two more models: the first one being a VAR model with fat-tailed errors (henceforth 'TVAR'), which features a scaled mixture of normals for the orthogonal errors; and the second one being the VAR model with stochastic volatility (SVOL-VAR).

5 Full sample estimation results

In this section we investigate the estimated regimes implied by our M2-VAR. The estimated values for α_i and σ_i, respectively the mean and the variance of shocks for each equation under the two regimes, are shown in Table 1.

Table 1: Estimated values for the coefficient α_i and σ_i in equation 2.4.

	Regime 1		Regime 2	
	α_i	σ_i	α_i	σ_i
ΔIP	-0.381	18.06	2.432	5.695
	(-1.057,-0.061)	(16.72,19.58)	(1.714,3.109)	(5.410,6.015)
π	-0.053	2.11	0.323	5.349
	(-0.157,-0.009)	(1.965,2.232)	(0.010,0.826)	(4.820,5.930)
R	-0.049	0.7579	0.007	0.1340
	(-0.086,-0.014)	(0.7025,0.8220)	(-0.006,0.021)	(0.1259,0.1240)
$\Delta SP500$	-23.50	73.56	19.48	38.95
	(-32.83,-13.46)	(66.72,82.27)	(15.17,23.98)	(36.96,40.97)

Note: Numbers in brackets indicate intervals of 10 and 90 percent.

It can be observed that regime 1 is associated with shocks of negative mean and high variance for ΔIP, R, and $\Delta SP500$, indicating that regime 1 can be interpreted as 'low mean and high variance' for the shocks of industrial growth rate, the short-term rate and the stock market returns. As for π, regime 1 is associated with shocks of negative mean and low variance, implying that regime 1 can be interpreted as 'low mean and low variance' for the inflation rate equation.

[1] We also consider models with four and five unobserved regimes, whose forecasting performances are very similar to our M3-VAR models. In the interest of space we do not report their results here.

Figure 3 plot the time-varying estimated regimes each of the variables. A few observations are in order:

- The model manages to capture the great moderation period for the *industrial production growth* equation, where the mean of shocks is high but the variance is low. This is indicated by the a prolonged period of regime 2 between the mid-1980s and 2007. In contrast, the immediate post-world war period, part of the 1970s and the 1980s, and the Great Recession are characterised by regime 1, where the shocks are of low mean and high variance;

- As for the *inflation rate* equation, the model indicates shocks of regime 2 (high mean and high variance) in the late 1940s and early 1950s, and various periods of 1970s, as well as around the Great Recession;

- The model captures negative interest rate shocks with regime 1 (negative mean and high variance) in periods where there is a downward trend in the short-term interest rate, noticeably at the end of 1950s, the early 1990s, late 1990s and the Great Recession. During the episode of 1970s and mid 1980s where the short-term rate is volatile, the estimated regime is also fluctuating a lot, with most of the time being in regime 1 which is characterised by high variance.

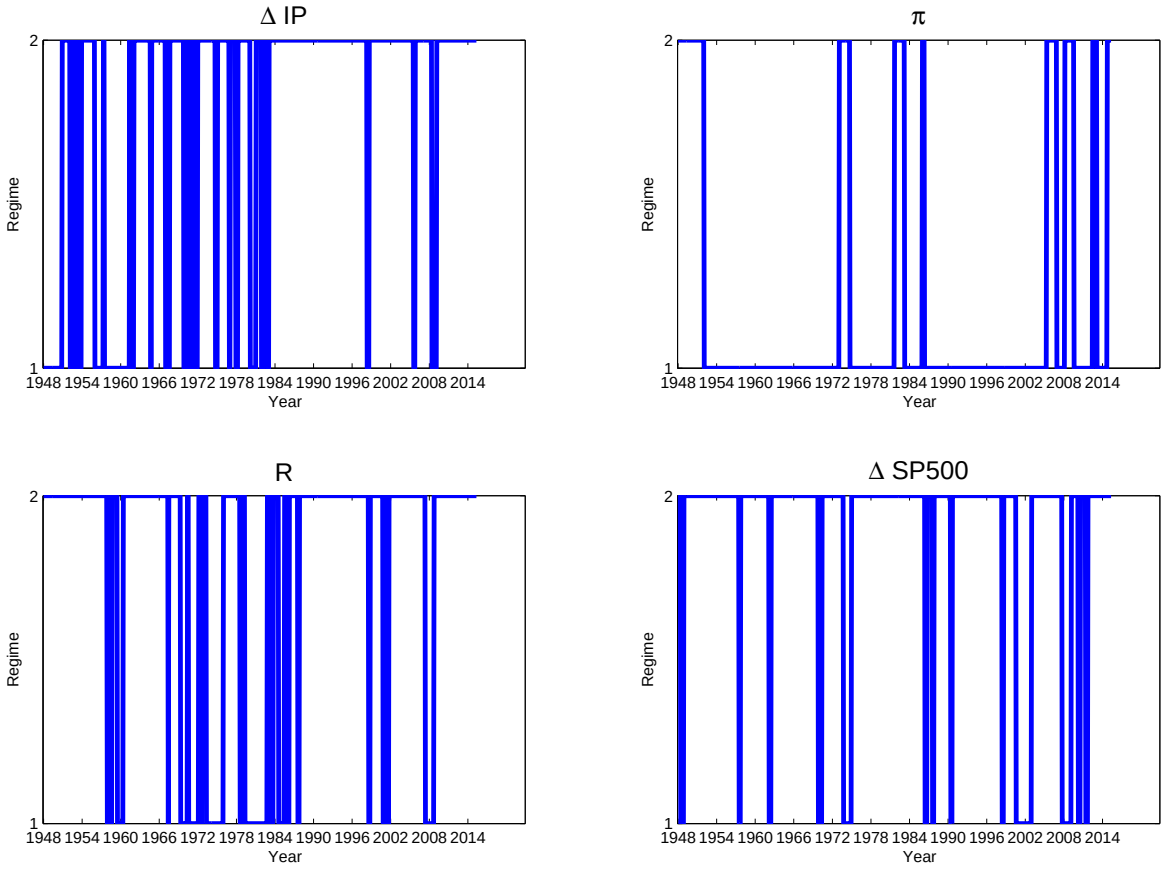
- As for the $\Delta SP500$ equation, the model is able to capture episodes of shocks with volatile and negative mean at times of high stock market volatility throughout the sample.

