A New Multivariate Model with an Unknown Number of Change-points

John Maheu* Yong Song†

March 18, 2012

Preliminary

Abstract

This paper develops a new efficient approach for multivariate time series data modeling and forecasting in the presence of an unknown number of change-points. The predictive density has a closed form by assuming conjugate priors for the parameters which characterize each regime. A Markov chain Monte Carlo method takes advantage of the conjugacy to integrate out the parameters which characterize each regime, treat the regime duration as a state variable and simulate the regime allocation of the data from its posterior distribution efficiently. Two priors, one is non-hierarchical for fast computation, the other is shrinkage hierarchical to exploit the information across regimes, are proposed. The model is applied to 7 U.S. macroeconomic time series and finds strong evidence for the existence of structural instability. A general pattern of the data is similar to the great moderation. However, we discover heterogeneous dynamics with infrequent volatility jumps for individual variables. The marginal likelihood comparison shows that our approach provides superior out-of-sample forecasting performance.

1 Introduction

Multivariate time series data analysis plays a central role in macroeconomic analysis and forecasting. Linear models such as vector auto regressions (VAR) are standard tools to calculate the impulse response function and forecast. 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 return among many. 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

*University of Toronto and RCEA. Email:jmaheu@chass.utoronto.ca.
†CenSoC, University of Technology, Sydney. Email:ysong1@gmail.com.

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. In this paper, we develop 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 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), Giordani 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 problem 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 all 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 $O(T^2)$, 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 computing $O(T^2)$ models numerically, the computational burden is even heavier than Koop and Potter's (2007) method. Extending these methodologies to a multivariate framework is empirically unrealistic, since a multivariate model requires much more computation as the number of variables increases.

Current multivariate change-point models include Jochmann and Koop (2011) and Cogley and Sargent (2005). A common feature of these models is that they need to pre-specify the number of regimes as exogenous. The full posterior distribution for the allocation of the data to their respective regimes is not explored because of the aforementioned 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. Then, the Bayesian averaging method can be applied to obtain the posterior distribution for the regime allocation. However, this solution is computationally brutal and the multimodal posterior density problem in each single estimation procedure may exist and 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 for the conditional posterior distribution and provides a closed form of the predictive density. This give us a huge gain in the 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.[1] 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 the multimodal problem caused by the usual Gibbs sampler is not present in our MCMC algorithm. For a simple univariate application, we find our approach can improve the relative numeric efficiency for the posterior mean of the number of regimes by more than 100 times compared to Koop and Potter's (2007) model.

Applying the conjugate priors to VAR was investigated by Kadiyala and Karlsson (1997) for the practitioners. Recent empirical work such as Carriero et al. (2011) 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 all applied to linear models without structural change. Nonetheless, they demonstrate that a conjugate prior is practically reasonable and a helpful starting building block for a structural break model.

Regarding to the 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. (2011), which is designed to approximate the Minnesota prior (Litterman (1986)). We use this prior for the parameters characterizing each regime in the structural break model, which is labeled as non-hierarchical SB model. The advantage of using this prior is the fast computational speed. With our MCMC algorithm, for a simulated data set with 7 variables and 600 observations, if we assume a VAR(1) model in each regime, 6000 times of simulation from the posterior distribution take less than 5 seconds. The approximation of the Minnesota prior is an informative prior but covers a reasonable range for the parameter values.

The second prior is featured by a hierarchical structure and shrinkage parameters. The hierarchical structure is on the parameters which characterize each regime. It is able to exploit the information across regimes. Pesaran et al. (2006) applied this method to a univariate model to improve forecasting. The shrinkage (e.g., Belmonte et al. (2011)) is another effective methodology for parsimony in Bayesian modelling. Out second prior has hyper parameters which shrink towards the first prior. We can control the tightness parameters to reflect the prior belief for the variation of the hierarchical structure. In our application to the U.S. economy, the tight hierarchical model provides superior forecasting than the non-hierarchical model and other alternatives including linear models.

From the view of computation, the aforementioned hierarchical structure is unaffordable for a time series model as Maheu and Gordon (2008) even for the univariate case. This is because their approach requires $O(T^2)$ times numeric approximation. Each time is associated with a MCMC estimation applied to a distinct subset of the data. For a univariate time series with 600 observations, it could take one day or even longer to estimate by using a regular PC. A simple hierarchical structure may easily increase the estimation time to months, or even years! This is obviously impractical even if we use some regular CPU parallel programming techniques.[2] Because our conjugate prior assumption guarantees an analytic form of the

[1] It is called Rao-Blackwellisation. See Casella and Robert (1996).

[2] A large scale parallel computing method for Bayesian modeling by using GPUs is discussed by Durham

predictive density conditional on the current regime duration, the numeric approximation with the MCMC algorithm is avoided and the computation speed is greatly improved. Hence, the hierarchical structure is affordable in our approach and the estimation can be done in a reasonable time.

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 the dimension from $O(2^T)$ to $O(T^2)$, but their approach simulates the regime allocation and the set of parameters which characterize each regime individually. To sample them jointly, theoretically we need to consider $O(2^T)$ scenarios. Each scenario has a distinct path of break points and can represent a state if the time-varying parameters characterizing each regime can be integrated out. However, it is impractical to estimate all $O(2^T)$ scenarios by using a current PC. This paper applies the assumption similar to Chib (1998) to reduce the dimension of the state space from $O(2^T)$ to $O(T)$. In detail, 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 characterizes 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 in our MCMC algorithm. This assumption frees the model from path dependence and enable the 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, we introduce a new efficient approach to model multivariate time series data in the presence of an unknown number of structural changes. Two challenges, one is the multi-dimensionality and the other is the unknown number of regimes, is solved by two reasonable assumptions. The first assumption is using a conditional conjugate prior for the parameters characterizing each regime to produce a closed form of the predictive density. This analytic form provides a super fast estimation procedure for a non-hierarchical model and allows us to explore the hierarchical structure, which is impractical for other existing models. The second assumption is that the data before a break point is uninformative for the current regime conditional on the prior for the parameters characterizing each regime. By using these two assumptions, we reduce the state space dimension to $O(T)$ and treat the regime duration as a state variable by integrating out the parameters which characterize each regime. In the MCMC algorithm, the regime allocation is jointly simulated with the parameters which characterize each regime from their posterior distribution.

Our approach has three features attractive to the practitioners. First, the number of regimes is estimated endogenously. This allows the full posterior distribution of the number of regimes as well as regime allocation to be explored. All time-varying parameters are sampled jointly, so the estimation is efficient in terms of mixing. Second, the conjugate prior makes

and Geweke (2011). However, this technique may not be appropriate for a multivariate structural change model for two reasons. First, detecting the sudden change of the regime and the sampling of the covariance matrices are challenging for the random walk Metropolis-Hastings algorithm, which is the only sampling method used in Durham and Geweke (2011). Second, the inference of the time-varying parameters conditional on the full sample is not available since this technique focuses on the real time forecasting.

the estimation of the non-hierarchial model very fast because no numeric approximation is involved. Lastly, the shrinkage hierarchical structure is parsimonious and able to exploit the information across regimes to improve forecasting. We find combining the learning of the priors and the shrinkage provides superior prediction in the application.

We apply our new approach to a VAR model with 7 variables. They are unemployment rate(UR), Core personal consumption expenditure(PCE), non-farm employment(EM), retail sales(Retail), housing starts level(Housing), industrial production index (IP) and the federal funds rate(FFR). The new model discovers very strong evidence for the existence of structural changes. Another interesting finding is that although a simple prior approximating the Minnesota prior is useful and competitive in out-of-sample forecasting as in Carriero et al. (2011), introducing the hierarchical structure helps to further improve the predictive(marginal) likelihood. This new finding implies that a Minnesota prior may miss capturing some information, which can be learned by the hierarchical structure.

The rest of the paper is organized as follows. Section 2 introduces the model. Section 3 apply the model to 7 U.S. macroeconomic variables. Section 4 concludes.

2 Model

In this section, we will first introduce the conjugate prior for a simple linear multivariate model. Then, we introduce the non-hierarchical and hierarchical structural break model together with the MCMC methods for their estimation. The prior elicitation is discussed in detail afterwards.

2.1 Linear multivariate model

A simple linear multivariate model has the following form:

$$y_t = \Phi' x_t + e_t,\ e_t \overset{\text{i.i.d.}}{\sim} \mathbf{N}(0, \Sigma). \tag{1}$$

y_t is a $N \times 1$ vector of the data at time t. x_t is a $M \times 1$ vector of the regressors. Φ is a $M \times N$ matrix of the coefficients. Each e_t is a $N \times 1$ zero mean i.i.d. normal random vector.

