An Efficient Bayesian Approach to Multiple Structural Change in Multivariate Time Series*

John M. Maheu† Yong Song‡

December 2015

Abstract

This paper provides a feasible approach to estimation and forecasting of multiple structural breaks for vector autoregressions and other multivariate models. Due to conjugate prior assumptions we obtain a very efficient sampler for the regime allocation variable. A new hierarchical prior is introduced to allow for learning over different structural breaks. Two empirical applications show the improvements the model has over benchmarks.

*Maheu thanks SSHRC of Canada for financial support
†DeGroote School of Business, McMaster University and RCEA. Email:maheujm@mcmaster.ca
‡University of Melbourne and RCEA. Email:yong.song@unimelb.com.au.
1 Introduction

Multivariate time series data plays a central role in macroeconomic analysis and prediction. Linear models such as vector auto regressions (VAR) are standard tools to calculate the impulse response function and forecasts. Recently, many papers highlight the importance of nonlinearity associated with structural instability for macroeconomic and financial variables such as GDP growth, real interest rate, inflation and equity returns among many others. However, because the estimation is usually involved with intensive computation, most of the change-point models are applied to univariate time series. Existing multivariate change-point models have restrictions to the number of regimes a priori. It is either fixed at a small number (2 or 3) as in Jochmann and Koop (2011) or assumed equal to the length of the data as in Cogley and Sargent (2005). A multivariate approach which can estimate and forecast in the presence of an unknown number of regimes is missing in the current literature. This paper develops a new multivariate time series model to fill the gap by exploring the full posterior distribution for the allocation of the data to their respective regimes. The estimation of the new approach is fast by using a conjugate prior for the parameters which characterize each regime. The simulation of the regime allocation of the data from its posterior distribution is very efficient, because the time-varying parameters for the conditional data density are integrated out. A new hierarchical structure is introduced to exploit the information across regimes.

Accounting for structural instability in macroeconomic and financial time series modeling and forecasting is important. Empirical applications by Clark and McCracken (2010), Gior-dani et al. (2007), Liu and Maheu (2008), Wang and Zivot (2000) and Stock and Watson (1996) among others demonstrate strong evidence for the existence of nonlinearity in the form of structural changes.

The challenges of estimation and forecasting in the presence of structural breaks has been recently addressed by Koop and Potter (2007), Maheu and Gordon (2008) and Pesaran et al. (2006) by using Bayesian methods. These approaches provide feasible solutions for univariate time series modeling, but they are computationally intensive. This is because there are too many combinations of the break points, exploring them exhaustively is impractical. For example, Koop and Potter’s (2007) model assumes path dependent time-varying parameters, which imply $O(2^T)$ possible change-points scenarios. Although they have reduced the state space from $O(2^T)$ to $O(T^2)$ in their MCMC algorithm, it is still computationally challenging to calculate the predictive density and the mixing property of their MCMC algorithm is left unanswered. Another approach with an unknown number of regimes is Maheu and Gordon (2008). Since their approach requires conducting $O(T^2)$ posterior inference numerically, the computational burden is even heavier than Koop and Potter’s (2007) method. Extending these methodologies to the multivariate framework is empirically unrealistic, since a multivariate model requires much more computation as the number of variables increases. Instead this paper extends the efficient posterior sampling methods for univariate structural break models in Maheu and Song (2014) to the multivariate setting.

Current multivariate change-point models include Cogley and Sargent (2005), Jochmann and Koop (2011) and Koop et al. (2011). A common feature of these models is that they need to fix the number of regimes a priori. The full posterior distribution for the allocation of the data to their respective regimes is not explored because of this restriction. One potential
solution to this problem is to estimate the model many times. For each time, the estimation is associated with a distinct number of regimes. However, this solution is computationally expensive and in each single estimation a multimodal posterior density may exist, which can cause slow mixing of the Markov chain and affect the inference.

To alleviate the computational burden, we use a conjugate prior for the parameters which characterize each regime. This assumption avoids the numeric approximation of the conditional posterior distribution and provides a closed-form predictive density. This results in a large gain in computational speed. Meanwhile, another advantage of this methodology is that the sampler of the regime allocation is very efficient since the parameters which characterize each regime can be integrated out as nuisance parameters. Different from the usual Gibbs sampling scheme for a hidden Markov model, in which the set of the regime indicators and the set of the parameters characterizing each regime are simulated conditional on each other, this assumption enables us to sample these time-varying parameters jointly. So a potential slow mixing problem in the usual Gibbs sampler associated with a multimodal posterior surface is not present in our MCMC algorithm.

Applying the conjugate priors to VAR was investigated by Kadiyala and Karlsson (1997) for the practitioners. Recent empirical work such as Carriero et al. (2015) has shown the usefulness of simple conjugate priors for the U.S. economy. Banbura et al. (2010) augment the conjugate prior by a shrinkage parameter to reflect subjective belief and show that it is competitive in forecasting. These methods are applied to linear models without structural change. They have demonstrated that a conjugate prior is practically reasonable and a good prior for our structural change models.

Regarding prior elicitation for the parameters which characterize each regime, we adopt two different but closely related approaches. The first is a slightly revised simple conjugate prior used in Carriero et al. (2015), which is designed to approximate the Minnesota prior (Litterman (1986)). This prior is informative but covers a reasonable range of the parameter space. The model using this prior is labelled as non-hierarchical SB model, where SB means structural break. The advantage of this prior is the fast computational speed. For instance, if we assume a VAR(1) model in each regime in a 7-variable system for 600 observations, it takes less than 5 seconds to simulate 6000 samples of model parameters from the posterior distribution in a regular desktop PC.

The second new prior features a hierarchical structure with shrinkage hyper parameters, which is labelled as the hierarchical structural break model. The hierarchical structure is on the parameters which characterize each regime. It is designed to exploit the information across regimes (Pesaran et al. (2006)). In addition, the shrinkage method (e.g., Belmonte et al. (2011)) makes the model parsimonious in the Bayesian framework. The shrinkage hyper parameters in our model can shrink the second prior towards the first one. It reflects the prior belief for the variation of the hierarchical structure.

The hierarchical structure in this paper, to the best of our knowledge, is new to the multivariate time series literature. Current literature of the hierarchical priors such as Pesaran et al. (2006) or Koop and Potter (2007) are on the univariate analysis. In our new approach, besides the ability to learn across regimes, the hierarchical prior is systematically calibrated by following the first prior, which approximates the Minnesota prior. This feature is very

---

1 This is called Rao-Blackwellisation. See Casella and Robert (1996).
important for multivariate models because of the overparameterization problem. In other words, the curse of the dimensionality may make a seemingly harmless hierarchical prior to have strong impact on the inference. Since our hierarchical prior is built on the Minnesota prior, it has a solid theoretic foundation and a reasonable range for the model parameters.

In order to apply the joint sampler for the time-varying parameters, assuming path independence is necessary to reduce the dimension of the state space. Koop and Potter (2007) applies a Gibbs sampler to reduce the state dimension from $O(2^T)$ to $O(T^2)$ in the posterior simulation, but their approach randomly draws the regime allocation and the set of parameters which characterize each regime individually. To sample them jointly, we need to consider $O(2^T)$ scenarios. Each scenario has a distinct path of break points and can represent a state after the time-varying parameters characterizing each regime are integrated out. However, it is impractical to estimate all $O(2^T)$ scenarios with existing regular hardware. This paper applies the assumption similar to Chib (1998) to reduce the dimension of the state space from $O(2^T)$ to $O(T)$. Specifically, we assume that the data before a break point is uninformative for the current regime conditional on the prior for the parameters characterizing each regime. For the non-hierarchical model, this assumption is equivalent to Chib (1998). For the hierarchical approach, the parameters which characterize each regime are dependent, because they share the same hierarchical prior and this prior is not exogenously fixed. However, they are independent conditional on one sample of the hierarchical prior parameters. This assumption frees the model from path dependence and enables an exhaustive exploration of the posterior for the regime allocation. By using this assumption, we have maximal $T$ paths for each observation, which can be evaluated very quickly after being combined with the conjugate prior assumption.