6 Forecasting Results

Table 2 presents the average point forecast performance measured by the RMSE and the average density forecast performance measured by the CRPS for each model relative to that obtained using the BVAR *for the full sample*. The relative forecast gains are presented in ratios, therefore ratios with values being less (more) than one suggesting superior (inferior) forecast performance relative to the BVAR.

The results suggest that our baseline models tend to outperform both the BVAR and the TVAR models in forecasting output and interest rates in terms of both point and density forecasts. For example, the M3-VAR model yields a 2-6% more accurate point forecast over the 1-3 month horizon relative to the BVAR, while the same model yields a 13% more accurate density forecast over the same horizon relative to the BVAR. In general, our model generates more accurate forecasts for these two variables than the TVAR. Interestingly, our proposed model delivers relatively little forecast gains for inflation and no forecast gains for stock returns compared to the BVAR model. Our forecasts are also generally competitive with those of the SVOL-VAR model.

Figure 3: Estimated regimes (regime 1 or 2) for each equation in the two-component model (M2-VAR). Sample period:1948-2014.



Note: Regime 1 is associated with shocks of low-mean and high variance for ΔIP, R and $\Delta SP500$. As for π, regime 1 is associated with shocks of low-mean and low-variance.

Table 2: Forecast Performance Relative to BVAR: Rolling window for 1961-2014

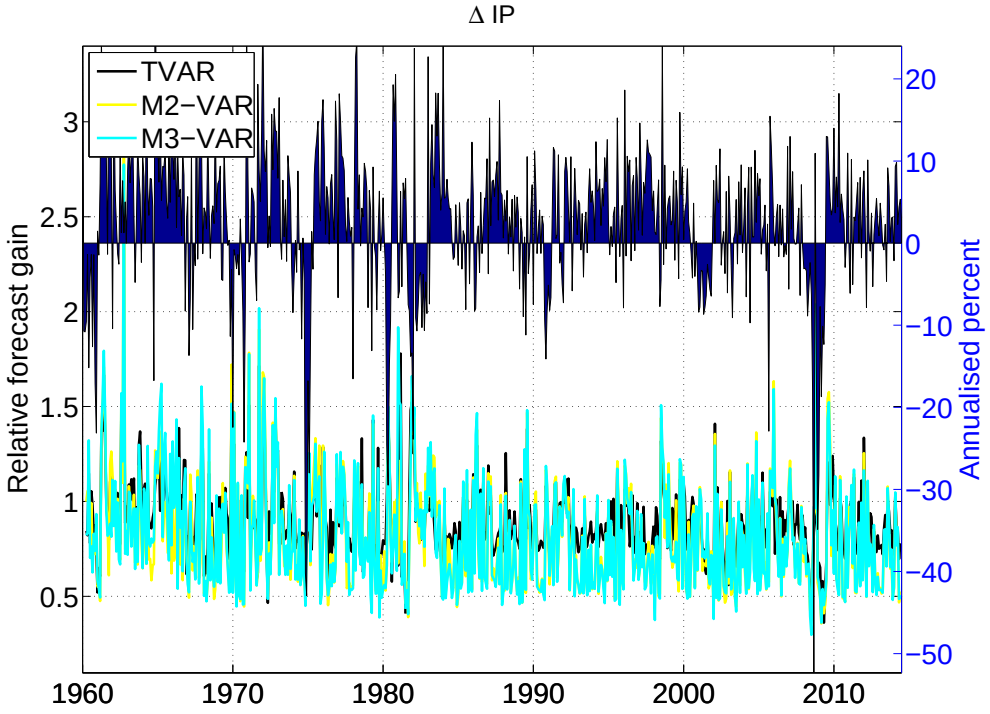
		RMSE				CRPS			
		1M	3M	6M	12M	1M	3M	6M	12M
ΔIP	TVAR	0.976	0.965	1.000	1.023	0.914	0.903	0.946	0.983
	M2-VAR	0.959	0.957	0.975	1.004	0.883	0.862	0.912	0.959
	M3-VAR	0.966	0.966	0.983	1.010	0.888	0.867	0.917	0.965
	SVOL-VAR	0.955	0.952	0.960	0.981	0.897	0.876	0.910	0.955
π	TVAR	0.996	0.998	1.052	1.108	0.972	0.976	0.991	1.015
	M2-VAR	1.015	1.005	1.015	1.028	0.967	0.977	0.997	1.038
	M3-VAR	1.010	0.998	1.003	1.022	0.970	0.971	0.983	1.039
	SVOL-VAR	1.010	0.998	1.001	1.006	0.963	0.960	0.966	0.987