Let T to 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 $E = (e_1, e_2, \ldots, e_T)'$ as the stacking up of y_t's, x_t's and e_t's, respectively. (1) can also be written as

$$Y = X\Phi + E,\ E \sim \mathbf{MN}(0, \Sigma, I), \tag{2}$$

where $\mathbf{MN}(0, \Sigma, I)$ means a matrix normal distribution. The first parameter, which is a $T \times N$ zero matrix, represents the mean of the error matrix E. The second parameter, the $N \times N$ matrix Σ, is proportional to the covariance matrix of each row of matrix E, namely, e_t. The last parameter, the $T \times T$ identity matrix I, is proportional to the covariance matrix of each column of the matrix E. The identity matrix I 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 $\text{vec}(E) \sim \mathbf{N}(0, \Sigma \otimes I)$ or $\text{vec}(E') \sim \mathbf{N}(0, I \otimes \Sigma)$.[3] Appendix A introduces the matrix normal distribution in detail.

[3] Σ and I are not identified up to a scalar. This does not affect any derivation or inference in this paper.

A special case is the VAR model, which is the focus of this paper. For a VAR(p) model, where p is the number of lags in the autoregression, $x_t = (1, y'_{t-1}, y'_{t-2}, \ldots, y'_{t-p})'$ and $M = Np+1$. Φ can be decomposed as $(\phi_0, \phi_1, \ldots, \phi_p)'$, where ϕ_0 is a $N \times 1$ vector of the intercepts and ϕ_i is the $N \times N$ coefficient matrix of y_{t-i} for $i = 1, \ldots, p$.

The inverse Wishart matrix normal distribution is used as the conjugate prior for the parameters (Φ, Σ):

$$\Sigma \sim IW(\underline{S}, \underline{\nu}), \tag{3}$$
$$\Phi \mid \Sigma \sim MN(\underline{\Phi}, \Sigma \otimes \underline{\Omega}). \tag{4}$$

An inverse Wishart distribution is a random distribution, from which each sample is a nonnegative definite matrix. The mean of Σ is $\mathbf{E}(\Sigma) = \frac{\underline{S}}{\underline{\nu}-N-1}$. See the appendix for the details of an inverse Wishart distribution.

The conjugacy shows that the posterior of Φ and Σ is still an inverse Wishart matrix normal distribution:

$$\Sigma \mid Y, X \sim IW(\overline{S}, \overline{\nu}) \tag{5}$$
$$\Phi \mid \Sigma, Y, X \sim MN(\overline{\Phi}, \Sigma \otimes \overline{\Omega}) \tag{6}$$

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

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

$$y_t \mid x_t \sim t(\underline{\Phi}' x_t, \frac{(1 + x'_t\underline{\Omega}x_t)\underline{S}}{\underline{\nu} + 1 - N}, \underline{\nu} + 1 - N) \tag{7}$$

Its probability density function is $p(y_t \mid x_t) = k^{-1} \left|1 + \frac{(y_t - \underline{\Phi}'x_t)'\underline{S}^{-1}(y_t - \underline{\Phi}'x_t)}{(1+x'_t\underline{\Omega}x_t)}\right|^{-\frac{\underline{\nu}+1}{2}}$, where $k = \pi^{N/2}(1 + x'_t\underline{\Omega}x_t)^{N/2}|\underline{S}|^{1/2}\frac{\Gamma((\underline{\nu}+1-N)/2)}{\Gamma((\underline{\nu}+1)/2)}$. The first two moments are $\mathbf{E}(y_t \mid x_t) = \underline{\Phi}'x_t$ and $\mathbf{Var}(y_t \mid x_t) = (1 + x'_t\underline{\Omega}x_t)\mathbf{E}(\Sigma)$.

If we use the posterior distribution, which is also an inverse Wishart matrix normal 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.

$$y_{T+1} \mid I_T \sim t(\overline{\Phi}' x_{T+1}, \frac{(1 + x'_{T+1}\overline{\Omega}x_{T+1})\overline{S}}{\overline{\nu} + 1 - N}, \overline{\nu} + 1 - N). \tag{8}$$

$I_T = (y_1, \ldots, y_T, x_1, \ldots, x_{T+1})$ represents the information available for the whole sample. Notice that we assume x_{T+1} is also known for the prediction purpose. In a VAR model, x_{T+1} is simply $y_T, y_{T-1}, \ldots, y_{T-p}$, which is consistent with the definition of I_T. For the rest of the paper, we also define $I_t = (y_1, \ldots, y_t, x_1, \ldots, x_{t+1})$ as the information up to time t, inclusive.

2.2 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 Φ_t and Σ_t to replace Φ and Σ to get

$$y_t = \Phi_t' x_t + e_t,\ e_t \overset{\text{i.i.d.}}{\sim} \mathbf{N}(0, \Sigma_t). \tag{9}$$

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 π for a structural change to occur. It the structural change happens, the new value of θ_t will be drawn from an aforementioned inverse Wishart matrix normal distribution independently. Otherwise, θ_t will stay the same as the value in the previous period.

The model is

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

$$\theta_t = \mathbf{1}(d_t = 1)\mathbf{F}_\theta + \mathbf{1}(d_t > 1)\delta_{\theta_{t-1}}. \tag{11}$$

$$y_t \mid \theta_t, x_t = \mathbf{N}(\Phi_t' x_t, \Sigma_t). \tag{12}$$

In (10), d_t is an implicitly defined time-varying parameter, which represents the regime duration up to time t. This variable will be shown to be very important and treated as the state variable for the predictive density. The regime duration d_t can take 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 θ_t is drawn from the inverse Wishart matrix normal distribution $\mathbf{F}_\theta$ as in (11). If no break appears in the previous period, the duration is increased by 1 and θ_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 θ_t.

Compared to the existing structural break models, this approach explores all the 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 θ_t is drawn from the distribution $\mathbf{F}_\theta$ independently from the value of θ_{t-1}. We make this assumption for two reasons. First of all, it is computationally feasible to calculate the predictive density by integrating out θ_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 $D \mid \pi, I_T$, which is obtainable only if the conjugate prior and the path independence are assumed. Then, conditional on the duration D, simulate Θ from the distribution $\Theta \mid D, \pi, I_T$. This is equivalent to the joint sampling from distribution $D, \Theta \mid \pi, I_T$, which is efficient based on Casella and Robert (1996).

The MCMC method in this paper is new to the existing literature and described here in details. The first step of sampling D from $D \mid \pi, I_T$ is done by taking advantage of

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.

In the current literature, a sample series of the regime indicator $S = (s_1, s_2, \ldots, s_T)$ 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. Obviously, there is a one-to-one relationship between D and S.

However, an individual value of s_t and d_t has different information content. The regime indicator s_t is able to tell how many regime 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 ith regime. By definition, $\theta_t = \tilde{\theta}_{s_t}$. On the other hand, d_t is able to tell how long the current regime lasts 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 θ_t. However, if the data in the past regime is uninformative to the current regime, the regime duration d_t can tell which sub-sample can be used to obtain the posterior and provides a predictive density by integrating out the parameters of the conditional data density in that regime, which cannot be done by using the regime indicator s_t.

In our approach, the assumption of independent sampling of new θ_t from $\mathbf{F}_\theta$ enables us to treat d_t as a state variable, because it is sufficient to produce the predictive density. Θ 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 $D \mid \pi, I_T$. The conjugate prior provides a closed form for the predictive density to accelerate the computational speed by a great amount, which makes the MCMC algorithm practical.

The forward filter is the following:

1. At $t = 1$, set $p(d_1 = 1 \mid \pi, I_1) = 1$, which is trivial.

2. The forecasting step:

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

3. The updating step:

$$p(d_t = j \mid \pi, I_t) = \frac{p(y_t \mid d_t = j, I_{t-1}) p(d_t = j \mid \pi, I_{t-1})}{p(y_t \mid \pi, I_{t-1})}$$

for $j = 1, \ldots, t$. The first term in the numerator is a student-t distribution density function as the following:

$$y_t \mid I_{t-1}, d_t \sim t(\hat{\Phi}' x_t, \frac{(1 + x_t' \hat{\Omega} x_t)\hat{S}}{\hat{\nu} + 1 - N}, \hat{\nu} + 1 - N) \tag{13}$$

with $\hat{\Phi} = \hat{\Omega}(\underline{\Omega}^{-1}\underline{\Phi} + X'_{t+1-d_t,t-1}Y_{t+1-d_t,t-1}), \hat{\Omega} = (\underline{\Omega}^{-1} + X'_{t+1-d_t,t-1}X_{t+1-d_t,t-1})^{-1}, \hat{\nu} = \underline{\nu} + d_t - 1$, and $\hat{S} = \underline{S} + Y'_{t+1-d_t,t-1}Y_{t+1-d_t,t-1} + \underline{\Phi}'\underline{\Omega}^{-1}\underline{\Phi} - \hat{\Phi}'\hat{\Omega}^{-1}\hat{\Phi}$. where $X_{t+1-d_t,t-1} = (x_{t+1-d_t}, \ldots, x_{t-2}, x_{t-1})'$ and $Y_{t+1-d_t,t-1} = (y_{t+1-d_t}, \ldots, y_{t-2}, y_{t-1})'$ are the data between time $t+1-d_t$ and $t-1$ inclusive. If $d_t = 1$, which means a break happens, we have the first subscript (t) less than the second subscript $(t-1)$. In this case, $X_{t+1-d_t,t-1}$ and $Y_{t+1-d_t,t-1}$ are empty sets and all the *hat* parameters $(\hat{\Phi}, \hat{\Omega}, \hat{\nu}, \hat{S})$ is replace by the prior parameters $(\underline{\Phi}, \underline{\Omega}, \underline{\nu}, \underline{S})$.