In summary, our approach has four attractive features for practitioners. First, the number of regimes is estimated endogenously and the regime allocation is explored from its posterior distribution exhaustively. All time-varying parameters are sampled jointly, so the estimation is efficient in terms of mixing. Second, the conjugate prior makes the estimation of the non-hierarchical model very fast because no numeric approximation is involved. Third, the hierarchical structure with shrinkage control is parsimonious and able to exploit the information across regimes to improve forecasting. Lastly, the priors are automatically adjusted to different normalization, because they are calibrated according to the Minnesota prior.

The model is applied to oil and real GDP and a 7 variable VAR. In each case the structural break model has superior forecasts and identifies several break points. In general, the hierarchical prior is significantly better in terms of Bayes factors compared to the structural break model without this prior specification.

The paper is organized as follows. The next section introduces the benchmark multivariate linear model, conjugate prior and posterior. Section 3 details the multivariate structural break model followed by Section 4 which extends the model with the hierarchical prior. Section 5 discusses out-of-sample forecasts that account for in-sample breaks as well as breaks out-of-sample. Two empirical applications to oil and real GDP and a 7 variable VAR are given in Section 6 and 7. Conclusion are found in Section 8 followed by the Appendix that collects additional details on posterior simulation.
2 Multivariate Linear Model

We start with a multivariate linear model for a $N \times 1$ vector $y_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})'$ as follows:

$$y_t = \Phi' x_t + e_t, \quad e_t \text{ i.i.d. } \sim \mathcal{N}(0, \Sigma),$$  \hspace{1cm} (1)

where $x_t$ is a $M \times 1$ vector of independent variables and $\Phi$ is a $M \times N$ matrix of coefficients. Each $e_t$ is a $N \times 1$ iid normal random vector with zero mean and a covariance matrix $\Sigma$.

Let $T$ represent the length of the time series data. Define $Y = (y_1, y_2, \ldots, y_T)'$, $X = (x_1, x_2, \ldots, x_T)'$ and $E = (e_1, e_2, \ldots, e_T)'$. Then, we can write (1) as

$$Y = X\Phi + E, \quad E \sim \mathcal{MN}(0, \Sigma, I_T).$$  \hspace{1cm} (2)

The data $Y$ is $T \times N$, $X$ is $T \times M$ and the error term $E$ is $T \times N$. The notation $\mathcal{MN}(0, \Sigma, I)$ means the matrix normal distribution. The first parameter is a $T \times N$ zero matrix representing the mean of the error matrix $E$. The second parameter $\Sigma$ is a $N \times N$ matrix and equals to the covariance of $e_t$. The last parameter, $I_T$, is a $T \times T$ identity matrix and proportional to the covariance matrix of each column of the matrix $E$. The identity matrix $I_T$ comes from the assumption that $e_t$ is i.i.d. If vectorizing the matrix $E$, the matrix normal distribution is equivalent to a multivariate normal distribution as $\text{vec}(E) \sim \mathcal{N}(0, \Sigma \otimes I)$ or $\text{vec}(E') \sim \mathcal{N}(0, I \otimes \Sigma)$.

This paper focuses on structural change in the VAR model. For a VAR($p$) model, where $p$ is the number of lags in the autoregression, it can be represented as

$$y_t = \phi_0 + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + e_t$$

and can be cast into the above model with $x_t = (1, y_{t-1}', y_{t-2}', \ldots, y_{t-p}')'$, $M = Np + 1$ and $\Phi = (\phi_0, \phi_1, \ldots, \phi_p)'$.

2.1 Prior

We assume an Inverse Wishart-Matrix Normal prior distribution for $(\Phi, \Sigma)$ as Carriero et al. (2015):

$$\Sigma \sim \mathcal{IW}(\mathcal{S}, \nu),$$  \hspace{1cm} (3)

$$\Phi|\Sigma \sim \mathcal{MN}(\Phi_0, \Sigma, \Omega).$$  \hspace{1cm} (4)

This prior can be set similar to the Minnesota prior (Litterman (1986)). We use the rules:

1. A stationary series has its regression coefficients centered around 0. A non-stationary series has its regression coefficients to approximate the random walk.

2. The prior for a distant lag is tighter than for a closer lag. In other words, the coefficients of the regressors shrinks to zero as their lag length increases.

$\Sigma$ and $I$ are not identified up to a scalar. This does not affect any derivation or inference in this paper.
3. The volatility and intercept are calibrated by using the univariate series information.

Define the variance of the error term from the best fit ARIMA model of \( y_i \) as \( \hat{\sigma}_i^2 \). The priors are set as follows:

1. We set \( \nu = N + 3.5, S_{ii} = (\nu - N - 1)\hat{\sigma}_i^2 \) and \( S_{ij} = 0 \) if \( i \neq j \). So we have

\[
E(\Sigma) = \begin{pmatrix}
\hat{\sigma}_1^2 & 0 & \cdots & 0 \\
0 & \hat{\sigma}_2^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \hat{\sigma}_N^2 
\end{pmatrix},
\]

with \( sd(\Sigma_{ii}) = 2\hat{\sigma}_i^2 \) and \( sd(\Sigma_{ij}) \approx 1.2\hat{\sigma}_i\hat{\sigma}_j \) if \( i \neq j \).

2. We set

\[
\Phi = \begin{pmatrix}
0 & 0 & \cdots & 0 \\
1(\text{nonstationary}) & 0 & \cdots & 0 \\
0 & 1(\text{nonstationary}) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1(\text{nonstationary}) \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 
\end{pmatrix}
\]

If a series \( y_i \) is nonstationary, we set the corresponding coefficient to 1 and 0 otherwise. Stationarity can be tested by using unit root tests or based on experience.

3. We set

\[
\Omega = \gamma \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{\hat{\sigma}_1^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{\hat{\sigma}_N^2} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{4\hat{\sigma}_1^2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4\hat{\sigma}_N^2} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{P^2\hat{\sigma}_N^2} 
\end{pmatrix},
\]

where scalar \( \gamma \) controls the shrinkage. Carriero et al. (2015) find that setting \( \gamma = 0.2 \) provides a good fit in their macro data application.

A representative variance of \( \Phi_{ij} \) is \( \text{Var}(\Phi_{ij}|\Sigma = E(\Sigma)) = \hat{\sigma}_j^2\Omega_{ii} \). For instance, \( \Phi_{21} \) is the coefficient of \( y_{1,t-1} \) in equation \( y_{1,t} \) and its variance is \( \text{Var}(\Phi_{21}|\Sigma = E(\Sigma)) = \hat{\sigma}_1^2\Omega_{22} = \gamma \).
For another example, if \( i < N \), \( \Phi_{N+1+i,j} \) is the coefficient of \( y_{i,t-2} \) in equation \( y_{j,t} \), we have \( \text{Var}(\Phi_{N+1+i,j}) = \hat{\theta}_j^2 \omega_{N+1+i,N+1+i} = \gamma \frac{\hat{\theta}_j^2}{2} \). The variance decreases as a quadratic function of the lag order.

The value of the top left element is set as 1 to imply a representative variance of the intercept \( \text{Var}(\Phi_{1j}|\Sigma = E(\Sigma)) = \gamma \hat{\theta}_j^2 \), which reflects a proper prior with a reasonable range over the parameter space.\(^3\)

### 2.2 Posterior

The posterior of \( \Phi \) and \( \Sigma \) is still an Inverse Wishart-Matrix Normal distribution by conjugacy:

\[
\Sigma|Y, X \sim IW(\mathbf{S}, \nu) \tag{5}
\]

\[
\Phi|\Sigma, Y, X \sim MN(\bar{\Phi}, \Sigma, \Omega) \tag{6}
\]

where \( \bar{\Phi} = \Omega(\Omega^{-1} \Phi + X'Y) \), \( \Omega = (\Omega^{-1} + X'X)^{-1}, \nu = \nu + T \) and \( \mathbf{S} = \mathbf{S} + Y'Y + \Phi'\Omega^{-1}\Phi - \bar{\Phi}\Omega^{-1}\bar{\Phi} \).