R	TVAR	0.988	0.950	1.291	1.283	0.938	0.919	0.932	0.969
	M2-VAR	0.975	0.932	0.915	0.919	0.890	0.875	0.896	0.951
	M3-VAR	0.981	0.941	0.929	0.948	0.870	0.867	0.907	0.997
	SVOL-VAR	0.982	0.931	0.905	0.902	0.854	0.830	0.845	0.905
ΔSP500	TVAR	0.999	0.996	1.066	1.058	1.000	0.998	0.999	1.003
	M2-VAR	1.020	1.011	1.013	1.016	1.018	1.005	1.016	1.023
	M3-VAR	1.015	1.013	1.015	1.020	1.022	1.018	1.024	1.036
	SVOL-VAR	1.001	0.998	1.000	1.002	0.998	0.998	1.005	1.017

Note: The table presents the average point and density forecast measures based on 655 recursive estimations of the five models. Sample period: 1961-2014.

To provide a pictorial representation of forecasting results, Figures 4–7 show the evoluation of CRSP measures of the three-month ahead forecasts for the TVAR, M2-VAR and M3-VAR models relative to the BVAR. The coloured lines (units in left axis) are constructed based on the recursive estimation of the forecasting models. When the line is below (above) one, then the given model delivers a superior (inferior) forecasting performance relative to the BVAR.

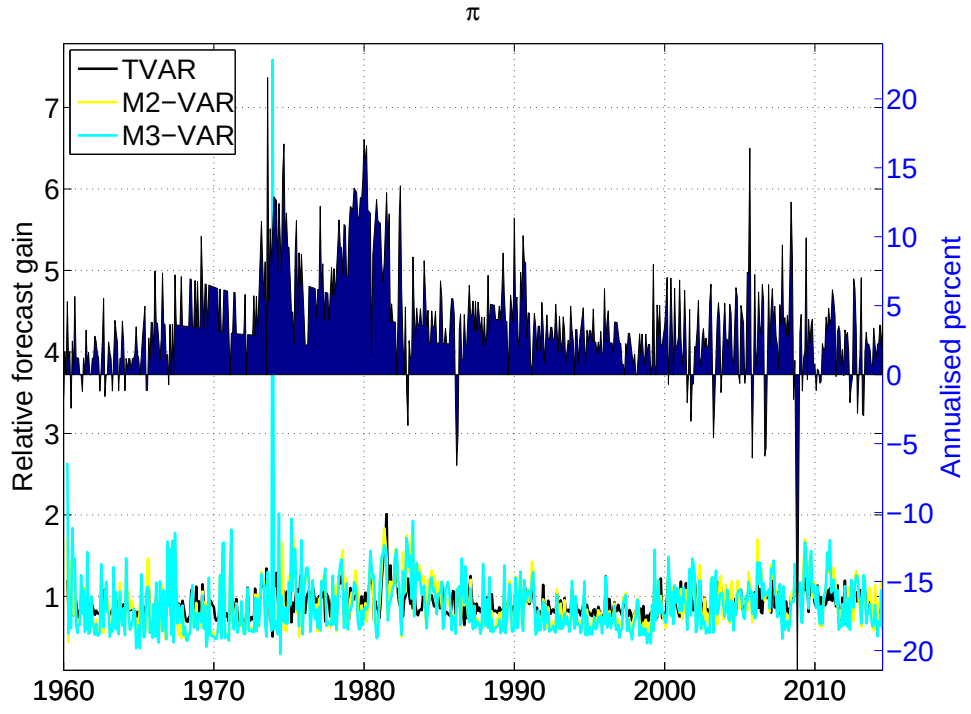
The results suggest that there is considerable asymmetry in forecast performance of the models depending on the time period in question. Specifically, there are at least two things to notice. First, there is evidence that the simpler linear BVAR model performs relatively better than our proposed models during the highly volatile 1970s and 1980s. However, the situation is reversed afterwards: our models do much better during the relatively more tranquil periods such as the Great Moderation and during the aftermath of the Great Recession. Second, allowing for fat-tails may not be enough to generate accurate density forecasts, and modelling asymmetries can result in additional forecast gains. The most striking example for this is demonstrated by Figure 6 which plots the results for the nominal interest rate. While both the TVAR and our mixture models dominate the BVAR during the recent period of the zero lower bound (2010-2015), the M2-VAR and M3-VAR models strongly dominate TVAR during this period.

