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

Maheu, John M and Song, Yong

McMaster University, University of Melbourne

May 2017

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

John M. Maheu† Yong Song‡

First draft Dec 2015
This draft May 2017

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. The model is extended to independent breaks in regression coefficients and the volatility parameters. Two empirical applications show the improvements the model has over benchmarks. In a macro application with 7 variables we empirically demonstrate the benefits from moving from a multivariate structural break model to a set of univariate structural break models to account for heterogeneous break patterns across data series.

*Maheu thanks SSHRC of Canada for financial support and the 2015 Eminent Research Scholarship, University of Melbourne.

†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 have highlighted 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 usually involves intensive computation, most of the change-point models are applied to univariate time series. Existing multivariate change-point models have restrictions on 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 speed of estimation of the new approach is increased 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), 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 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 so many combinations of break points that exploring them exhaustively is impractical.[1] 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), Koop et al. (2011) and Pettenuzzo and Timmermann (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

[1]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.

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.[2]

Some may regard the conjugate prior as a drawback of our approach. This assumption is necessary to achieve the closed form results and computational benefits. 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 with 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 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 on a regular desktop PC.[3]

The second new prior features a hierarchical structure with shrinkage hyper parameters. 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. (2014)) 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 for a univariate setting. 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 important for multivariate models because of the overparameterization problem. In other words, the curse of dimensionality may make a seemingly harmless hierarchical prior have a 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.

[2]This is called Rao-Blackwellisation. See Casella and Robert (1996).

[3]For instance, computations were done on a CPU Intel Core i5-6500 3.2GHz quad-core and 8GB memory.

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. 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 five 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. Multiperiod out-of-sample forecasts incorporate the probability of multiple structural breaks out-of-sample. Lastly, the priors are automatically adjusted to different normalizations, because they are calibrated according to the Minnesota prior.

Besides the need for a conjugate prior, a second potential issue is that all parameters entering the data density are assumed to break simultaneously.[4] We extend the model to allow for independent breaks in regression parameters and covariance parameters[5], nevertheless, if all or a subset of parameters are breaking independently a time-varying parameter models may perform better as we show in one application.

Two empirical applications are considered. The first is an application to oil and real GDP. The multivariate hierarchical structural break models produce superior density forecasts for several forecast horizons compared to a number of popular benchmark specification. Several break points are identified and in general, the hierarchical prior is significantly better in terms of Bayes factors compared to the structural break model without this prior specification. In the second application to a macro model with 7 variables we find the univariate hierarchical structural break model of Maheu and Song (2014) applied to each data series independently produces the best density forecasts. Point forecasts from the structural break models are competitive and are generally best from the univariate specifications. These results underscore the importance of allowing parameters to break at different times. Posterior estimates of break patterns in the different univariate models display considerable heterogeneity.

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

[4]We thank a referee for pointing out the importance of this issue.

[5]Other papers that decouple structural change in the conditional mean and conditional variance are Maheu and Song (2014) and Bauwens et al. (2017).

break model followed by Section 4 which extends the model with an hierarchical prior as well as independent break processes for the regression parameters and covariance matrix. 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. Conclusions 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 $y_t = (y_{1t}, y_{2t}, ..., y_{Nt})'$ as follows:

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

where x_t is a $M \times 1$ vector of independent variables and Φ 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 Σ.

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

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

The data Y is $T \times N$, X is $T \times M$ and the error term E is $T \times N$. The notation $MN(0, \Sigma, I)$ 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 Σ 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 $\text{vec}(E) \sim N(0, \Sigma \otimes I)$ or $\text{vec}(E') \sim N(0, I \otimes \Sigma)$.

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

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

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 (Φ, Σ) as Carriero et al. (2015):

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

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 approximating 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 shrink to zero as their lag length increases.

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 $\hat{v}_i^2$. The priors are set as follows:

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

$$E(\Sigma) = \begin{pmatrix} \hat{v}_1^2 & 0 & ... & 0 \\ 0 & \hat{v}_2^2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \hat{v}_N^2 \end{pmatrix},$$

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

2. We set

$$\underline{\Phi} = \begin{pmatrix} 0 & 0 & ... & 0 \\ \mathbf{1}(nonstationary) & 0 & ... & 0 \\ 0 & \mathbf{1}(nonstationary) & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \mathbf{1}(nonstationary) \\ 0 & 0 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & 0 \end{pmatrix}$$

   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

$$\underline{\Omega} = \gamma \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{\hat{v}_1^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\hat{v}_N^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4\hat{v}_1^2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4\hat{v}_N^2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{p^2\hat{v}_N^2} \end{pmatrix},$$

where scalar γ 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 Φ_{ij} is $\text{Var}(\Phi_{ij}|\Sigma = E(\Sigma)) = \hat{v}_j^2 \underline{\Omega}_{ii}$. For instance, Φ_{21} is the coefficient of $y_{1,t-1}$ in equation $y_{1,t}$ and its variance is $\text{Var}(\Phi_{21}|\Sigma = E(\Sigma)) = \hat{v}_1^2 \underline{\Omega}_{22} = \gamma$. 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 $\text{Var}(\Phi_{N+1+i,j}) = \hat{v}_j^2 \underline{\Omega}_{N+1+i,N+1+i} = \gamma \frac{\hat{v}_j^2}{4\hat{v}_i^2}$. 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 $\text{Var}(\Phi_{1j}|\Sigma = E(\Sigma)) = \gamma \hat{v}_j^2$, which reflects a proper prior with a reasonable range over the parameter space.[6]