The inverse Wishart matrix normal prior also provides a closed-form solution to the predictive density of \( y_t \), which is a multivariate Student-t distribution. For example, if only the prior is used, we have

\[
y_t|x_t \sim t\left(\Phi'x_t, \frac{(1 + x_t'\Omega x_t)\mathbf{S}}{\nu + 1 - N}, \nu + 1 - N \right) . \tag{7}
\]

where \( E(y_t|x_t) = \Phi'x_t \) and \( \text{Var}(y_t|x_t) = (1 + x_t'\Omega x_t)E(\Sigma), E(\Sigma) = \mathbf{S}/(\nu - N - 1) \), for \( \nu + 1 - N > 2 \). The probability density function is \( p(y_t|x_t) = k^{-1}\left|1 + \frac{(y_t - \Phi'x_t)'\mathbf{S}^{-1}(y_t - \Phi'x_t)}{1 + x_t'\mathbf{S}^{-1}x_t}\right|^{-\nu + 1} \frac{\nu + 1}{2} \).

If we use the posterior distribution, the out-of-sample predictive density of \( y_{T+1} \) is obtained by replacing the prior parameters in Equation 7 by the posterior parameters.

\[
y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T} \sim t\left(\bar{\Phi}'x_{T+1}, \frac{(1 + x_{T+1}'\bar{\Omega} x_{T+1})\mathbf{S}}{\nu + 1 - N}, \nu + 1 - N \right) , \tag{8}
\]

where \( Y_{1,T} = (y_1, \ldots, y_T) \) and \( X_{1,T} = (x_1, \ldots, x_T) \). In a VAR(p) model, \( x_{T+1} \) includes \( y_T, y_{T-1}, \ldots, y_{T-p} \).

### 3 Non-hierarchical structural break model

The difference between a linear model and the structural break model in this paper is that the parameters in the aforementioned linear model are time-varying instead of constant. In other words, we use \( \Phi_t \) and \( \Sigma_t \) to replace \( \Phi \) and \( \Sigma \) to get

\[
y_t = \Phi_t'x_t + e_t, \quad e_t \sim i.i.d. N(0, \Sigma_t) . \tag{9}
\]

\(^3\)It can be changed to a much larger value such as 1.0e10. For a linear model, it is equivalent to Carriero et al. (2015) from the empirical point of view, but their approach needs a training sample because their prior is improper.
Define $\theta_t = (\Phi_t, \Sigma_t)$ as the time-varying parameters which characterize the conditional data density at time $t$. At each time $t$, there is a positive probability $\pi$ for a structural change to occur. If a structural change happens, the new value of $\theta_t$ is drawn from the inverse Wishart matrix normal distribution. Otherwise, $\theta_t$ stays the same as the value in the previous period.

The model is

$$d_t = \begin{cases} d_{t-1} + 1, & \text{w.p. } 1 - \pi \\ 1, & \text{w.p. } \pi \end{cases} \tag{10}$$

$$\theta_t = \begin{cases} \sim F_{\theta}, & \text{if } d_t = 1 \\ \theta_{t-1}, & \text{o.w.} \end{cases} \tag{11}$$

$$y_t|\theta_t, x_t \sim \mathcal{N}(\Phi_t'x_t, \Sigma_t). \tag{12}$$

In (10), $d_t$ is an implicitly defined time-varying parameter, which represents the regime duration up to time $t$. This variable is very important and treated as the state variable for the predictive density. The regime duration $d_t$ takes values of $1, \ldots, t$. The last period $T$ has the maximal number of possible values for $d_t$ (from 1 to $T$). If $d_t = 1$, a structural change happens and $\theta_t$ is drawn from the inverse Wishart matrix normal distribution $F_{\theta}$ as in (11). If no break appears in the previous period, the duration is increased by 1 and $\theta_t$ stays the same as value in the previous period. In each regime, the dynamics of $y_t$ follows a linear representation as in (1) conditional on $\theta_t$.

Compared to existing structural break models, this approach explores all possible change-points as Koop and Potter (2007) and Giordani et al. (2007). The difference is that if there is a structural change ($d_t = 1$), we assume that the new parameter $\theta_t$ is drawn from the distribution $F_{\theta}$ independently from the value of $\theta_{t-1}$. We make this assumption for two reasons. First, it is computationally feasible to calculate the predictive density by integrating out $\theta_t$’s. It reduces the effective number of paths from $O(2^t)$ to $O(t)$ at each period $t$. Second, from an empirical point, it is reasonable or even preferable for some macroeconomic variables to have a sudden change of the parameters.

The parameters to be estimated in this model include the regime durations $D = (d_1, \ldots, d_T)$ and the conditional data density parameters $\Theta = (\theta_1, \ldots, \theta_T)$. Existing MCMC methods usually apply a sampler to randomly draw the regime allocation and the parameters characterizing each regimes conditional on each other. This paper proposes to jointly simulate these time-varying parameters from their posterior distribution. First, randomly sample the regime duration $D$ from its marginal distribution $D|\pi, Y_{1,T}, X_{1,T}$, which is obtainable only if the conjugate prior and the path independence are assumed. Then, conditional on the duration $D$, simulate $\Theta$ from the distribution $\Theta|D, \pi, Y_{1,T}, X_{1,T}$. This is equivalent to the joint sampling from distribution $D, \Theta|\pi, Y_{1,T}, X_{1,T}$, which is efficient based on Casella and Robert (1996).

### 3.1 Prior

We assume that $F_{\theta}$ is the same as the linear model. The break probability has a beta distribution $\pi \sim B(\pi_a, \pi_b)$. 

3.2 MCMC

The MCMC method in this paper is new to the existing literature, so we delineate it in this section. The first step of sampling $D$ from $D|\pi, Y_{1:T}, X_{1:T}$ is done by using the forward filtering and backward sampling method of Chib (1998). In our new approach, the duration $d_t$ is treated as the state variable instead of a regime indicator in the current literature, where a sample series of the regime indicators $S = (s_1, s_2, \ldots, s_T)$ defines the regime allocation of the data and is always in a non-decreasing order. For example, $S = (1, 1, 1, 2, 2, 3, 3, 3, 3)$ means that the first 3 periods are in the first regime, the 4th and 5th periods are in the second regime and the last 4 periods are in the third regime. This sample path is equivalent to a sample path of the regime durations $D = (1, 2, 3, 1, 2, 1, 2, 3, 4)$. For each time $t$ with $d_t = 1$, the data enter into a new regime, otherwise no regime change happens. Clearly, there is a one-to-one relationship between $D$ and $S$.

Each individual value of $s_t$ and $d_t$ has different information content. The regime indicator $s_t$ is able to tell how many regimes there are before time $t$, but is unable to show how long the current regime is. Drawing $s_t$ from its posterior distribution is usually done conditional on the distinct regime dependent parameters $\tilde{\theta}_i$, where subscript $i$ represents the $i$th regime. By definition, we have $\theta_t = \tilde{\theta}_{s_t}$. On the other hand, $d_t$ tells the current regime’s duration but contains no information about how many regimes appear before time $t$. So if one only knows $d_t$ and all the distinct values of $\tilde{\theta}_i$’s, he cannot tell the current value of $\theta_t$. However, if the data in the past regime is uninformative to the current regime, as we assume, then the regime duration $d_t$ is sufficient to obtain the posterior and predictive density by integrating out the parameters of the conditional data density in that regime. This cannot be done by using the regime indicator $s_t$.

In our approach, the assumption of independent sampling of new $\theta_t$ from $F_\theta$ enables us to treat $d_t$ as a state variable, because it is sufficient to produce the predictive density. $\Theta$ is integrated out as a set of nuisance parameters and the MCMC posterior sampler simulates directly from the marginal posterior distribution of the regime durations $D|\pi, Y_{1:T}, X_{1:T}$. The conjugate prior provides a closed form for the predictive density and accelerates the computational speed considerably, making the MCMC algorithm practical for multivariate applications.