The second term is obtained from step 2.

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

$$p(y_t \mid \pi, I_{t-1}) = \sum_{d_t=1}^{t} p(y_t \mid d_t, I_{t-1}) p(d_t \mid \pi, I_{t-1}). \tag{14}$$

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 $d_T \mid \pi, I_T$, 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 $d_{t-1} \mid I_{t-1}$. This is because $d_t = 1$ implies a structural change at time t. Hence, for any $\tau \geq t$, the data y_τ is in a new regime and independent of d_{t-1}. The distribution $d_{t-1} \mid d_t = 1, \pi, I_T$ is equivalent to $d_{t-1} \mid d_t = 1, \pi, I_{t-1}$.

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

After obtaining the durations D, simulating Θ from $\Theta \mid D, I_T$ is simply done by using the conjugacy property of (5) and (6). First convert D to a series of the aforementioned 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 τ 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 labeling 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 Θ looks more straightforward. Because Θ 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 θ_i^* is independent from the other θ_j^*'s, we can simulate each θ_i^* only conditional on the data allocated to the ith regime implied by S. In detail , θ_i^* is randomly drawn from the following distribution.

$$\Sigma_i^* \sim \mathbf{IW}(\overline{S}_i, \overline{\nu}_i) \tag{15}$$

$$\Phi_i^* \mid \Sigma_i^* \sim \mathbf{MN}(\overline{\Phi}_i, \Sigma_i^* \otimes \overline{\Omega}_i) \tag{16}$$

with $\overline{\Phi}_i = \overline{\Omega}_i(\underline{\Omega}^{-1}\underline{\Phi} + X_i^{*\prime}Y_i^*), \overline{\Omega}_i = (\underline{\Omega}^{-1} + X_i^{*\prime}X_i^*)^{-1}, \overline{\nu}_i = \underline{\nu} + d_i^*$, and $\overline{S}_i = \underline{S} + Y_i^{*\prime}Y_i^* + \underline{\Phi}'\underline{\Omega}^{-1}\underline{\Phi} - \overline{\Phi}_i'\overline{\Omega}^{-1}\overline{\Phi}_i$. 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 ith regime, respectively. d_i^* is the duration of the ith regime.

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

1. Sample $D, \Theta \mid \pi, I_T$.
2. Sample $\pi \mid D$.

2.3 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 one hundred variables in a few minutes or even seconds. Section 2.4 proposes a reasonable conjugate prior to approximate the Minnesota prior. For the application in Section 3, this prior works well both in terms of marginal likelihood and predictive likelihood.

Meanwhile, the fast computational speed gives us the privilege to adventure more structures and exploit more information from the data. For a simple example, we can try thousands of different priors for sensitivity check. In this paper, we pursue a more systematical way by proposing a hierarchical structure to exploit the information across regimes. It is also a natural solution to the prior robustness check and intrinsically more objective than the Minnesota prior.

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

These priors are assumed as follows:

$$\underline{\Omega} \sim \mathbf{IW}(\Omega_0, \omega_0), \tag{17}$$

$$\underline{\Phi} \mid \underline{\Omega} \sim \mathbf{MN}(M_0, \Lambda_0 \otimes \underline{\Omega}), \tag{18}$$

$$\underline{S} \sim \mathbf{W}(S_0, \tau_0), \tag{19}$$

$$\underline{\nu} \sim \mathbf{G}(a_0, b_0)\mathbf{1}(\underline{\nu} \geq N + 2). \tag{20}$$

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

1. Sample $D, \Theta \mid \pi, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu}, I_T$ by using the joint sampler in the non-hierarchical model.
2. Sample $\pi \mid D$.
3. Sample $\underline{\Phi}, \underline{\Omega} \mid D, \Theta$
4. Sample $\underline{S} \mid D, \Theta, \underline{\nu}$.

5. Sample $\underline{\nu} \mid D, \Theta, \underline{S}$.

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 $(\underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu})$.

2.4 Priors for the non-hierarchical model

The importance of the prior elicitation for multivariate Bayesian models has been addressed by many papers. This is because a multivariate model usually involves many parameters. A seemingly harmless prior may be very informative and severely distort the inference. The worse part is that this kind of prior may be left unnoticed by the applicant. It is like a black box, from which we try to avoid.

In this paper, the prior for the non-hierarchical model is made to approximate the Minnesota prior (Litterman (1986)) for a linear VAR model. Since our approach has a linear representation for each regime, the Minnesota prior is a natural candidate for the non-hierarchical model. Notice that the Minnesota prior is not a conjugate prior, nonetheless, its essence can be captured in a systematical way by having the following properties.

1. An uninformative prior for the intercept ϕ_0.
2. A stationary series has its regression coefficients centered around 0. Meanwhile, a non-stationary series has its regression coefficients to approximate the random walk.
3. 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.
4. The volatility is calibrated by using the univariate series information.

In detail:

1. $\underline{\mathbf{\Phi}}$:

   It is the prior mean of the regression coefficient Φ_t's. In the VAR framework, $\underline{\mathbf{\Phi}}$ can be written as $(\underline{\phi}_0, \underline{\phi}_1, \ldots, \underline{\phi}_p)'$, where $\underline{\phi}_0$ is the prior mean of the intercept vectors and $\underline{\phi}_i$ is the prior mean of the coefficient matrix for y_{t-i}. We set $\underline{\mathbf{\Phi}}$ equal to 0 except $\underline{\phi}_1^{(ii)}$, which is the coefficient of the first lag of the ith variable in the ith equation. For example, if $\underline{\phi}_1^{(11)} = 1$, the prior mean implies the first variable $y_t^{(1)}$ is a random walk process, or $y_t^{(1)} = y_{t-1}^{(1)} + e_t^{(1)}$.

   Let $\underline{\phi}_1^{(ii)} = 1$ if the process is non-stationary and 0 otherwise. The judgement can be done by using a formal statistical test or based on experience.

2. $\underline{S}$ and $\underline{\nu}$:

   Estimate a univariate AR model for each variable to get the estimated residual variance $\hat{\sigma}_i^2$ for $i = 1, \ldots, N$. Then, set the prior mean of Σ as $\text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2)$. Specifically,

   $$\underline{S} = (\underline{\nu} - N - 1) diag(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2)$$

$$\underline{\nu} = N + 2$$

The value of $\underline{\nu}$ guarantees the existence of the second moment of y_t. It is also necessary for the numerical stability in the MCMC algorithm.

3. $\underline{\mathbf{\Omega}}$:

We assume $\text{Var}(\phi_k^{(ij)}) = \gamma \frac{\sigma_i^2}{k^2 \sigma_j^2}$, where the superscript (ij) and subscript k means that $\phi_k^{(ij)}$ is the coefficient of the kth lag of the jth variable in the ith equation. γ controls the global tightness of the prior and k^2 in the denominator shows the variance shrinks towards 0 as the lag length increases. The ratio $\frac{\sigma_i^2}{\sigma_j^2}$ is for normalization.

The matrix normal assumption implies $\text{Var}(\phi_k^{(ij)}) = \sigma_i^2 \underline{\Omega}_{1+N(k-1)+j,1+N(k-1)+j}$. So we set $\underline{\Omega}_{1+N(k-1)+j,1+N(k-1)+j} = \gamma \frac{1}{k^2 \sigma_j^2}$ to meet the assumption of $\text{Var}(\phi_k^{(ij)}) = \gamma \frac{\sigma_i^2}{k^2 \sigma_j^2}$. The $M \times M$ matrix $\underline{\Omega}$ is then given by

$$diag(100, \frac{\gamma}{\sigma_1^2}, \ldots, \frac{\gamma}{\sigma_N^2}, \frac{\gamma}{4\sigma_1^2}, \ldots, \frac{\gamma}{4\sigma_N^2}, \ldots, \frac{\gamma}{p^2\sigma_1^2}, \ldots, \frac{\gamma}{p^2\sigma_N^2},)$$

or

$$\begin{bmatrix}
100 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{\gamma}{\sigma_1^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{\gamma}{\sigma_N^2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{\gamma}{4\sigma_1^2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{\gamma}{4\sigma_N^2} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\gamma}{p^2\sigma_N^2}
\end{bmatrix}$$

The value of the top left element is set as 100 to imply $\text{Var}(\phi_0^{(i)}) = 100\sigma_i^2$, which reflects an uninformative but appropriate prior.[4]

2.5 Priors for the hierarchical model

The prior for the hierarchial model is related to that of the non-hierarchical model in the sense that the hierarchical prior is set to center around the non-hierarchical prior and can be controlled to shrink towards it. The first feature is the hierarchical structure. It allows us to estimate these hyper parameters instead of fixing them exogenously. Hence, we can learn from the information across regimes. The second feature is shrinkage. This is necessary since

[4]It can be changed to a much larger value such as $1.0e10$. For a linear model, it is equivalent to Carriero et al. (2011) from the empirical point of view, but their approach needs a training sample because the prior is inappropriate.

we want the model to be parsimonious, especially in the multivariate framework. An over dispersed prior may harm the forecasting or even contaminate the in-sample estimation.