Figure 4: Time-series of CRSP of the industrial production growth rate



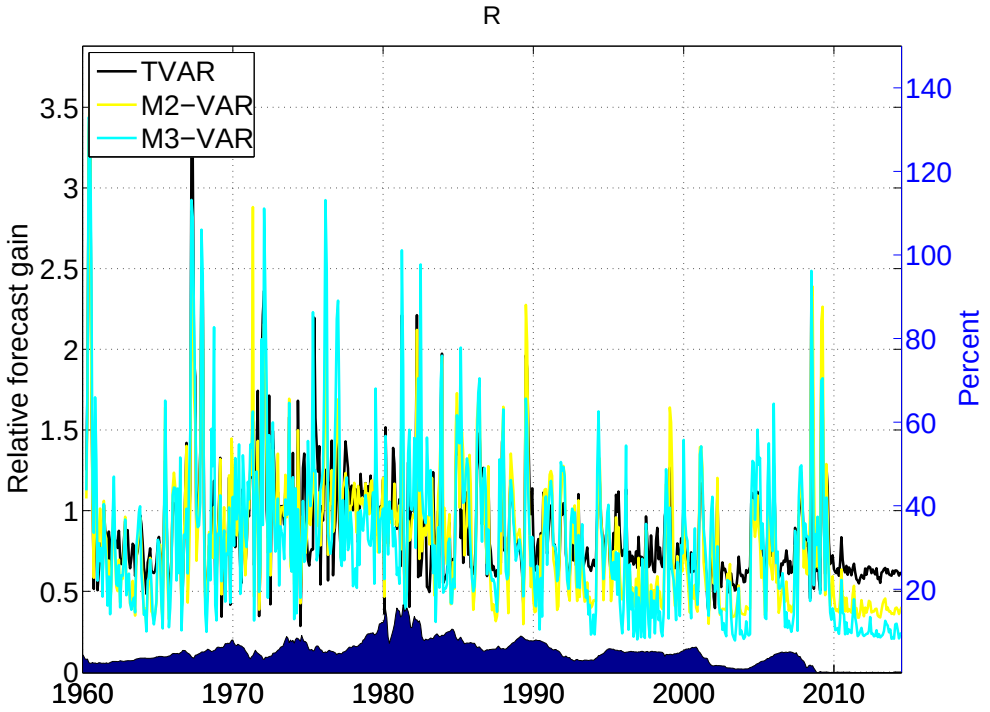
Note: The figure plots the three-month ahead density forecast performance (measured by CRSP) of the TVAR (black line), M2-VAR (yellow line), M3-VAR (green line) relative to the BVAR model. The left axis shows the units of the relative forecast gain with values being less (more) than one suggesting superior (inferior) forecast performance relative to the BVAR. The dark blue lines are the data whose units are measured by the right axis.