Exact block sampling from $D|\pi, Y_{1:T}, X_{1:T}$ follows from the forward filtering and backward sampling steps. The forward filter:

1. At $t = 1$, set $p(d_1 = 1|\pi) = 1$.

2. The forecasting step:

$$p(d_t = j|\pi, Y_{1:t-1}, X_{1:t-1}) = \begin{cases} p(d_{t-1} = j - 1|\pi, Y_{1:t-1}, X_{1:t-1})(1 - \pi), & \text{for } j = 2, \ldots, t; \\ \pi, & \text{for } j = 1. \end{cases}$$

When $t = 1$, $Y_{1:t-1}$ and $X_{1:t-1}$ are empty sets.

3. The updating step:

$$p(d_t = j|\pi, Y_{1:t}, X_{1:t}) = \frac{p(y_t|d_t = j, x_t, Y_{1:t-1}, X_{1:t-1})p(d_t = j|\pi, Y_{1:t-1}, X_{1:t-1})}{p(y_t|\pi, x_t, Y_{1:t-1}, X_{1:t-1})}.$$
for $j = 1, \ldots, t$. The first term in the numerator is a multivariate Student-t distribution density function since

$$y_t|d_t = j, x_t, Y_{1,t-1}, X_{1,t-1} \sim t \left( \Phi' x_t, (1 + x_t' \hat{\Omega} x_t) \hat{S}, \hat{\nu} + 1 - N \right)$$

(13)

with $\Phi = \hat{\Omega} (\Omega^{-1} \Phi + X'_{t+1-j,t-1} Y_{t+1-j,t-1}), \hat{\Omega} = (\Omega^{-1} + X'_{t+1-j,t-1} X_{t+1-j,t-1})^{-1}, \hat{\nu} = \nu + j - 1$, and $\hat{S} = S + Y'_{t+1-j,t-1} Y_{t+1-j,t-1} + \Phi' \hat{\Omega}^{-1} \Phi - \Phi' \hat{\Omega}^{-1} \Phi$. If $d_t = 1$, which means a structural change, then $X_{t+1-j,t-1}$ and $Y_{t+1-j,t-1}$ are empty sets and all hat parameters $(\hat{\Phi}, \hat{\Omega}, \hat{\nu}, \hat{S})$ are replace by the prior parameters $(\Phi, \Omega, \nu, S)$.

The second term of the numerator is obtained from step 2.

The predictive likelihood in the denominator, $p(y_t|\pi, x_t, Y_{1,t-1}, X_{1,t-1})$, is computed by summing over all values of the duration $d_t$

$$p(y_t|\pi, x_t, Y_{1,t-1}, X_{1,t-1}) = \sum_{d_t=1}^{t} p(y_t|d_t, x_t, Y_{1,t-1}, X_{1,t-1}) p(d_t|\pi, Y_{1,t-1}, X_{1,t-1}).$$

(14)

4. Iterate over step 2 and 3 until the last period $T$.

The backward sampler of the duration vector $D$ is the following:

1. Sample the last period duration $d_T$ from the distribution $d_T|\pi, Y_{1,T}, X_{1,T}$, which is obtained from the last iteration of the forward-filtering step.

2. If $d_t > 1$, then $d_{t-1} = d_t - 1$.

3. If $d_t = 1$, then sample $d_{t-1}$ from the distribution $d_{t-1}|\pi, Y_{1,t-1}, X_{1,t-1}$. This is because $d_t = 1$ implies a structural change at time $t$. Hence, for any $\tau \geq t$, the data $y_\tau$ is in a new regime and independent of $d_{t-1}$. The distribution $d_{t-1}|d_t = 1, \pi, Y_{1,T}, X_{1,T}$ is equivalent to $d_{t-1}|d_t = 1, \pi, Y_{1,t-1}, X_{1,t-1}$.

4. Iterate step 2 and 3 until the first period $t = 1$.

After obtaining the durations $D$, simulating $\Theta$ from $\Theta|D, Y_{1,T}, X_{1,T}$ is simply done by using the conjugacy property of (5) and (6). First convert $D$ to a series of regime indicators $S = (s_1, \ldots, s_T)$. This is done by calculating the number of regimes $K$ and index the regimes by $1, \ldots, K$. Label $s_1 = 1$ and $s_t = 1$ for $t > 1$ until at some time $\tau$ with $d_\tau = 1$, which implies there is a break and the data is in a new regime. Then, set $s_\tau = 2$ at this break point. Iterate this labelling procedure until the last period with $s_T = K$.

We know that a sample series of $D$ and $S$ are equivalent. The reason of introducing $S$ is to help the sampling of $\Theta$ looks more straightforward. Because $\Theta$ can only takes $K$ possible values implied by a sample path of $S$ ($K$ can be different for other sample paths of $S$), we can define its distinct values as $\Theta^* = (\theta^*_1, \ldots, \theta^*_K)$. Because each $\theta^*_i$ is independent from the other $\theta^*_j$’s, we can simulate each $\theta^*_i$ only conditional on the data allocated to the $i$th regime implied by $S$. In detail, $\theta^*_i$ is randomly drawn from the following distribution.

$$\Sigma^*_i \sim IW(\mathcal{S}_i, \nu_i)$$

(15)
\[ \Phi_i^* \Sigma_i^* \sim MN(\Phi_i, \Sigma_i^* \otimes \Omega_i) \]  

with \( \Phi_i = \Omega_i (\Omega^{-1} \Phi + X_i^* Y_i^*)^{-1}, \Omega_i = (\Omega^{-1} + X_i^* X_i^*)^{-1}, \nu_i = \nu + d_i^*, \) and \( \Sigma_i = S + Y_i^* Y_i^* + \Phi_i^* \Omega_i^{-1} \Phi_i - \Phi_i^* \Omega_i^{-1} \Phi_i. \) The data \( X_i^* = (x_{t_0}, \ldots, x_{t_1})' \) and \( Y_i^* = (y_{t_0}, \ldots, y_{t_1})' \), where \( s_t = i \) if and only if \( t_0 \leq t \leq t_1 \), are the collection of \( x_t \) and \( y_t \) being allocated to the \( i \)th regime, respectively. \( d_i^* \) is the duration of the \( i \)th regime.

The above algorithm is based on a fixed break probability \( \pi \). If we have a prior for \( \pi \) as a beta distribution \( B(\pi_a, \pi_b) \), the conditional posterior of \( \pi \) is \( \pi | D \sim B(\pi_a + K - 1, \pi_b + T - K) \) by conjugacy. This can be combined with the previous steps to form a Gibbs sampler as follows:

1. Sample \( D, \Theta | \pi, Y_{1,T}, X_{1,T} \).
2. Sample \( \pi | D \).

### 4 Hierarchical structural break model

The advantage of the non-hierarchical structural break model is that the estimation time is almost negligible. We can estimate a model with a thousand observations in a few minutes.

Meanwhile, the fast computational speed allows for a more complex structure that learns about structural breaks. In this paper, we propose a hierarchical structure to govern the data density parameters and exploit information across regimes. It is also a natural solution to the prior sensitivity check and may be more useful than the Minnesota prior in out-of-sample forecasting.

In the non-hierarchical model (10)-(12), the distinct parameters \( \theta_i^* \)'s are drawn from the pre-specified distribution \( F_\theta \). In this section, We propose to use these values to learn \( F_\theta \) instead of assuming it as exogenous. This can be translated to proposing a prior for \( (\Phi, \Omega, S, \nu) \), which are the parameters of the distribution \( F_\theta \).