2.2 Posterior

The posterior of Φ and Σ is still an Inverse Wishart-Matrix Normal distribution by conjugacy:

$$\Sigma|Y, X \sim IW(\overline{S}, \overline{\nu}) \tag{5}$$

$$\Phi|\Sigma, Y, X \sim MN(\overline{\Phi}, \Sigma, \overline{\Omega}) \tag{6}$$

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 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

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

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

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.

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

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 $y_T, y_{T-1}, \ldots, y_{T-p}$.

[6] 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.

3 Non-hierarchical structural break model (SB-VAR)

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 obtain

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

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. If a structural change happens, the new value of θ_t is drawn from the inverse Wishart matrix normal distribution. Otherwise, θ_t stays the same as the value in the previous period.

The SB-VAR model is:

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

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

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

In (10), d_t is an implicitly defined time-varying parameter, which represents the regime duration up to time t. This variable 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 θ_t is drawn from the inverse Wishart matrix normal distribution F_θ 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 existing structural break models, this approach explores all possible change-points as do 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 F_θ independently from the value of θ_{t-1}. We make this assumption for two reasons. First, 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 standpoint, it is reasonable or even preferable for some macroeconomic variables to have a sudden change of the parameters.

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.

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, we will 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, we will simulate Θ 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). Finally, sample $\pi|D$.

3.1 Prior

We assume that F_θ 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). However, in our setting the transition matrix is of dimension T allowing for up to T structural breaks. Each row of the transition matrix contains $1 - \pi$ and π with remaining terms 0. In contrast to Chib (1998), only one run of the sampler is required to estimate the number of breaks and associated parameters and estimation does not impose how many breaks should occur in-sample.

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 ith 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 θ_t. However, if the data in the past regime are 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 θ_t from F_θ 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 $D|\pi, Y_{1,T}, X_{1,T}$. 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

$$y_t|d_t = j, x_t, Y_{1,t-1}, X_{1,t-1} \sim t\left(\hat{\Phi}'x_t, \frac{(1 + x_t'\hat{\Omega}x_t)\hat{S}}{\hat{\nu} + 1 - N}, \hat{\nu} + 1 - N\right) \tag{13}$$

   with $\hat{\Phi} = \hat{\Omega}(\underline{\Omega}^{-1}\underline{\Phi} + X_{t+1-j,t-1}'Y_{t+1-j,t-1}), \hat{\Omega} = (\underline{\Omega}^{-1} + X_{t+1-j,t-1}'X_{t+1-j,t-1})^{-1}, \hat{\nu} = \underline{\nu} + j - 1$, and $\hat{S} = \underline{S} + Y_{t+1-j,t-1}'Y_{t+1-j,t-1} + \underline{\Phi}'\underline{\Omega}^{-1}\underline{\Phi} - \hat{\Phi}'\hat{\Omega}^{-1}\hat{\Phi}$. 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 $(\underline{\Phi}, \underline{\Omega}, \underline{\nu}, \underline{S})$.

   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

$$p(y_t|\pi, x_t, Y_{1,t-1}, X_{1,t-1}) = \sum_{d_t=1}^{t} p(y_t|d_t, x_t, Y_{1,t-1}, X_{1,t-1})p(d_t|\pi, Y_{1,t-1}, X_{1,t-1}). \tag{14}$$

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 $d_T|\pi, Y_{1,T}, X_{1,T}$, 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 $d_{t-1}|\pi, Y_{1,t-1}, X_{1,t-1}$, step 3 of the forward filter. 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 $d_{t-1}|d_t = 1, \pi, Y_{1,T}, X_{1,T}$ 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 Θ 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 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 for introducing S is to help the sampling of Θ look more straightforward. Because Θ can only take 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:

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

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

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 $B(\pi_a, \pi_b)$, the conditional posterior of π 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 (H-SB-VAR)

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. Moreover, 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 θ_i^*'s are drawn from the pre-specified distribution F_θ. In this section, we propose to use these values to learn F_θ 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 F_θ.