Figure 5: Time-series of CRSP of the inflation rate



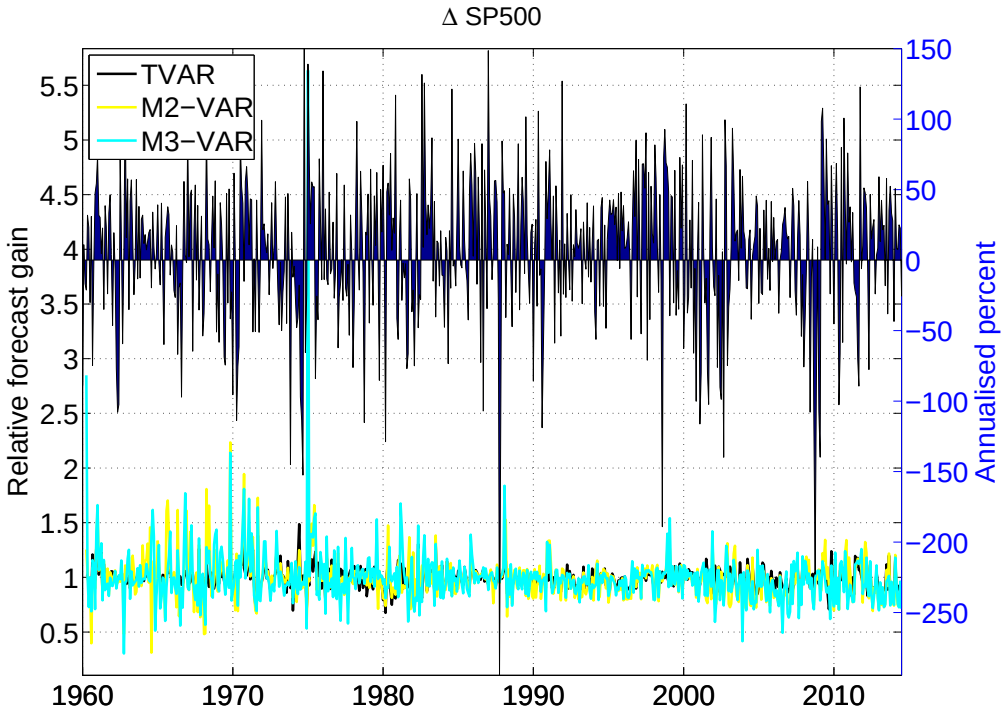
Note: The figure plots the three-month ahead density forecast performance (measured by CRSP) of the TVAR (black line), M2-VAR (yellow line), M3-VAR (green line) relative to the BVAR model. The left axis shows the units of the relative forecast gain with values being less (more) than one suggesting superior (inferior) forecast performance relative to the BVAR. The dark blue lines are the data whose units are measured by the right axis.

Figure 6: Time-series of CRSP of the short-term interest rate



Note: The figure plots the three-month ahead density forecast performance (measured by CRSP) of the TVAR (black line), M2-VAR (yellow line), M3-VAR (green line) relative to the BVAR model. The left axis shows the units of the relative forecast gain with values being less (more) than one suggesting superior (inferior) forecast performance relative to the BVAR. The dark blue lines are the data whose units are measured by the right axis.

Figure 7: Time-series of CRSP of SP500 returns



Note: The figure plots the three-month ahead density forecast performance (measured by CRSP) of the TVAR (black line), M2-VAR (yellow line), M3-VAR (green line) relative to the BVAR model. The left axis shows the units of the relative forecast gain with values being less (more) than one suggesting superior (inferior) forecast performance relative to the BVAR. The dark blue lines are the data whose units are measured by the right axis.

To check whether our proposed models do indeed better at forecasting during tranquil period, we compute the average point and density forecast measures of Table 2 for the subsample 1990-2014. Table 3 confirms that indeed all our mixture models deliver more accurate point and density forecasts over virtually all horizons during this relatively more tranquil period.

The results indicate improvement in forecasting all of the four variables relative to the BVAR in this sample. But the improvement is far more substantial for output growth and interest rates. For example, the M3-VAR model delivers a 35-50% more accurate point forecast and a 20-55% more accurate density forecast of the nominal interest rate compared to the BVAR. We argue that this is due to the peculiar shape of the interest rate distribution (Figure 1) that models with Normal distribution cannot easily capture. Our flexible modelling of the shock distribution can do a much better job in fitting the data, which results in significant improvements in out-of-sample forecast performance. There is also evidence that our proposed models outperform the SVOL-VAR in forecasting output growth in this subsample.

Table 3: Forecast Performance Relative to BVAR: Rolling window for 1990-2014

		RMSE				CRPS			
		1M	3M	6M	12M	1M	3M	6M	12M
	TVAR	0.828	0.787	0.794	0.814	0.775	0.734	0.784	0.817
	M2-VAR	0.814	0.780	0.784	0.797	0.736	0.684	0.727	0.750
ΔIP	M3-VAR	0.810	0.782	0.786	0.801	0.734	0.688	0.732	0.751
	SVOL-VAR	0.824	0.799	0.808	0.829	0.753	0.726	0.779	0.811
	TVAR	0.993	0.996	0.963	0.940	0.981	0.939	0.890	0.860
	M2-VAR	1.017	1.004	0.978	0.962	0.981	0.945	0.902	0.880
π	M3-VAR	0.996	0.982	0.952	0.936	0.963	0.925	0.977	0.857
	SVOL-VAR	1.002	0.988	0.955	0.934	0.970	0.923	0.879	0.854
	TVAR	0.521	0.519	0.562	0.661	0.484	0.520	0.612	0.781
	M2-VAR	0.498	0.492	0.535	0.630	0.463	0.469	0.559	0.733
R	M3-VAR	0.506	0.501	0.551	0.657	0.448	0.468	0.579	0.791
	SVOL-VAR	0.508	0.502	0.543	0.635	0.439	0.457	0.551	0.733
	TVAR	0.992	0.981	0.975	0.971	0.997	0.994	0.978	0.985
	M2-VAR	0.996	0.987	0.980	0.972	0.994	0.997	0.986	0.993
ΔSP500	M3-VAR	0.984	0.977	0.971	0.967	0.982	0.995	0.987	0.995
	SVOL-VAR	0.990	0.978	0.972	0.968	0.982	0.980	0.977	0.995

Note: The table presents the average point and density forecast measures based recursive estimations of the five models. Sample period: 1990-2014.