These priors are assumed as follows:

\[ \Omega \sim IW(\Omega_0, \omega_0), \]  

\[ \Phi|\Omega \sim MN(M_0, \Lambda_0, \Omega), \]  

\[ S \sim W(S_0, \tau_0), \]  

\[ \nu \sim G(a_0, b_0)1(\nu \geq N + 1). \]

The detailed MCMC procedure to draw the model parameters from the posterior distribution is in the appendix. A simple list of steps is as follows:

1. Sample \( D, \Theta | \pi, \Phi, \Omega, S, \nu, Y_{1,T}, X_{1,T} \) by using the joint sampler in the non-hierarchical model.
2. Sample \( \pi | D \).
3. Sample \( \Phi, \Omega | D, \Theta \)
4. Sample \( S | D, \Theta, \nu \).
5. Sample $\nu|D, \Theta, S$.

The path independence and conjugacy assumptions greatly facilitate the computation of Step 1, so the MCMC algorithm can iterate for thousands of times to obtain the numeric approximation for the posterior of the hierarchical parameters $(\Phi, \Omega, S, \nu)$.

4.1 Prior

The prior for the hierarchical model is related to that of the non-hierarchical model in the sense that the hierarchical prior is set to be centered around the non-hierarchical prior and can be controlled to shrink towards it. One advantage of this hierarchical structure is that it allows estimation of the hyper parameters instead of fixing them exogenously. Hence, we can learn from the information across regimes. The second attractive feature is that shrinkage can make the model parsimonious which is especially useful in the multivariate framework.

In detail, we use the following.

- Set $\omega_0 = M + 3.5$ and $\Omega_0 = (\omega_0 - M - 1)\Omega_{\text{non-hie}}$, so we have $E(\Omega) = \Omega_{\text{non-hie}}$ and $sd(\Omega_{ii}) = 2\Omega_{\text{non-hie},ii}$. Notice that $\omega_0$ can be adjusted (increased) to reflect shrinkage. We need $\omega_0 > M + 1$.

- Set $M_0 = \Phi_{\text{non-hie}}$ and $\Lambda_0 = \lambda \begin{pmatrix} \hat{\nu}_1^2 & 0 & \ldots & 0 \\ 0 & \hat{\nu}_2^2 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \hat{\nu}_N^2 \end{pmatrix}$, where $\lambda$ controls the shrinkage. This implies that $Var(\Phi_{ij}|\Omega) = \lambda \hat{\nu}_j^2 \Omega_{ii}$.

- Set $S_0 = \frac{1}{\tau_0} S_{\text{non-hie}}$ and $\tau_0 = N$, so we have $E(S) = S_{\text{non-hie}}$ and $sd(S_{ii}) = \left(\frac{2}{\tau_0}\right)^{1/2} S_{\text{non-hie},ii}$. The $\tau_0$ can be increased to reflect shrinkage. We need $\tau_0 > N - 1$.

- For $\nu > N + 1$ set $a_0 = 10$ and $b_0 = \frac{a_0}{\nu_{\text{non-hie}}}$. If no restriction, the prior mean of $\nu$ is $\nu_{\text{non-hie}}$. As $a_0 \to \infty$, the $\nu$ shrinks to $\nu_{\text{non-hie}}$.

5 Forecasts

Forecasts of the model naturally take into account past structural changes as well as structural changes occurring out-of-sample. This is done by integrating over all possible change points. Let $\Psi = \{D, \Theta, \pi, \Phi, \Omega, S, \nu\}$ denote the posterior draws from the hierarchical structural break model. Then the predictive density one-period ahead conditional on $\Psi$ is

$$p(y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}, \Psi) = \sum_{d_{T+1}=1}^{T+1} p(y_{T+1}|d_{T+1}, x_{1,T+1}, Y_{1,T}, Y_{1,T}, \Psi)p(d_{T+1}|Y_{1,T}, X_{1,T}, \Psi). (21)$$

This is computed as in (14). The value of $d_{T+1} = 1$ is a structural break out-of-sample. The final estimate is obtained after all parameter uncertainty is integrated out. For example,
given \( R \) MCMC draws obtained after dropping a suitable number of initial burn-in draws, the predictive density estimate is

\[
p(y_{T+1}|x_{T+1}, Y_{1:T}, X_{1:T}) \approx \frac{1}{R} \sum_{i=1}^{R} p(y_{T+1}|x_{T+1}, Y_{1:T}, X_{1:T}, \Psi^{(i)}).
\] (22)

The log-marginal likelihood can be computed from this by estimating the predictive density at each time \( t \) and evaluating it as the associated data \( y_t \). For instance, the log-marginal likelihood for hierarchical structural break model given the data \( Y_{1:T} \) is

\[
LML = \sum_{t=1}^{T} \log(p(y_t|x_t, Y_{1:t-1}, X_{1:t-1})).
\] (23)

In a similar way the log-predictive likelihood can be computed over the data \( y_{\tau_s}, \ldots, y_{\tau_e} \) where \( \tau_s \leq \tau_e \). In this case the summation \( t = 1, \ldots, T \) in (23) is replaced with \( t = \tau_s, \ldots, \tau_e \). The marginal likelihood or predictive likelihood are the key ingredients in Bayesian model comparison. Similar results to these hold for the non-hierarchical model.

Finally, the predictive mean can be computed based on (21) which integrates over all possible break points,

\[
E[y_{T+1}|x_{T+1}, Y_{1:T}, X_{1:T}, \Psi] = \sum_{d_{T+1}=1}^{T+1} E[y_{T+1}|d_{T+1}, x_{T+1}, Y_{1:T}, X_{1:T}, \Psi] p(d_{T+1}|Y_{1:T}, X_{1:T}, \Psi).
\] (24)

Each of the terms \( E[y_{T+1}|d_{T+1}, x_{T+1}, Y_{1:T}, X_{1:T}, \Psi] \) are obtained from the conditional mean in (13) given \( \Psi \) and weighted by the probability of duration \( d_{T+1} \). The final estimate with parameter uncertainty accounted for is

\[
E[y_{T+1}|x_{T+1}, Y_{1:T}, X_{1:T}] \approx \frac{1}{R} \sum_{i=1}^{R} E[y_{T+1}|x_{T+1}, Y_{1:T}, X_{1:T}, \Psi^{(i)}].
\] (25)

### 6 Oil and Real GDP

Oil price dynamics are nonlinear and feature heteroskedasticity (Baumeister and Peersman 2013) and model and parameter instability (Baumeister and Kilian 2015, Kilian 2009). There is a large body of literature that investigates the changing relationship between oil and the economy (Guo and Kliesen 2005, Hamilton 2009; 2011, Hooker 1996). This complex relationship is of considerable interest to both academics and policy makers.

We use our model to study the interrelationship between the oil prices and output and to compare out-of-sample forecasts with benchmark models. We obtain the oil price from Citibase as the composite refiner’s acquisition cost (data label: EEPRPC). The real GDP growth rate (chained, seasonally adjusted) is downloaded from Bureau of Economic Analysis. We transform the oil price level to growth by taking the difference of the log prices and multiply it by 100. The data span from 1974Q2-2015Q2 (165 observations). We estimate a VAR, the non-hierarchical structural change model and the hierarchical structural change model.
In order to account for potential heteroskedasticity, we also estimated a vector diagonal GARCH model in the VAR setting, denoted by VAR-VDGARCH. Specifically,

\[ y_t = \Phi' x_t + e_t, \quad e_t \sim N(0, H_t) \quad (26) \]
\[ H_t = CC' + AA' \odot e_{t-1}' e_{t-1} + BB' \odot H_{t-1}, \quad (27) \]

where \(\odot\) is element-by-element multiplication (Hadamard product). This model essentially implies a univariate GARCH process for each element of the conditional covariance. We assume that \(AA'\) and \(BB'\) are scalars \(a\) and \(b\) for simplicity. \(C\) is a lower triangular matrix, which is estimated.