In detail, the prior in (17) is set as

$$\Omega_0 = (\omega_0 - M - 1)\underline{\Omega}_{\text{non-hie}},$$
$$\omega_0 = M + 2,$$

where $\underline{\Omega}_{\text{non-hie}}$ is the pre-specified value of $\underline{\Omega}$ in the non-hierarchical model. So we have $\mathbf{E}(\underline{\Omega}) = \underline{\Omega}_{\text{non-hie}}$. ω_0 is chosen to be greater than or equal to $M + 2$ for numeric stability in the MCMC algorithm. Increasing ω_0 shrinks the prior of $\underline{\Omega}$ towards the constant matrix $\underline{\Omega}_{\text{non-hie}}$.

For the prior in (18), we assume

$$M_0 = \underline{\Phi}_{\text{non-hie}},$$
$$\Lambda_0 = \lambda \mathrm{E}(\Sigma_{\text{non-hie}}),$$

where λ is a positive scalar representing the tightness of the prior for $\underline{\Phi}$. $\mathbf{E}(\Sigma_{\text{non-hie}})$ is the prior mean of the covariance matrix Σ in the non-hierarchical model. This is similar to the prior $\underline{\Phi}_t$ in the non-hierarchical model except that it does not depend on Σ_t. The reason is that we do not want the high dimensionality and different normalization of the variables to bring us any unrealistic prior without our notice. In other words, we try to avoid the aforementioned black box problem. The prior is obviously centered at $\underline{\Phi}_{\text{non-hie}}$. As λ decreases, it shrinks towards the value of the non-hierarchical model.

For the prior in (19), we set

$$S_0 = \frac{1}{\tau_0}\mathbf{E}(\Sigma_{\text{non-hie}}),$$

where $\tau_0 = N + 2$, or any positive scalar $\geq N + 2$. This prior has a mean of $\mathbf{E}(\Sigma_{\text{non-hie}})$. It shrinks towards the mean as τ_0 increases.

The last parameter $\underline{\nu}$ in (20) has a truncated gamma distribution as $\underline{\nu} \sim \mathbf{G}(\nu_a, \nu_b)\mathbf{1}(\underline{\nu} \geq N + 2)$. If $\nu_a \to \infty$ and $\frac{\nu_b}{\nu_a} \to$ some constant c, $\underline{\nu}$ shrinks towards that constant. In the application, we set $\nu_a = \nu_b = 5$.

3 Application to the U.S. economy

3.1 Basic Results

We apply our model to a system with 7 variables downloaded from CITIBASE. They are: unemployment rate (UR), Core PCE (1200 $\times$ log difference of the level), nonfarm employment (1200 $\times$ log difference of the level), retail sales (1200 $\times$ log difference of the level), housing starts level (100 $\times$ log difference of the level), industrial production index (1200 $\times$ log difference of the level), federal funds rate.[5] There are 625 observations from 1959M02 to 2011M02. Summary statistics are shown in Table 1. We can notice that the variables are normalized differently from the variance column. This is not a problem to us since it is automatically corrected in the prior elicitation procedure.

[5]This is the same set of variables used in Carriero et al. (2011).

Table 1: 7-variable VAR: summary statistics

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

The models used are the hierarchical SB-VAR(1) and the non-hierarchical SB-VAR(1), where SB means structural break and VAR(1) means each regime has a VAR(1) representation.[6] For the non-hierarchical model, we estimate two versions. The first one fixes the structural break probability $p = 0.01$, while the second one estimates the p by assuming a prior $p \sim \mathbf{B}(1, 9)$ as a beta distribution. For the hierarchical SB-VAR(1) model, we estimate three versions to investigate the effect of shrinkage in the multivariate setting with structural instability. The first and the second assumes a tight ($\lambda = 0.1$), a loose ($\lambda = 1$) and a black box ($\Lambda_0 = \mathbf{I}$) prior for $\underline{\Phi}$. Linear VAR(p) models are applied as the benchmark models.

The model comparison is based on Kass and Raftery (1995). They suggest to compare the model $\mathcal{M}_i$ and $\mathcal{M}_j$ by the log Bayes factors $\log(BF_{ij})$, where $BF_{ij} = \frac{p(Y_{1,T}|\mathcal{M}_i)}{p(Y_{1,T}|\mathcal{M}_j)}$ is the ratio of the marginal likelihoods. A positive value of $\log(BF_{ij})$ supports model $\mathcal{M}_i$ against $\mathcal{M}_j$. Quantitatively, Kass and Raftery (1995) consider the results barely worth a mention for $0 \leq \log(BF_{ij}) < 1$; positive for $1 \leq \log(BF_{ij}) < 3$; strong for $3 \leq \log(BF_{ij}) < 5$; and very strong for $\log(BF_{ij}) \geq 5$.

Geweke and Amisano (2010) have shown the marginal likelihood can be written as the product of one period ahead predictive likelihoods $p(Y_{1,T}) = \prod_{t=1}^{T} p(y_t \mid Y_{1,t-1})$. Hence the marginal likelihood in essence is based on the out-of-sample forecasting. The model comparison by the Bayes factor automatically penalize over parametrization and abides by Ockham's razor.

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 previous literature. Second, the volatility has a decreasing pattern in general and is in line with the great moderation. Meanwhile, some volatility jumps exist. Lastly, our approach find the number of regimes is different from most of the current models. Current model 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 10 regimes, which is new to the multivariate analysis of the U.S. economy.

Figure 1 plots the posterior probabilities of structural changes implied by the non-hierarchical model.[7] It shows a visible structural change at 1987M03 and some evidence

[6] The first data point is truncated as the regressor.

[7] The two versions of the non-hierarchical models produce similar figures. We produce the figure generated

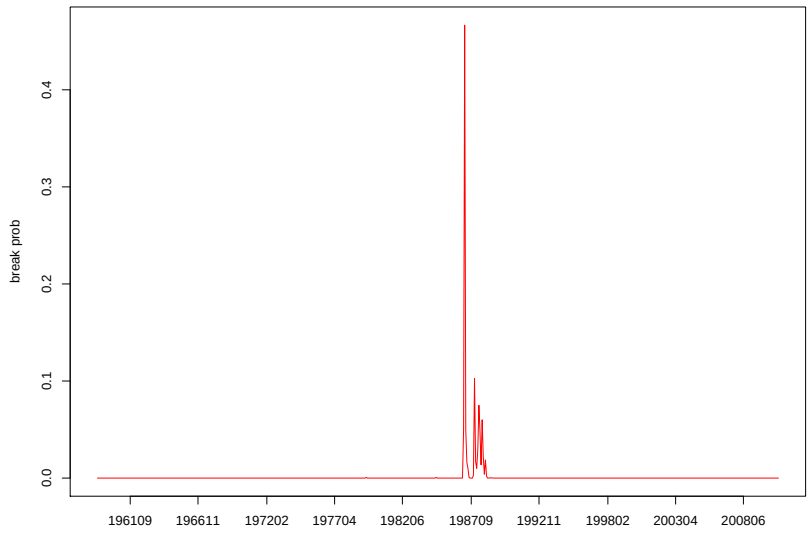

Figure 1: 7-variable VAR, non-hierarchical model: break probability

of structural instability in the end of 1987 and early 1988.

To understand the structural change in the multivariate system. Figure 2 shows the posterior mean of the volatility of each individual variable ($\sigma_t^{(i)} = \sqrt{\Sigma_t^{(ii)}}$, for $i = 1, \ldots, 7$ and $t = 1, \ldots, T$). All variables are featured by 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).

We report the results implied by the hierarchical model with the tight prior, which is the optimal model in Table 2.[8] Figure 3 shows the smoothed break probability over time. The hierarchical model finds more than two regimes implied by the non-hierarchical model. Define a break happens if the posterior break probability $p(d_t = 1 \mid I_T) > 0.5$, the model identifies 1960M06, 1979M10, 1982M12 and 2009M01 as the change-points. If using $p(d_t = 1 \mid I_T) > 0.2$ 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 of is consistent with Koop and Potter (2007) in their univariate analysis of U.S. GDP growth and inflation data.

Figure 4 plots the posterior mean of $\phi_{1,t}^{(ii)}$ 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.

by the model in which p is estimated.

[8]All three versions of the hierarchical structural break models produce similar posterior mean of the time-varying parameters.