7 Conclusion

This paper proposed a new, flexible approach to modelling shocks in VAR models, whereby we approximated the disturbances with a finite mixture of normal distributions, and allowed for regime switching among the different components of the mixture of normals. The model is thereby more flexible than existing models of stochastic volatility and more general than existing models with Markov-switching. We showed that the pro-

posed model can generate substantial out-of-sample forecast gains relative to standard BVAR models, especially during tranquil periods such as the Great Moderation and the aftermath of the Great Recession.

References

Banbura, M., D. Giannone, and L. Reichlin (2010): "Large Bayesian vector auto regressions," *Journal of Applied Econometrics*, 25(1), 71–92.

Carriero, A., T. E. Clark, and M. Marcellino (2016): "Common Drifting Volatility in Large Bayesian VARs," *Journal of Business and Economic Statistics*, forthcoming.

Chan, J. 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.

Chiu, C.-W. J., H. Mumtaz, and G. Pinter (2015): "Forecasting with VAR models: fat tails and stochastic volatility," Bank of England working papers 528, Bank of England.

Clark, T. E., and F. Ravazzolo (2015): "Macroeconomic Forecasting Performance Under Alternative Specifications of Time-varying Volatility," *Journal of Applied Econometrics*, Forthcoming.

Cogley, T., and T. 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, V., M. del Negro, and D. L. Greenwald (2014): "Rare Shocks, Great Recessions," *Journal of Applied Econometrics*, 29(7), 1031–1052.

DAgostino, A., L. Gambetti, and D. Giannone (2013): "Macroeconomic forecasting and structural change," *Journal of Applied Econometrics*, 28(1), 82–101.

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

——— (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Gneiting, T., and A. E. Raftery (2007): "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102(477), 359–378.

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

Hamilton, J. D. (1994): *Time Series Analysis*. Princeton University Press, 1 edn.

Hersbach, H. (2010): "Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems," *Weather and Forecasting*, 15(5).

Hubrich, K., and R. J. Tetlow (2014): "Financial stress and economic dynamics: the transmission of crises," Working Paper Series 1728, European Central Bank.

Jacquier, E., N. G. Polson, and P. E. Rossi (2004): "Bayesian analysis of stochastic volatility models with fat-tails and correlated errors," *Journal of Econometrics*, 122(1), 185–212.

Jolliffee, I. T., and D. B. Stephenson (2003): *Forecast Verification A Practitioner Guide in Atmospheric Science*. John Wiley and Sons.

Kalliovirta, L., M. Meitz, and P. Saikkonen (2014): "Gaussian Mixture Vector Autoregression," Discussion Paper 386, HECER.

Kim, C. J., and C. R. Nelson (1999): *State-Space Models with Regime Switching*. MIT Press.

Kim, S., N. Shephard, and S. Chib (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *The Review of Economic Studies*, 65(3), 361–393.

Koop, G. (2003): *Bayesian Econometrics*. Wiley.

Ni, S., and D. Sun (2005): "Bayesian Estimates for Vector Autoregressive Models," *Journal of Business and Economic Statistics*, 23(1), pp. 105–117.

Primiceri, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *Review of Economic Studies*, 72(3), 821–852.

Shrestha, D. L. (2014): "Continuous rank probability score (CRPS)," Matlab code, CSIRO Land and Water, Highett.

Sims, C. A., D. F. Waggoner, and T. Zha (2008): "Methods for inference in large multiple-equation Markov-switching models," *Journal of Econometrics*, 146(2), 255 – 274, Honoring the research contributions of Charles R. Nelson.

Sims, C. A., and T. Zha (2006): "Were There Regime Switches in U.S. Monetary Policy?," *American Economic Review*, 96(1), 54–81.

Smith, M. S., and S. P. Vahey (2015): "Asymmetric Forecast Densities for U.S. Macroeconomic Variables from a Gaussian Copula Model of Cross-Sectional and Serial Dependence," mimeo, Melbourne Business School.