The prior of \(\Phi\) is the same as that of the linear model when \(\Sigma\) is fixed at its mean. The prior of \(C\), \(a\) and \(b\) are set to let the stationary mean of \(H_t (CC'/(1 - a - b))\) equals to the prior mean of the covariance matrix of the linear model. The prior mode of \(a\) and \(b\) are set as 0.2 and 0.7, respectively. We restrict \(a > 0\), \(b > 0\) and \(a + b < 1\) and assign an approximately uniform prior to them. The details can be found in the appendix. The Particle Markov Chain Monte Carlo (PMCMC) method of Herbst and Schorfheide (2014) is applied to compute the predictive likelihood at each period.

Table 1 shows the log-marginal likelihoods for various lag lengths. There are improvements in adding GARCH dynamics and moving to the structural break specifications. The improvements are significant. The best model is the hierarchical structural change specification with \(p = 2\). The log-Bayes factor for this model versus the non-hierarchical version is 6.0 while it is 18 against the VAR-VDGARCH. The structural change models dominate the VAR and VAR-VDGARCH models strongly for any lag length.

Because computation is fast, we can produce the cumulative log-predictive likelihoods and plot them in Figure 1. Each model uses the number of lags that has the largest marginal likelihood. The benchmark model is VAR(3). The red line is the difference between the VAR(3)-VDGARCH and VAR(3). The green line represents the difference between the non-hierarchical structural change model with \(p = 2\) and VAR(3) and the blue line represents the difference between the hierarchical structural change model with \(p = 2\) and VAR(3). The figure shows that the VAR-VDGARCH is gaining some prediction power over time because it incorporates heteroskedasticity. The structural change models are still the best in general over time. The data support structural changes, for instance, at the financial crisis in 2009. We can observe a jump-up of the green and blue lines at that period, evidence of the data supporting structural change models against the linear model.

Table 2 shows the root mean square forecast errors for the predictive mean. The last 10 years (120 observations) are used for out-of-sample forecasts. For the oil price change, the VAR and hierarchical structural change model has the best fit. Although the non-hierarchical model has better density forecasts than the VAR model, allowing for structural change does not improve point forecasts. Meanwhile, if we allow the hierarchical structure to learn from the data, the point forecast can be improved. It is also true for the GDP growth point forecast. The best forecasting models are the hierarchical structural change model and the VAR. And the hierarchical structural change model is robust to the number of lags included.

Figure 2 plots the structural change probability. It shows a strong pattern of parameter uncertainties. There are a significant number of breaks that are identified with over 0.8 probability.
Figure 3 shows the data and in-sample posterior predictive means implied by the best hierarchical structural change model by using the full sample. The top panel is the oil price change and the bottom is the real GDP growth. The in-sample fit is very good from the model with the best out-of-sample density forecast.

Figure 4 shows the standard deviations and the correlation coefficients of the error term over time. The top panel and middle panel plot the standard deviations of the error terms of the oil price change and the real GDP growth, respectively. We can see a clear spike in 2008Q4, which is coincide with the financial crisis. The correlation changes over time, but the deviations, both positive and negative from zero, are short lived. There is a large swing in 2008Q4 to 0.8 for the correlation coefficient, but it quickly reverts back to around zero.

7 A VAR for the U.S. Economy

In this application, we apply the structural break models to a medium size system with 7 variables downloaded from CITIBASE. They are: unemployment rate (UR), Core PCE (1200 × log difference of the level), nonfarm employment (1200 × log difference of the level), retail sales (1200 × log difference of the level), housing starts level (100 × log difference of the level), industrial production index (1200 × log difference of the level), and the federal funds rate. There are 625 observations from 1959M02 to 2011M02. Summary statistics are shown in Table 3. We can notice that the variables are normalized differently from the variance column. This is not a problem since scaling is automatically corrected in the prior elicitation procedure. The models used are a VAR and the hierarchical SB-VAR model with several lag lengths and the non-hierarchical SB-VAR(1).

Three features are discovered in this application. First, we find structural instability is an important feature for the U.S. macroeconomic variables, which is consistent with the literature. Second, volatility has a decreasing pattern in general and is in line with the great moderation. Meanwhile, some volatility jumps exist. Lastly, our approach finds the number of regimes is different from most of the current literature. Existing models either assume a small number of regimes (2 or 3) or structural change at each time (T). We find the best model supports more than 5 regimes.

Table 4 shows the predictive likelihoods and the root mean square forecast errors for the last 10 years of the sample. The second column is the log-predictive likelihoods. According to this the best VAR has a lag length of 4 and log-predictive likelihood of -1751.8. All of the non-hierarchical structural break models significantly improve upon the VAR with log-predictive Bayes factors that range from 30.7 to 56.2.

The hierarchical structural break specification has the best density forecasts. The log-predictive Bayes factor is 52.3 against the best non-hierarchical version and 108.5 against the best VAR specification. Learning through the hierarchical structure is clearly important to improving density forecasts.

The remaining columns of the table report the root mean square forecast error of the predictive mean forecasts. The best mean forecast is represented by the bold font. There is no clear evidence of which model performs the best in point forecasting. The hierarchical

---

4 This is the same set of variables used in Carriero et al. (2015).
model tends to have lower forecast errors than the non-hierarchical version but often the best point forecasts are from a VAR.

Figure 5 plots the posterior probabilities of structural changes implied by the non-hierarchical model. It shows a visible structural change at 1987M03 and some evidence of structural instability in the end of 1987 and early 1988.

To understand the structural change in the multivariate system Figure 6 shows the posterior mean of the volatility of each individual variable \( \sigma_{it}^{(i)} = \sqrt{\Sigma_{ii}^{(ii)}} \), for \( i = 1, \ldots, 7 \) and \( t = 1, \ldots, T \). All variables feature a volatility decrease after the structural change, which is consistent with the great moderation. However, the timing is different from the current literature, which is considered to start in early 1980’s as in Kim and Nelson (1999). The non-hierarchical model identifies one important break in the data.

We report the results implied by the preferred hierarchical model which are very different than the non-hierarchical model. Figure 7 shows the smoothed break probability over time. The hierarchical model finds more than the two regimes identified by the non-hierarchical model. Defining a break if the posterior break probability \( p(d_t = 1|I_T) > 0.5 \), the model identifies 1960M06, 1979M10, 1982M12 and 2009M01 as the change-points. If using \( p(d_t = 1|I_T) > 0.2 \) as the criteria of the structural change, 1979M09, 1984M03, 1987M12, 1995M05, 2001M01, 2001M11, 2007M12 and 2009M11 can also be considered as change-points. This finding is consistent with Koop and Potter (2007) in their univariate analysis of U.S. GDP growth and inflation data.

Figure 8 plots the posterior mean of \( \phi_{i,t}^{(ii)} \) over time. It represents the average effect of the first lag of the variable on itself. The unemployment rate (UR) and the federal fund rate (FFR) is very persistent for most of the time, while the rest of the variables are mean reverting.

Figure 9 plots the posterior mean of the volatility \( \sigma_{i,t}^{(i)} \) of the innovations over time. All variables except Core PCE have a trend of decreasing volatility over time. We have identified late 1979 as structural change point. Meanwhile, these dynamics are not exactly the same as implied by the great moderation, because heterogeneous dynamics exist for these macroeconomic variables. For example, some variables such as the unemployment rate and the federal fund rate had a volatility increase instead of a decrease after the 1979 break. Volatility of retail sales decreased after 1979, but after 1984M03 volatility jumped up. Industrial production had a volatility decrease after early 80’s, however, a volatility increase appeared during the most recent financial crisis.