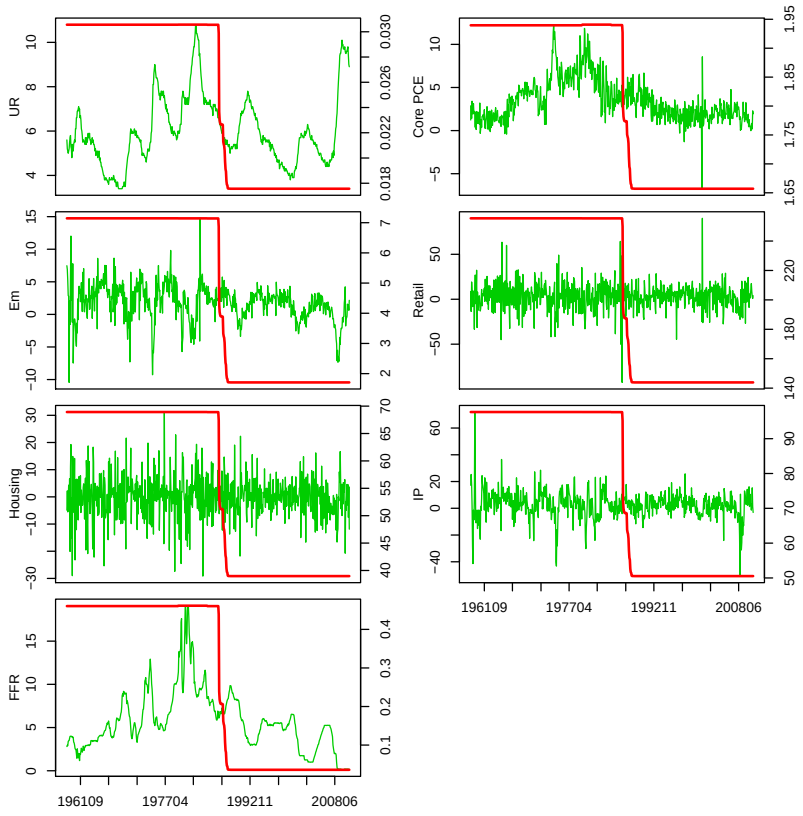

Figure 2: 7-variable VAR, non-hierarchical model: volatility

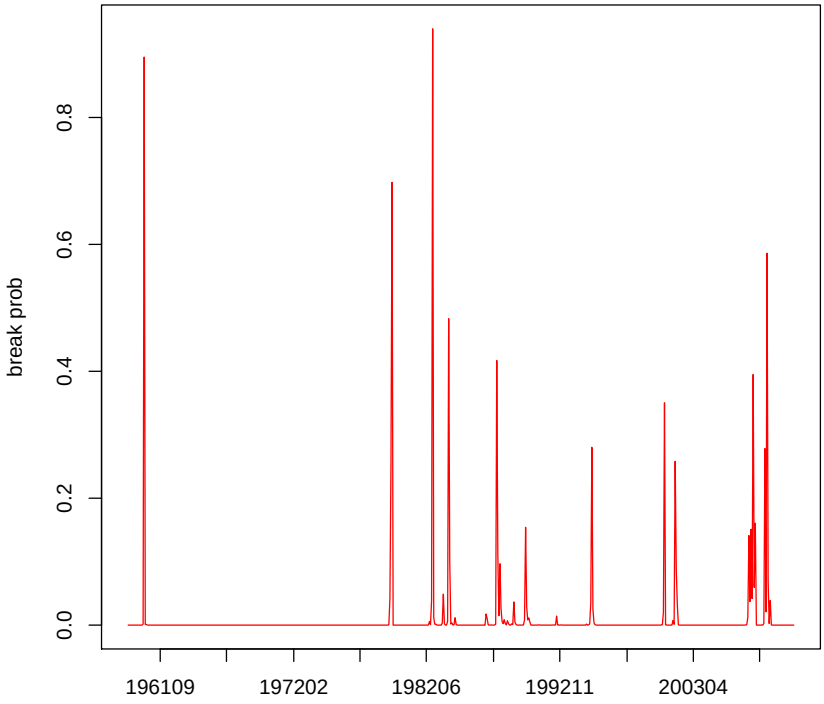

Figure 3: 7-variable VAR, hierarchical model: break probability

Figure 5 plots the posterior mean of the volatility $\sigma_t^{(i)}$ over time. All variables except Core PCE have a trend of decreasing volatility over time. And we have identified late 1979 as structural change points. Meanwhile, these dynamics are not exact 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 their volatility increased instead of decreased after the 1979 break. The retail sales decreased its volatility after 1979, but after 1984M03 the volatility jumped up. The industrial production has its volatility decreased after early 80's, however, a volatility uproar appeared during the most recent financial crisis.

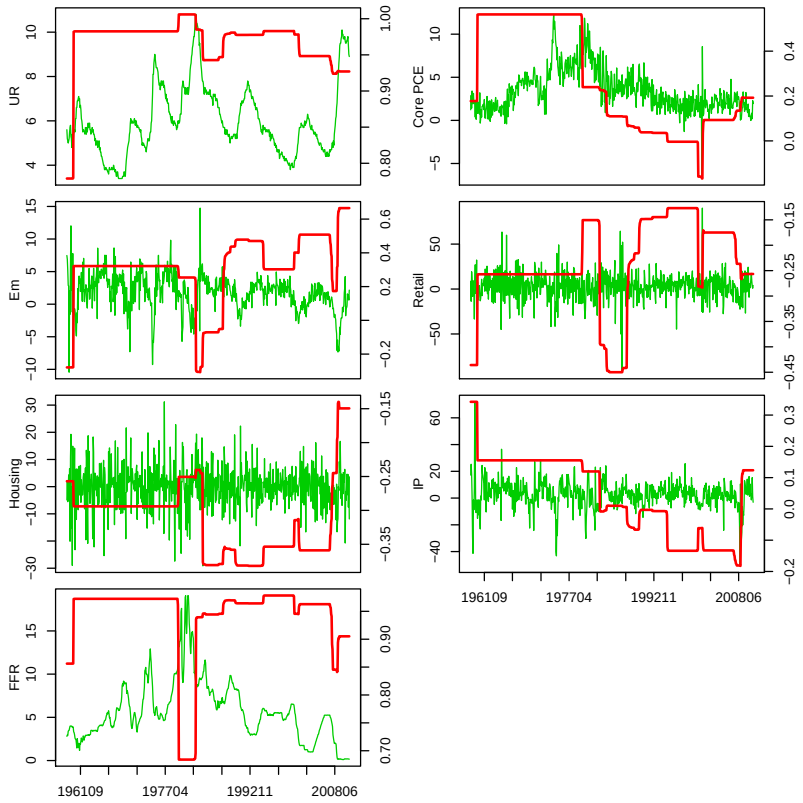

Figure 4: 7-variable VAR, hierarchical model: AR(1) coefficient

Table 2 shows the marginal likelihoods for model comparison. Structural break models outperform the linear models strongly. The log Bayes factor between the worst structural break model (-9568.1) and the best linear model (-9643.5) is 75.4, which is a very strong evidence supporting the structural break model. Among the structural break models, the best is the hierarchical model with a tight prior on the mean of the regression coefficients $\underline{\Phi}$. The hierarchical model with the loose prior or the black box prior for $\underline{\Phi}$ has similar performance to both non-hierarchical models. The application confirms the power of hierarchical structure and shrinkage in terms of forecasting.

Table 3 shows the predictive likelihoods and the predictive means for the last 10 years of the sample. The first panel is the predictive likelihoods. It shows that the hierarchical

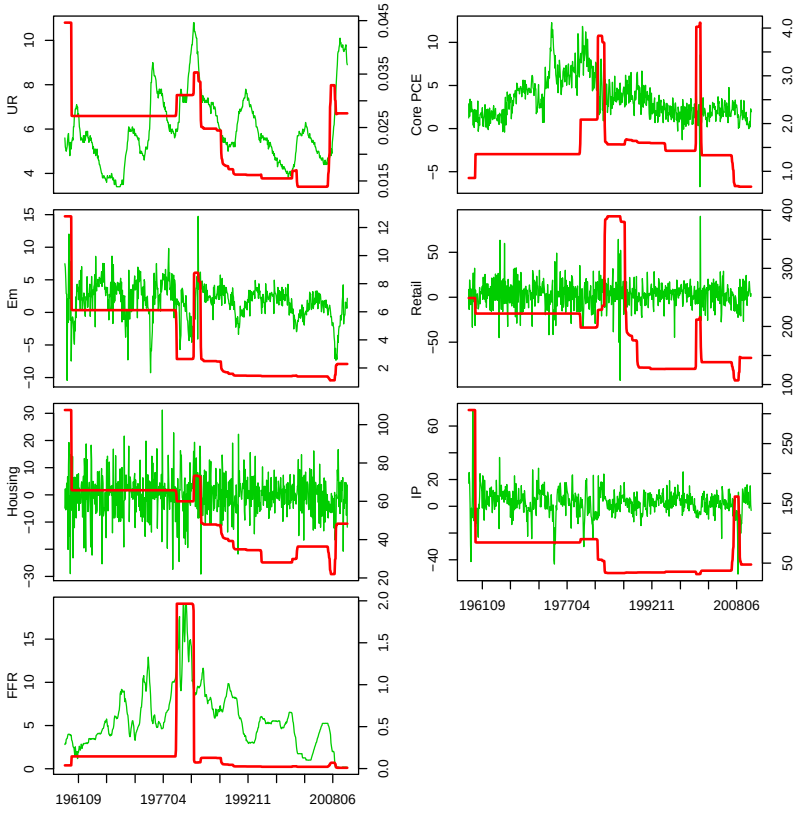

Figure 5: 7-variable VAR, hierarchical model: volatility

Table 2: 7-variable VAR, Log marginal likelihood

	Marginal likelihood
VAR(1)	-9768.1
VAR(2)	-9650.8
VAR(3)	-9643.5
VAR(4)	-9660.4
Non-hierarchical SB-VAR(1): $p = 0.01$	-9567.4
Non-hierarchical SB-VAR(1)	-9568.1
Hierarchical SB-VAR(1): $\Lambda_0 = I$	-9564.0
Hierarchical SB-VAR(1): $\Lambda_0 = E(\Sigma_{\text{non-hie}})$	-9569.9
Hierarchical SB-VAR(1): $\Lambda_0 = 0.1E(\Sigma_{\text{non-hie}})$	**-9531.0**