8 Technical Appendix

8.1 Monte-Carlo Experiment

In order to assess the Gibbs sampling algorithm for the benchmark model we conduct a simple Monte-Carlo experiment. Data is generated from the following VAR(1) model

$$y_t = B_1 y_{t-1} + u_t$$

$$\begin{aligned} cov(u_t) &= \Sigma = A^{-1} H A^{-1'} \\ e_t &= A u_t \\ e_{it} &= \alpha_{i,S_{it}} + \sigma_{i,S_{it}} \varepsilon_{it}, \varepsilon_{it} \sim N(0,1) \end{aligned}$$

where $i = 1, 2, ..., 4$ and S_{it} denotes the state-variable that follows a three state Markov Chain with transition probability matrix:

$$\begin{pmatrix} 0.95 & 0.025 & 0.025 \\ 0.025 & 0.95 & 0.025 \\ 0.025 & 0.025 & 0.95 \end{pmatrix}$$

The VAR coefficient vector is defined as

$$B_1 = diag([0.5; 0.5; 0.5; 0.5])$$

and

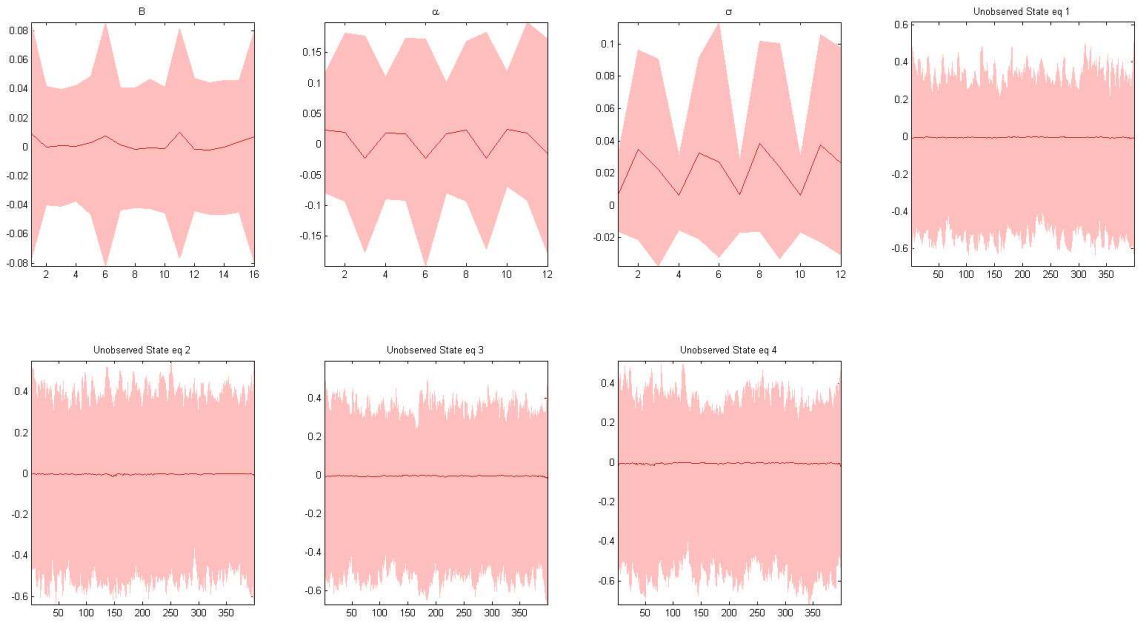
$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -0.1 & 1 & 0 & 0 \\ -0.1 & -0.1 & 1 & 0 \\ -0.1 & -0.1 & -0.1 & 1 \end{pmatrix}$$

The regime specific coefficients are defined as

$$\begin{aligned} \alpha_{i,S_{it}=1} &= -0.5, \quad \alpha_{i,S_{it}=2} = 0, \quad \alpha_{i,S_{it}=3} = 0.5 \\ \sigma_{i,S_{it}=1} &= 0.1^{1/2}, \quad \sigma_{i,S_{it}=1} = 0.2^{1/2}, \quad \sigma_{i,S_{it}=1} = 0.3^{1/2} \end{aligned}$$

We generate 500 observations and discard the first 100 to minimise the influence of starting values. For each artificial dataset, the MCMC algorithm described above is used to estimate the model and the experiment is repeated 500 times. Figure 8 shows the difference between the true values of key parameters and their estimated counterparts. It is clear from the figure that the estimated bias across 500 replications is centered around zero and the algorithm delivers reasonable estimates of the model parameters.

Figure 8: Monte-Carlo experiment. Difference between true and estimated values.



Notes: The red line is the median difference, while the shaded area represents the 10th and the 90th percentile of the difference across 500 iterations.