8 Conclusion

This paper develops a new multivariate time series model that allows for multiple structural breaks in-sample and incorporates structural breaks into out-of-sample forecasts. The estimation is fast by using a conjugate prior for the parameters which characterize each regime. The simulation of the regime allocation of the data from its posterior distribution is very efficient, because the time-varying parameters for the conditional data density are integrated out. A new hierarchical structure is introduced to exploit the information across regimes. Two applications show the usefulness of our model to multivariate time series.
9 Appendix

9.1 Inverse Wishart - Matrix Normal prior

1. $\Sigma$:

   The error covariance matrix $\Sigma$ has a Inverse-Wishart distribution. Its prior mean is
   \[ E(\Sigma) = \frac{S}{\nu - N - 1} \]
   The variance of each element
   \[ Var(\Sigma_{ij}) = \frac{(\nu - N + 1)S_{ij}^2 + (\nu - N - 1)S_{ii}S_{jj}}{(\nu - N)(\nu - N - 1)^2(\nu - N - 3)} \]
   Its density function is given by
   \[ p(\Sigma) = \frac{|S|^{\nu/2}|\Sigma|^{-(\nu+N+1)/2}etr\{-\frac{1}{2}S\Sigma^{-1}\}}{2^{\nu N/2}\Gamma_N(\nu/2)} \]

   $\Gamma_p$ is multivariate gamma function, which is $\Gamma_p(a) = \int_{S > 0}etr\{-S\}|S|^{a-1/2}dS$ where $S > 0$ means $S$ is $p \times p$ positive definite matrix, or $\Gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^p \Gamma(a+(1-j)/2)$

   A special case is when $N = 1$. Then $\Sigma = \sigma^2$ as a scalar and
   \[ p(\sigma^2) = \frac{s^{\nu/2}(\sigma^2)^{\nu/2-1}exp\{-\frac{s}{2}\sigma^{-2}\} \exp\{-\frac{s^2}{2}\sigma^{-2}\}}{2^{\nu/2}\Gamma(\nu/2)} \]

   So $\sigma^2$ has an inverse-gamma distribution with a shape parameter $\nu/2$ and a scale parameter $\frac{s}{2}$. The mean and the variance of the $\sigma^2$ equal to $\frac{s}{\nu - 2}$ and $\frac{2s^2}{(\nu - 2)^2(\nu - 4)}$, respectively.

   The precision matrix $P$, which is the inverse of the covariance matrix $\Sigma$, has a Wishart distribution $W(P, \nu)$, where $P = S^{-1}$. It has density
   \[ p(P) = \frac{|P|^{\nu/2}P^{(\nu-N-1)/2}etr\{-\frac{1}{2}P^{-1}P\}}{2^{\nu N/2}\Gamma_N(\nu/2)} \]

   A special case is when $N = 1$, then $P = \sigma^{-2}$ has a gamma distribution with
   \[ p(\sigma^{-2}) = \frac{s^{\nu/2}(\sigma^{-2})^{\nu/2-1}\exp\{-\frac{s}{2}\sigma^{-2}\}}{2^{\nu/2}\Gamma(\nu/2)} \]

   The mean and variance of $\sigma^{-2}$ are $\frac{\nu}{2}$ and $\frac{2\nu}{\nu - 2}$.

2. $\Phi$:

   The regression coefficient matrix $\Phi$ has a matrix normal distribution. Each column of $\Phi$, $\Phi_j$, is the regression coefficients for the $j$th equation and has a multivariate normal distribution
   \[ \Phi_j|\Sigma \sim N(\Phi_j, \Sigma_{jj}\Omega) \]
Each row of $\Phi_i, \Phi_{ii}$, is the coefficients of impact from the same source across equations.

$$\Phi_{ii} | \Sigma \sim N(\Phi_{ii}, \Sigma \Omega_{ii})$$

The density function is

$$p(\Phi | \Sigma) = \frac{\text{etr}\{-\frac{1}{2} \Sigma^{-1} (\Phi - \Phi') \Omega^{-1} (\Phi - \Phi)\}}{(2\pi)^{MN/2} |\Sigma|^{M/2} |\Omega|^{N/2}}$$

### 9.2 Sample from a matrix Gaussian

For $\Phi | \Sigma \sim MN(\Phi, \Sigma \otimes \Omega)$, to generate a sample of $\Phi$, first get lower triangular matrices $\Sigma^{1/2}$ and $\Omega^{1/2}$ through Cholesky decomposition. Then, generate $C \sim WN(0, I \otimes I)$. $\Phi$ is generated from

$$\Phi = \Omega^{1/2} C \Sigma^{1/2'},$$

since $vec(\Omega^{1/2} C \Sigma^{1/2'}) = \Sigma^{1/2} \otimes \Omega^{1/2} vec(C)$. So the variance of $vec(C)$ is $\Sigma^{1/2} \otimes \Omega^{1/2} (\Sigma^{1/2} \otimes \Omega^{1/2})' = \Sigma^{1/2} \otimes \Omega^{1/2} (\Sigma^{1/2'} \otimes \Omega^{1/2'}) = (\Sigma^{1/2} \Sigma^{1/2'}) \otimes (\Omega^{1/2} \Omega^{1/2'}) = \Sigma \otimes \Omega$

### 9.3 Sample from an Inverse-Wishart distribution

Generate $\Sigma$ from a Inverse-Wishart, $IW(S, \nu)$, by

$$\Sigma = S^{1/2} C^{-1} S^{1/2'}$$

where $S^{1/2}$ is the lower triangular matrix from the Cholesky decomposition of $S$ and $C$ is drawn from a Wishart $W(I, \nu)$.

### 9.4 Sample the hierarchical prior

1. $\Phi$ and $\Omega$:

   The prior is matrix normal and inverse-Wishart.

   $$\Omega \sim IW(\Omega_0, \omega_0)$$

   $$\Phi | \Omega \sim MN(M_0, \Lambda_0 \otimes \Omega)$$

   The conditional posterior $\Phi, \Omega | \{\Sigma_i, \Phi_i\}^K_{i=1}$ is

   $$\Omega | \{\Sigma_i, \Phi_i\}^K_{i=1} \sim IW(\Omega_1, \omega_1)$$

   $$\Phi | \Omega, \{\Sigma_i, \Phi_i\}^K_{i=1} \sim MN(M_1, \Lambda_1 \otimes \Omega)$$

   with

   $$\Omega_1 = \Omega_0 + \sum_{i=1}^K \Phi_i \Sigma_i^{-1} \Phi_i' + M_0 \Lambda_0^{-1} M_0' - M_1 \Lambda_1^{-1} M_1'$$
\[ \omega_1 = \omega_0 + KN \]
\[ M_1 = (M_0 \Lambda_0^{-1} + \sum_{i=1}^{K} \Phi_i \Sigma_i^{-1}) \Lambda_1 \]
\[ \Lambda_1 = (\Lambda_0^{-1} + \sum_{i=1}^{K} \Sigma_i^{-1})^{-1} \]

2. \( S \):

The prior of \( S \) is a Wishart \( W(S_0, \tau_0) \). The conditional posterior is also Wishart.
\[ S|\nu, \{\Sigma_i\}_{i=1}^{K} \sim W(S_1, \tau_1) \]

with
\[ S_1^{-1} = S_0^{-1} + \sum_{i=1}^{K} \Sigma_i^{-1} \]
\[ \tau_1 = \tau_0 + K\nu \]

3. \( \nu \):

The prior is a Gamma \( G(a_0, b_0) \). The conditional posterior has no convenient form.
\[
p(\nu|S, \{\Sigma_i\}_{i=1}^{K}) = p_G(\nu; a_0, b_0) \prod_{i=1}^{K} p(\Sigma_i|S, \nu) \\
\propto p_G(\nu; a_0, b_0) \prod_{i=1}^{K} \left\{ \frac{|S|^{\nu/2}}{2^{\nu N/2} \Gamma_N(\nu/2)} |\Sigma_i|^{-\nu_i + N + 1/2} \right\} \\
\propto \nu^{a_0-1} e^{-\frac{b_0}{\nu}} \frac{|S|^{K\nu/2}}{2^{K\nu N/2} \Gamma_N^{K}(\nu/2)} \prod_{i=1}^{K} |\Sigma_i|^{-\nu_i + N + 1/2} \]

The log of the last equation (after discarding more constants) is
\[
\frac{K \log(|S|) - 2b_0 - KN \log(2) - \sum_{i=1}^{K} \log(|\Sigma_i|)}{2} - K \log(\Gamma_{N}(\nu/2)) + (a_0 - 1) \log(\nu) 
\]

The sampling method of \( \nu \) is a M-H step with a proposal distribution of
\[ \nu^{(i)} \sim G(\xi, \xi/\nu^{(i-1)}) \]
9.5 VAR-VDGARCH Prior

1. The prior of the regression coefficients $\Phi$ is set the same as the linear model.

2. The prior of $a$ and $b$ is given by the following transformation

$$a = \frac{e^{\theta_1}}{1 + e^{\theta_1} + e^{\theta_2}}, \quad b = \frac{e^{\theta_2}}{1 + e^{\theta_1} + e^{\theta_2}}$$

We set the mean of $\theta_1$ and $\theta_2$ to satisfy $a = 0.2$ and $b = 0.7$. The prior of $\theta_1$ and $\theta_2$ are assumed to have normal distributions with the aforementioned mean and variance of 10.