structural break models perform better than the non-hierarchical structural break models and the linear models. The ability of learning for the hierarchical structure shows its advantage after a training sample. In the second panel, each row represents the root mean squared error of the mean forecasts. The best mean forecast is represented by the bold font. Not as the linear models, the hierarchical structural break models always provide a reasonable mean forecasting. For every variable, the hierarchical SB-VAR(1) models do not perform the worst in predictive means.

Table 3: 7-variable VAR, Predictive Likelihood and RMSE

	PL	UR	Core PCE	Nonfarm Em	Retail	Housing	IP	FFR
VAR(1)	-1783.7	0.149	1.667	2.060	14.398	7.489	9.059	0.202
VAR(2)	-1760.9	0.144	1.637	1.687	14.315	7.982	8.834	0.**186**
VAR(3)	-1756.1	0.**144**	1.569	1.604	14.388	8.156	**8.715**	0.196
VAR(4)	-1755.6	0.146	**1.530**	1.579	**14.276**	8.045	8.800	0.209
SB-VAR(1): Non-hie $p = 0.01$	-1714.2	0.148	2.795	**1.315**	19.146	**7.239**	12.199	0.198
SB-VAR(1): Non-hie	-1707.2	0.148	2.775	1.331	18.913	7.333	11.369	0.190
SB-VAR(1): $\Lambda_0 = I$	**-1695.5**	0.149	2.635	1.392	18.533	7.349	10.829	0.191
SB-VAR(1): $\lambda = 1$	-1719.8	0.156	2.911	1.332	19.364	7.269	12.694	0.218
SB-VAR(1): $\lambda = 0.1$	-1696.5	0.148	2.636	1.381	18.767	7.402	10.811	0.190

Forecast the last 10 year.

3.2 Model Selection

Maheu and Song (2012) proposed a univariate model by using this methodology and applied it to Canada inflation forecasting. They found adding more lags can not improve out-of-sample forecasts if the structural instability is controlled for. This section investigates whether this is true in our multivariate application.

Table 4 shows that the marginal likelihood are strongly improved by moving from VAR(1) to VAR(2) in each regime. The best model is the hierarchial SB-VAR(2) with the tight priors.

This is also supported by the predictive likelihoods of the last 10 years in Table 5. The difference is that when the prior is very tight $\lambda = 0.01$, it performs better in the beginning of the sample because of the shrinkage. Meanwhile, setting $\lambda = 0.1$ provides a more flexible learning ability.

Table 4: 7-variable VAR, Log marginal likelihood

	Marginal likelihood
Non-hierarchical SB-VAR(1): $p = 0.01$	-9567.4
Non-hierarchical SB-VAR(2): $p = 0.01$	-9505.9
Non-hierarchical SB-VAR(3): $p = 0.01$	-9537.5
Non-hierarchical SB-VAR(4): $p = 0.01$	-9577.7
Non-hierarchical SB-VAR(1)	-9568.1
Non-hierarchical SB-VAR(2)	-9503.9
Non-hierarchical SB-VAR(3)	-9535.5
Non-hierarchical SB-VAR(4)	-9575.8
Hierarchical SB-VAR(1): $\Lambda_0 = I$	-9564.0
Hierarchical SB-VAR(1): $\Lambda_0 = E(\Sigma_{\text{non-hie}})$	-9569.9
Hierarchical SB-VAR(1): $\Lambda_0 = 0.1E(\Sigma_{\text{non-hie}})$	-9531.0
Hierarchical SB-VAR(1): $\Lambda_0 = 0.01E(\Sigma_{\text{non-hie}})$	-9520.0
Hierarchical SB-VAR(2): $\Lambda_0 = I$	-9492.3
Hierarchical SB-VAR(2): $\Lambda_0 = E(\Sigma_{\text{non-hie}})$	-9494.5
Hierarchical SB-VAR(2): $\Lambda_0 = 0.1E(\Sigma_{\text{non-hie}})$	-9445.1
Hierarchical SB-VAR(2): $\Lambda_0 = 0.01E(\Sigma_{\text{non-hie}})$	**-9433.9**

Figure 6 and 7 plots the structural break probabilities and he posterior means of the volatilities over time. We can see that the hierarchical SB-VAR(2) identifies more break points than the non-hierarchical SB-VAR(1) but less than the hierarchical SB-VAR(1). One major break point at 1987 is the same as in Figure 1. And we can see the volatilities decrease after that period for all variables. Meanwhile, A structural change is consistent with the recent financial crisis.

4 Conclusion

This paper develops a new efficient approach for multivariate time series data modeling and forecasting in the presence of an unknown number of change-points. The predictive density has a closed form by assuming conjugate priors for the parameters which characterize each regime. A Markov chain Monte Carlo method takes advantage of the conjugacy to integrate out the parameters which characterize each regime, treat the regime duration as a state variable and simulate the regime allocation of the data from its posterior distribution efficiently.