8.2 Gibbs Algorithm for the VAR Model with Fat-tailed Residuals

The VAR is defined as

$$\begin{aligned} Y_t &= c + \sum_{j=1}^{P} b_j Y_{t-j} + \Sigma_i^{1/2} e_t, e \sim N(0,1) \\ \Sigma_i &= A^{-1} H_i A^{-1'} \end{aligned} \tag{8.1}$$

where A is a lower triangular matrix and H is a diagonal matrix: $H_i = diag(\sigma_1 \varpi_{1,i}, \sigma_2 \varpi_{2,i}, ..\sigma_N \varpi_{N,i}).\varpi_i$ is a vector of unknown parameters.

8.2.1 Priors

Following Koop, the prior for $\lambda_{k,i} = 1/\varpi_{k,i}$ is gamma

$$\begin{aligned} p(\lambda_k) &\sim \Gamma(1, v_{\lambda,k}) \\ p(v_{\lambda,k}) &\sim \Gamma(v_0, 2) \end{aligned}$$

where $v_0 = 20$ and $\Gamma(a,b)$ is the gamma density with mean a and degree of freedom b. The remaining priors are set in an identical fashion to the benchmark model. The Gibbs algorithm samples from the following conditional posterior distributions:

8.2.2 Conditional Posteriors

8.2.2.1 $G(\lambda_k \backslash \Psi)$ As shown in Koop (2003), $G(\lambda_k \backslash \Psi)$ is gamma with mean $\frac{v_{\lambda,k}+1}{\frac{1}{\sigma_K} e_{K,t}^2 + v_{\lambda,k}}$ and degrees of freedom $v_{\lambda,k}+1$ for $K = 1, 2, ...N$. Note that $\mathrm{e}_{K,t}$ is the kth column of $e = Av$ and Ψ denotes all other parameters.

8.2.2.2 $G\left(v_{\lambda,k} \backslash \lambda_k\right)$ As shown in Koop (2003), this conditional is the non-standard and given by

$$G\left(v_{\lambda,k} \backslash \lambda_k\right) \propto \left(\frac{v_{\lambda,k}}{2}\right)^{\frac{T v_{\lambda,k}}{2}} \Gamma\left(\frac{v_{\lambda,k}}{2}\right)^{-N} \exp\left(-\left(\frac{1}{v_0} + 0.5 \sum_{t=1}^{T} \left[\ln\left(\lambda_{t,K}^{-1}\right) + \lambda_{t,K}\right]\right) v_{\lambda,k}\right) \tag{8.2}$$

Koop uses a random walk metropolis step to draw from this conditional. In particular for each K we draw $v_{\lambda,k}^{new} = v_{\lambda,k}^{old} + c^{1/2}\epsilon$ with $\epsilon \sim N(0,1)$. The draw is accepted with probability $\frac{G\left(v_{\lambda,k}^{new} \backslash \lambda_k\right)}{G\left(v_{\lambda,k}^{old} \backslash \lambda_k\right)}$ with c chosen to keep the acceptance rate around 40%.

8.2.2.3 $G\left(\sigma_K \backslash \Psi\right)$ The conditional posterior of σ_K is inverse Gamma. The posterior scale parameter is $D_0 + e_{K,t}^{*\prime} e_{K,t}^{*}$ where $e_{K,t}^{*} = e_{K,t} \cdot \lambda_K^{1/2}$ and degrees of freedom $T + T_0$ where D_0 and T_0 are the prior scale parameter and degrees of freedom, respectively.

8.2.2.4 $G\left(A \backslash \Psi\right)$ Conditional on the VAR coefficients, the system can be re-written as

$$\begin{pmatrix} v_t \\ v_{2t} + v_{1t} a_1 \\ v_{3t} + v_{2t} a_2 + v_{1t} a_3 \\ v_{4t} + v_{3t} a_4 + v_{2t} a_5 + v_{1t} a_6 \end{pmatrix} = \begin{pmatrix} \left(\sigma_1 \varpi_{1,i}\right)^{1/2} e_{1t} \\ \left(\sigma_2 \varpi_{2,i}\right)^{1/2} e_{2t} \\ \left(\sigma_3 \varpi_{3,i}\right)^{1/2} e_{3t} \\ \left(\sigma_4 \varpi_{4,i}\right)^{1/2} e_{4t} \end{pmatrix}$$

conditional on λ_K and σ_K the elements of A have a normal posterior and formulas for linear regressions apply.

8.2.2.5 $G\left(B \backslash \Psi\right)$ Conditional on $\Sigma_i = A^{-1} H_i A^{-1\prime}$ equation 8.1 is a VAR with heteroscedastic disturbances. The distribution of the VAR coefficients is linear and Gaussian. $G\left(B \backslash \Psi\right) \sim N\left(B_{T \backslash T}, P_{T \backslash T}\right)$. We use the Kalman filter to estimate $B_{T \backslash T}$ and $P_{T \backslash T}$ where we account for the fact that the covariance matrix of the VAR residuals changes through time. The final iteration of the filter delivers $B_{T \backslash T}$ and $P_{T \backslash T}$.