3. The prior of $C$ is set as normal distribution with the following parametrization:

   (a) Let the stationary covariance matrix $H = \frac{CC'}{1-a-b}$ equals to the prior mean of the covariance matrix in the VAR model, where $a$ and $b$ are set as their prior means. Solve $C$ by using Cholesky decomposition.

   (b) Take log of the diagonal elements of $C$ and vectorize the lower triangular of $C$ (notice that $C$ is a lower triangular matrix). This vector is set as the prior mean.

   (c) The covariance matrix is set as 10 times an identity matrix.

References


Table 1: Oil and Real GDP: Log Marginal Likelihood

<table>
<thead>
<tr>
<th></th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR</td>
<td>-844.2</td>
<td>-841.5</td>
<td>-841.1</td>
<td>-841.4</td>
</tr>
<tr>
<td>VAR-VDGARCH</td>
<td>-832.3</td>
<td>-826.0</td>
<td>-824.7</td>
<td>-825.9</td>
</tr>
<tr>
<td>Non-Hie SB</td>
<td>-817.0</td>
<td>-814.0</td>
<td>-815.7</td>
<td>-816.8</td>
</tr>
<tr>
<td>Hie SB</td>
<td>-811.2</td>
<td><strong>-808.0</strong></td>
<td>-808.2</td>
<td>-808.1</td>
</tr>
</tbody>
</table>

$p$ is the lag length and the bold entry is the largest value.

Table 2: Oil and Real GDP: Forecasts

<table>
<thead>
<tr>
<th></th>
<th>RMSFE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 1$</td>
</tr>
<tr>
<td>Oil</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>14.3</td>
</tr>
<tr>
<td>VAR-VDGARCH</td>
<td>14.3</td>
</tr>
<tr>
<td>Non-Hie SB</td>
<td>15.6</td>
</tr>
<tr>
<td>Hie SB</td>
<td>14.5</td>
</tr>
<tr>
<td>GDP</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td><strong>0.76</strong></td>
</tr>
<tr>
<td>VAR-VDGARCH</td>
<td>1.03</td>
</tr>
<tr>
<td>Non-Hie SB</td>
<td>0.78</td>
</tr>
<tr>
<td>Hie SB</td>
<td><strong>0.76</strong></td>
</tr>
</tbody>
</table>

RMSFE is the root mean square forecast error based in the predictive mean from a model calculated from the last 10 years of data (120 observations). $p$ is the lag length. Bold text identifies the smallest entry in each panel.
Table 3: 7-variable VAR: summary statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>UR</td>
<td>5.99</td>
<td>3.40</td>
<td>10.80</td>
<td>2.45</td>
</tr>
<tr>
<td>Core PCE</td>
<td>3.44</td>
<td>-6.74</td>
<td>12.29</td>
<td>5.80</td>
</tr>
<tr>
<td>Em</td>
<td>1.75</td>
<td>-10.44</td>
<td>14.74</td>
<td>7.93</td>
</tr>
<tr>
<td>Retail</td>
<td>3.18</td>
<td>-92.54</td>
<td>90.04</td>
<td>230.9</td>
</tr>
<tr>
<td>Housing</td>
<td>-0.20</td>
<td>-29.15</td>
<td>31.22</td>
<td>62.22</td>
</tr>
<tr>
<td>IP</td>
<td>2.77</td>
<td>-50.71</td>
<td>71.98</td>
<td>101.3</td>
</tr>
<tr>
<td>FFR</td>
<td>5.70</td>
<td>0.11</td>
<td>19.10</td>
<td>11.76</td>
</tr>
</tbody>
</table>

Table 4: 7-variable VAR, Predictive Likelihood and RMSFE

<table>
<thead>
<tr>
<th></th>
<th>log-PL</th>
<th>UR</th>
<th>Core PCE</th>
<th>Nonfarm Em</th>
<th>Retail</th>
<th>Housing</th>
<th>IP</th>
<th>FFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VAR(1)</td>
<td>-1782.6</td>
<td>0.149</td>
<td>1.665</td>
<td>2.047</td>
<td>14.397</td>
<td>7.496</td>
<td>9.047</td>
<td>0.202</td>
</tr>
<tr>
<td>VAR(2)</td>
<td>-1759.6</td>
<td>0.143</td>
<td>1.635</td>
<td>1.688</td>
<td>14.305</td>
<td>7.964</td>
<td>8.830</td>
<td>0.185</td>
</tr>
<tr>
<td>VAR(3)</td>
<td>-1753.9</td>
<td>0.143</td>
<td>1.567</td>
<td>1.568</td>
<td>14.356</td>
<td>8.116</td>
<td>8.710</td>
<td>0.194</td>
</tr>
<tr>
<td>VAR(4)</td>
<td>-1751.8</td>
<td>0.144</td>
<td>1.530</td>
<td>1.568</td>
<td>14.210</td>
<td>8.007</td>
<td>8.789</td>
<td>0.206</td>
</tr>
<tr>
<td>Non-hie SB-VAR(1)</td>
<td>-1708.2</td>
<td>0.155</td>
<td>2.025</td>
<td>1.321</td>
<td>17.753</td>
<td>7.150</td>
<td>10.425</td>
<td>0.201</td>
</tr>
<tr>
<td>Non-hie SB-VAR(2)</td>
<td>-1713.1</td>
<td>0.152</td>
<td>2.200</td>
<td>1.258</td>
<td>18.515</td>
<td>7.360</td>
<td>10.235</td>
<td>0.200</td>
</tr>
<tr>
<td>Non-hie SB-VAR(3)</td>
<td>-1695.6</td>
<td>0.153</td>
<td>2.300</td>
<td>1.230</td>
<td>18.123</td>
<td>7.298</td>
<td>10.152</td>
<td>0.205</td>
</tr>
<tr>
<td>Non-hie SB-VAR(4)</td>
<td>-1721.1</td>
<td>0.153</td>
<td>2.319</td>
<td>1.220</td>
<td>18.257</td>
<td>7.300</td>
<td>10.091</td>
<td>0.203</td>
</tr>
<tr>
<td>Hie SB-VAR(1)</td>
<td>-1643.3</td>
<td>0.152</td>
<td>1.667</td>
<td>1.772</td>
<td>14.246</td>
<td>7.368</td>
<td>8.931</td>
<td>0.190</td>
</tr>
</tbody>
</table>

The last 10 years data are used in forecasting. RMSFE means the root mean square forecast error.
Figure 1: Oil and GDP: Cumulative log-predictive likelihoods
Figure 2: Oil and GDP: Structural change probability.
Figure 3: Oil and GDP: Posterior Mean
Figure 4: Oil and GDP: Posterior mean of standard deviation and correlation from error terms
Figure 5: 7-variable VAR(1), non-hierarchical model: break probability

Figure 6: 7-variable VAR(1), non-hierarchical model: red is volatility, green is the data
Figure 7: 7-variable VAR(1), hierarchical model: break probability
Figure 8: 7-variable VAR(1), hierarchical model: red is the persistence parameter, green is the data.
Figure 9: 7-variable VAR(1), hierarchical model: red is the volatility, green is the data