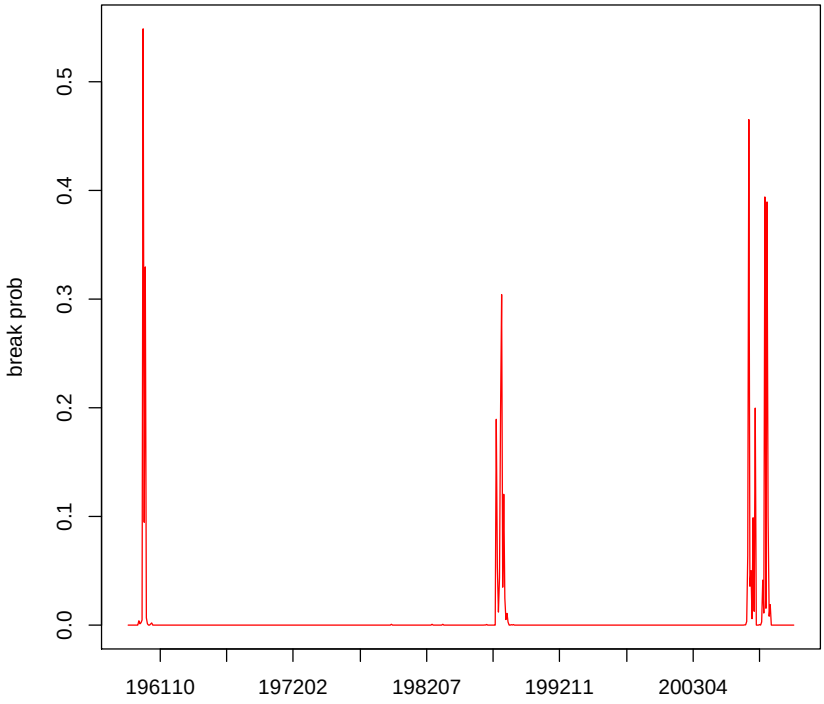

Figure 6: 7-variable VAR(2), hierarchical model: break probability

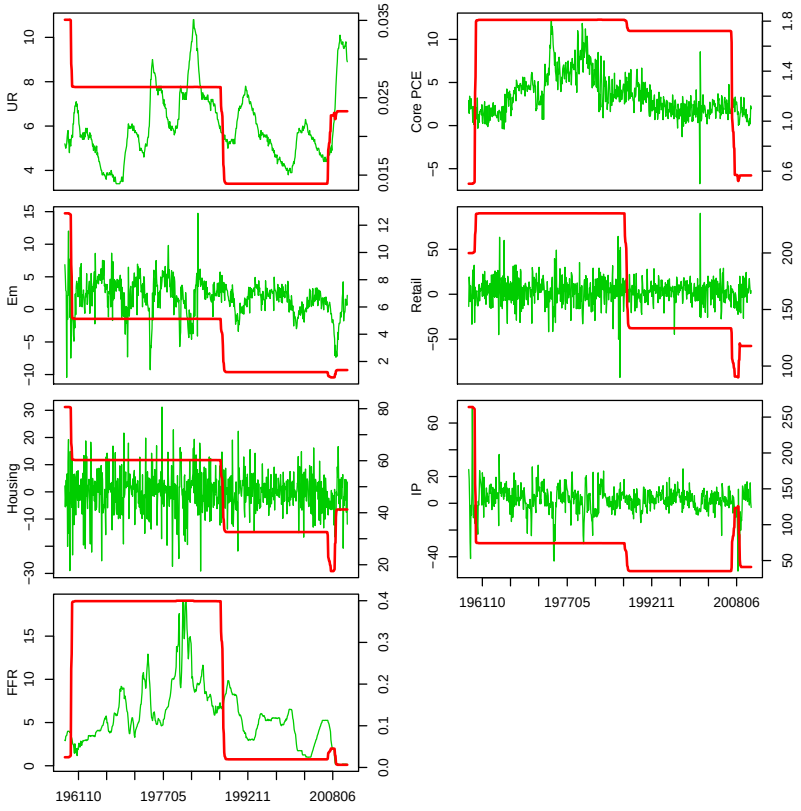

Figure 7: 7-variable VAR(2), hierarchical model: volatility

Table 5: 7-variable VAR, Predictive Likelihood and RMSE

	PL	UR	Core PCE	Nonfarm Em	Retail	Housing	IP	FFR
SB-VAR(1): Non-hie $p = 0.01$	-1714.2	0.148	2.795	1.315	19.146	**7.239**	12.199	0.198
SB-VAR(2): Non-hie $p = 0.01$	-1691.4	0.140	2.669	1.164	17.972	7.611	10.070	0.177
SB-VAR(3): Non-hie $p = 0.01$	-1704.1	0.144	2.682	1.139	18.332	7.761	9.747	0.176
SB-VAR(4): Non-hie $p = 0.01$	-1714.4	0.144	2.597	1.236	17.314	7.864	9.519	0.174
SB-VAR(1): Non-hie	-1707.2	0.148	2.775	1.331	18.913	7.333	11.369	0.190
SB-VAR(2): Non-hie	-1685.8	0.140	2.642	1.158	17.945	7.649	9.209	0.172
SB-VAR(3): Non-hie	-1700.3	0.144	2.682	**1.135**	18.469	7.847	9.110	0.173
SB-VAR(4): Non-hie	-1716.0	0.144	2.630	1.247	17.479	7.919	9.129	0.172
SB-VAR(1): $\Lambda_0 = I$	-1695.5	0.149	2.635	1.392	18.533	7.349	10.829	0.191
SB-VAR(1): $\lambda = 1$	-1719.8	0.156	2.911	1.332	19.364	7.269	12.694	0.218
SB-VAR(1): $\lambda = 0.1$	-1696.5	0.148	2.636	1.381	18.767	7.402	10.811	0.190
SB-VAR(1): $\lambda = 0.01$	-1721.5	0.153	2.897	1.358	19.771	7.312	13.342	0.220
SB-VAR(2): $\Lambda_0 = I$	-1689.7	**0.140**	2.794	1.171	18.139	7.678	9.327	0.179
SB-VAR(2): $\lambda = 1$	-1692.4	0.140	2.750	1.200	18.140	7.723	8.759	0.172
SB-VAR(2): $\lambda = 0.1$	**-1682.2**	0.140	2.806	1.183	18.978	7.655	9.802	0.172
SB-VAR(2): $\lambda = 0.01$	-1698.7	0.143	2.750	1.200	18.839	7.679	11.220	0.192

Forecast the last 10 year.

Two priors are proposed for model estimation. The first prior is non-hierarchical and approximates the Minnesota prior. Its advantage is the super fast computationally speed. The second prior assumes a hierarchical structure to exploit the information across regimes and a control over shrinkage for parsimony.

The new approach is applied to 7 U.S. macroeconomic time series. The structural break models strongly dominate the linear alternatives. The best model is the one with the hierarchical prior and tighter shrinkage. It identifies more regimes than what has been implied by the existing literature. A general trend of volatility decrease is consistent with the great moderation. However, we find heterogeneous dynamics with infrequent volatility jumps for different variables.

A Inverse Wishart - Matrix Normal prior

1. **$\boldsymbol{\Sigma}$**:

   The error covariance matrix Σ has a Inverse-Wishart distribution. Its prior mean is

   $$E(\Sigma) = \frac{\underline{S}}{\underline{\nu} - N - 1}$$

   The variance of each element

   $$Var(\Sigma_{ij}) = \frac{(\underline{\nu} - N + 1)\underline{S}_{ij}^2 + (\underline{\nu} - N - 1)\underline{S}_{ii}\underline{S}_{jj}}{(\underline{\nu} - N)(\underline{\nu} - N - 1)^2(\underline{\nu} - N - 3)}$$

   Its density function is given by

   $$p(\Sigma) = \frac{|\underline{S}|^{\underline{\nu}/2}|\Sigma|^{-(\underline{\nu}+N+1)/2} etr\{-\frac{1}{2}\underline{S}\Sigma^{-1}\}}{2^{\underline{\nu}N/2}\Gamma_N(\underline{\nu}/2)}$$

   Γ_p is multivariate gamma function, which is$\Gamma_p(a) = \int_{S>0} etr\{-S\}|S|^{a-(p+1)/2}dS$ 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

   $$p(\sigma^2) = \frac{\underline{s}^{\underline{\nu}/2}(\sigma^2)^{-\underline{\nu}/2-1}\exp\{-\frac{\underline{s}}{2}\sigma^{-2}\}}{2^{\underline{\nu}/2}\Gamma(\underline{\nu}/2)}.$$

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

   The precision matrix P, which is the inverse of the covariance matrix Σ, has a Wishart distribution $W(\underline{P}, \underline{\nu})$, where $\underline{P} = \underline{S}^{-1}$. It has density

   $$p(P) = \frac{|\underline{P}|^{-\underline{\nu}/2}|P|^{(\underline{\nu}-N-1)/2} etr\{-\frac{1}{2}\underline{P}^{-1}P\}}{2^{\underline{\nu}N/2}\Gamma_N(\underline{\nu}/2)}$$

   A special case is when $N = 1$, then $P = \sigma^{-2}$ has a gamma distribution with

   $$p(\sigma^{-2}) = \frac{\underline{s}^{\underline{\nu}/2}(\sigma^{-2})^{\underline{\nu}/2-1}\exp\{-\frac{\underline{s}}{2}\sigma^{-2}\}}{2^{\underline{\nu}/2}\Gamma(\underline{\nu}/2)}.$$

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

2. **$\boldsymbol{\Phi}$**:

   The regression coefficient matrix Φ has a matrix normal distribution. Each column of Φ, $\Phi_{\cdot j}$, is the regression coefficients for the jth equation and has a multivariate normal distribution

   $$\Phi_{\cdot j} \mid \Sigma \sim N(\underline{\Phi}_{\cdot j}, \Sigma_{jj}\underline{\Omega})$$

Each row of Φ, $\Phi_{i.}$, is the coefficients of impact from the same source across equations.

$$\Phi_{i.} \mid \Sigma \sim N(\underline{\Phi}_{i.}, \Sigma\underline{\Omega}_{ii})$$

The density function is

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

B Sample from a matrix Gaussian

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

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

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

C Sample from an Inverse-Wishart distribution

Generate Σ from a Inverse-Wishart, $IW(\underline{S}, \underline{\nu})$, by

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

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

D Sample the hierarchical prior

1. $\underline{\mathbf{\Phi}}$ and $\underline{\mathbf{\Omega}}$:

   The prior is matrix normal and inverse-Wishart.

   $$\begin{aligned} \underline{\Omega} &\sim IW(\Omega_0, \omega_0) \\ \underline{\Phi} \mid \underline{\Omega} &\sim MN(M_0, \Lambda_0 \otimes \underline{\Omega}) \end{aligned}$$

   The conditional posterior $\underline{\mathbf{\Phi}}, \underline{\mathbf{\Omega}} \mid \{\Sigma_i, \Phi_i\}_{i=1}^K$ is

   $$\begin{aligned} \underline{\Omega} \mid \{\Sigma_i, \Phi_i\}_{i=1}^K &\sim IW(\Omega_1, \omega_1) \\ \underline{\Phi} \mid \underline{\Omega}, \{\Sigma_i, \Phi_i\}_{i=1}^K &\sim MN(M_1, \Lambda_1 \otimes \underline{\Omega}) \end{aligned}$$

   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. $\underline{\mathbf{S}}$:

   The prior of $\underline{S}$ is a Wishart $W(S_0, \tau_0)$. The conditional posterior is also Wishart.

   $$\underline{S} \mid \underline{\nu}, \{\Sigma_i\}_{i=1}^{K} \sim W(S_1, \tau_1)$$

   with

   $$S_1^{-1} = S_0^{-1} + \sum_{i=1}^{K} \Sigma_i^{-1}$$
   $$\tau_1 = \tau_0 + K\underline{\nu}$$

3. $\underline{\nu}$:

   The prior is a Gamma $G(a_0, b_0)$. The conditional posterior has no convenient form.

   $$\begin{aligned} p(\underline{\nu} \mid \underline{S}, \{\Sigma_i\}_{i=1}^{K}) &= p_G(\underline{\nu}; a_0, b_0) \prod_{i=1}^{K} p(\Sigma_i \mid \underline{S}, \underline{\nu}) \\ &\propto p_G(\underline{\nu}; a_0, b_0) \prod_{i=1}^{K} \left\{ \frac{|\underline{S}|^{\underline{\nu}/2}}{2^{\underline{\nu}N/2}\Gamma_N(\underline{\nu}/2)} |\Sigma_i|^{-\frac{\underline{\nu}+N+1}{2}} \right\} \\ &\propto \underline{\nu}^{a_0-1} e^{-b_0\underline{\nu}} \frac{|\underline{S}|^{K\underline{\nu}/2}}{2^{K\underline{\nu}N/2}\Gamma_N^K(\underline{\nu}/2)} \prod_{i=1}^{K} \left\{ |\Sigma_i|^{-\frac{\underline{\nu}+N+1}{2}} \right\} \end{aligned}$$

   The log of the last equation (after discarding more constants) is

   $$\frac{K\log(|\underline{S}|) - 2b_0 - KN\log(2) - \sum_{i=1}^{K}\log(|\Sigma_i|)}{2}\underline{\nu} - K\log(\Gamma_N(\underline{\nu}/2)) + (a_0 - 1)\log(\underline{\nu}).$$

   The sampling method of $\underline{\nu}$ is a M-H step with a proposal distribution of

   $$\underline{\nu}^{(i)} \sim G(\xi, \xi/\underline{\nu}^{(i-1)})$$

E Marginal likelihood

The marginal likelihood is calculated by using the bridge-sampling estimator in Frühwirth-Schnatter (2004), Meng and Wong (1996).

$$\hat{p}_t(Y) = \hat{p}_{t-1}(Y) \frac{L^{-1} \sum_{l=1}^{L} \frac{\hat{p}(\tilde{\phi}^{(l)}|Y)}{Lq(\tilde{\phi}^{(l)})+M\hat{p}(\tilde{\phi}^{(l)}|Y)}}{M^{-1} \sum_{l=1}^{M} \frac{q(\tilde{\phi}^{(m)})}{Lq(\tilde{\phi}^{(m)})+M\hat{p}(\tilde{\phi}^{(m)}|Y)}},$$

and

$$\hat{p}(\phi \mid Y) = \frac{p^*(\phi \mid Y)}{\hat{p}_{t-1}(Y)} = \frac{p(Y \mid \phi)p(\phi)}{\hat{p}_{t-1}(Y)}$$

where ϕ represents the parameters of a model. $\phi^{(l)}$'s are simulated from an importance density q; and $\phi^{(m)}$'s are the posterior samples from the MCMC sampler. The above two procedures are iterated until convergence.

This method is from $L^{-1} \sum_{l=1}^{L} \hat{p}(\tilde{\phi}^{(l)} \mid Y) \to \int p(\phi \mid Y) q(\phi) d\phi$ and $M^{-1} \sum_{l=1}^{M} q(\tilde{\phi}^{(m)}) \to \int q(\phi) p(\phi \mid Y) d\phi$ are equivalent. Frühwirth-Schnatter (2004) showed the mean-squared error of $\log \hat{p}(Y)$ is approximated by

$$\frac{1}{L} \frac{V_q\left(\frac{p(\phi|Y)}{\omega q(\phi)+(1-\omega)p(\phi|Y)}\right)}{E_q^2\left(\frac{p(\phi|Y)}{\omega q(\phi)+(1-\omega)p(\phi|Y)}\right)} + \frac{\rho_f(0)}{M} \frac{V_p\left(\frac{q(\phi)}{\omega q(\phi)+(1-\omega)p(\phi|Y)}\right)}{E_p^2\left(\frac{q(\phi)}{\omega q(\phi)+(1-\omega)p(\phi|Y)}\right)},$$

where $\omega = \frac{L}{L+M}$ and $\rho_f(0)$ is the normalized spectral density of $f = \frac{q(\phi)}{\omega q(\phi)+(1-\omega)p(\phi|Y)}$ at frequency 0.

$$\hat{\rho}_f(0) = 1 + 2 \sum_{s=1}^{S} \left(1 - \frac{s}{S+1}\right) r_s$$

and

$$r_s = \frac{1}{M} \sum_{m=s+1}^{M} \frac{(f^{(m)} - \overline{f})(f^{(m-s)} - \overline{f})}{s_f^2}.$$

$\overline{f}$ and s_f^2 are the sample mean and sample variance of f.

For the MSB-LSV model, $\phi = (p, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu})$. The importance density for p is a beta density implied by the posterior of mean of K, $\tilde{K}$. $q(\underline{\Omega})$ is inverse Wishart with parameters $\tilde{\Omega}_1, \tilde{\omega}_1$, which are the posterior means of Ω_1 and ω_1. $q(\underline{\Phi} \mid \underline{\Omega})$ is matrix normal with parameter $\tilde{M}_1, \tilde{\Lambda}_1$ which are the posterior means of M_1 and Λ_1. $q(\underline{S})$ is a Wishart with parameters $\tilde{S}_1, \tilde{\tau}_1$, which are the posterior means of S_1 and τ_1. $q(\underline{\nu})$ is a gamma with mean and variance matching the moments of the posterior.

$$q(p) = \mathbf{B}(\tilde{K} - 1, T - \tilde{K})$$

$$
\begin{aligned}
q(\underline{\Omega}) &= \mathbf{IW}(\tilde{\Omega}_1, \tilde{\omega}_1) \\
q(\underline{\Phi} \mid \underline{\Omega}) &= \mathbf{MN}(\tilde{M}_1, \tilde{\Lambda}_1 \otimes \underline{\Omega}) \\
q(\underline{S}) &= \mathbf{W}(\tilde{S}_1, \tilde{\tau}_1) \\
q(\underline{\nu}) &= \mathbf{G}(\tilde{\nu}_a, \tilde{\nu}_b),
\end{aligned}
$$

where $\frac{\tilde{\nu}_a}{\tilde{\nu}_b}$ and $\frac{\tilde{\nu}_a}{\tilde{\nu}_b^2}$ match the posterior mean and variance. So $\tilde{\nu}_b = \frac{E(\underline{\nu}|Y)}{V(\underline{\nu}|Y)}$ and $\tilde{\nu}_a = \frac{E^2(\underline{\nu}|Y)}{V(\underline{\nu}|Y)}$

References

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

Belmonte, M., Koop, G., and Korobilis, D. Hierarchical shrinkage in time-varying parameter models. 2011.

Carriero, A., Clark, T., and Marcellino, M. Bayesian vars: specification choices and forecast accuracy. *FRB of Cleveland Working Paper No. 1112*, 2011.

Casella, G. and Robert, C.P. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1): 81, 1996.

Chib, S. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

Clark, T.E. and McCracken, M.W. Averaging forecasts from vars with uncertain instabilities. *Journal of Applied Econometrics*, 25(1):5–29, 2010.

Cogley, T. and Sargent, T.J. Drifts and volatilities: monetary policies and outcomes in the post WWII US. *Review of Economic Dynamics*, 8(2):262–302, 2005.

Durham, G. and Geweke, J. Massively parallel sequential monte carlo for bayesian inference. 2011.

Frühwirth-Schnatter, S. Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. *Econometrics Journal*, 7(1):143–167, 2004.

Geweke, J. and Amisano, G. Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting*, 2010.

Giordani, P., Kohn, R., and Van Dijk, D. A unified approach to nonlinearity, structural change, and outliers. *Journal of Econometrics*, 137(1):112–133, 2007.

Jochmann, M. and Koop, G. Regime-switching cointegration. *Working Papers*, 2011.

Kadiyala, K. and Karlsson, S. Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics*, 12(2):99–132, 1997.

Kass, R.E. and Raftery, A.E. Bayes factors. *Journal of the American Statistical Association*, 90(430), 1995.

Kim, C.J. and Nelson, C.R. Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. *Review of Economics and Statistics*, 81(4):608–616, 1999.

Koop, G. and Potter, S.M. Estimation and forecasting in models with multiple breaks. *Review of Economic Studies*, 74(3):763, 2007.

Litterman, R.B. Forecasting with bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, pages 25–38, 1986.

Liu, C. and Maheu, J.M. Are there structural breaks in realized volatility? *Journal of Financial Econometrics*, 6(3):326–360, 2008.

Maheu, J. and Song, Y. A new structural break model with application to canadian inflation forecasting. *working paper*, 2012.

Maheu, J.M. and Gordon, S. Learning, forecasting and structural breaks. *Journal of Applied Econometrics*, 23(5):553–584, 2008.

Meng, X.L. and Wong, W.H. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

Pesaran, M.H., Pettenuzzo, D., and Timmermann, A. Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73(4):1057–1084, 2006.

Stock, J.H. and Watson, M.W. Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics*, pages 11–30, 1996.

Wang, J. and Zivot, E. A Bayesian time series model of multiple structural changes in level, trend, and variance. *Journal of Business & Economic Statistics*, 18(3):374–386, 2000.