These priors are assumed as follows:

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

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

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

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

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

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

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

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 for 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 $\Omega_0 = (\omega_0 - M - 1)\underline{\Omega}_{\text{non-hie}}$, so we have $E(\underline{\Omega}) = \underline{\Omega}_{\text{non-hie}}$ and $sd(\underline{\Omega}_{ii}) = 2\underline{\Omega}_{\text{non-hie},ii}$. Notice that ω_0 can be adjusted(increased) to reflect shrinkage. We need $\omega_0 > M + 1$.
- Set $M_0 = \underline{\Phi}_{\text{non-hie}}$ and $\Lambda_0 = \lambda \begin{pmatrix} \hat{v}_1^2 & 0 & ... & 0 \\ 0 & \hat{v}_2^2 & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \hat{v}_N^2 \end{pmatrix}$, where λ controls the shrinkage. This implies that $Var(\underline{\Phi}_{ij}|\underline{\Omega}) = \lambda \hat{v}_j^2 \underline{\Omega}_{ii}$.
- Set $S_0 = \frac{1}{\tau_0}\underline{S}_{\text{non-hie}}$ and $\tau_0 = N$, so we have $E(\underline{S}) = \underline{S}_{\text{non-hie}}$ and $sd(\underline{S})_{ii} = \left(\frac{2}{\tau_0}\right)^{1/2} \underline{S}_{\text{non-hie},ii}$. The τ_0 can be increased to reflect shrinkage. We need $\tau_0 > N - 1$.
- For $\underline{\nu} > N + 1$ set $a_0 = 10$ and $b_0 = \frac{a_0}{\underline{\nu}_{\text{non-hie}}}$. If no restriction, the prior mean of $\underline{\nu}$ is $\underline{\nu}_{\text{non-hie}}$. As $a_0 \to \infty$, the $\underline{\nu}$ shrinks to $\underline{\nu}_{\text{non-hie}}$.

4.2 Independent Hierarchical Structural Break Model

Maheu and Song (2014) has shown how to estimate a univariate change-point model with independent breaks in regression coefficients and the volatility. This section shows how to estimate such model in the multivariate framework.

The conditional data density is the same as (9). The difference is that there are two duration counters: one (d_t^{Φ}) represents the duration of the regression coefficients and the other (d_t^{Σ}) represents the duration of the covariance matrix. Specifically,

$$d_t^{\Phi} = \begin{cases} d_{t-1}^{\Phi} + 1, & \text{w.p. } 1 - \pi_{\Phi} \\ 1, & \text{w.p. } \pi_{\Phi} \end{cases} \tag{21}$$

$$\Phi_t = \begin{cases} \sim F_{\Phi}, & \text{if } d_t^{\Phi} = 1 \\ \Phi_{t-1}, & \text{o.w.} \end{cases} \tag{22}$$

$$d_t^{\Sigma} = \begin{cases} d_{t-1}^{\Sigma} + 1, & \text{w.p. } 1 - \pi_{\Sigma} \\ 1, & \text{w.p. } \pi_{\Sigma} \end{cases} \tag{23}$$

$$\Sigma_t = \begin{cases} \sim F_{\Sigma}, & \text{if } d_t^{\Sigma} = 1 \\ \Sigma_{t-1}, & \text{o.w.} \end{cases} \tag{24}$$

$$y_t | \Phi_t, \Sigma_t, x_t \sim \mathbf{N}(\Phi_t' x_t, \Sigma_t). \tag{25}$$

π_{Φ} and π_{Σ} are the probabilities of a structural break in the regression parameters and the covariance matrix, respectively. Because we cannot apply the conjugate prior in this model, it is not feasible to integrate Φ_t and Σ_t out simultaneously. Meanwhile, we can still integrate out one series of Φ_t or Σ_t conditional on the other. The meta distributions F_{Φ} and F_{Σ} are assumed to be:

$$F_{\Sigma} \sim IW(\underline{S}, \underline{\nu}) \tag{26}$$

$$F_{\Phi} \sim MN(\underline{\Phi}, \Sigma^*, \underline{\Omega}). \tag{27}$$

The distribution (26) is the same as (3), but (27) is different from (4) and we lose conjugacy. A new value of Φ_t does not depend on the value of Σ_t.

A hierarchical prior on F_{ϕ} and F_{Σ} is the same as (17)-(20). The parameter space is $D^{\Phi} = (d_1^{\Phi}, ..., d_T^{\Phi})$, $D^{\Sigma} = (d_1^{\Sigma}, ..., d_T^{\Sigma})$, $\Phi = (\Phi_1, ..., \Phi_T)$, $\Sigma = (\Sigma_1, ..., \Sigma_T)$, π_{Φ}, π_{Σ}, $\underline{\Phi}$, $\underline{\Omega}$, $\underline{S}$, and $\underline{\nu}$. 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 $D^{\Phi}, \Phi \mid D^{\Sigma}, \Sigma, \pi_{\Phi}, \pi_{\Sigma}, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu}, Y_{1,T}, X_{1,T}$ by using the joint sampler in the non-hierarchical model.

2. Sample $D^{\Sigma}, \Sigma \mid D^{\Phi}, \Phi, \pi_{\Phi}, \pi_{\Sigma}, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu}, Y_{1,T}, X_{1,T}$ by using the joint sampler in the non-hierarchical model.

3. Sample $\pi_{\Phi} \mid D^{\Phi}$.

4. Sample $\pi_{\Sigma} \mid D^{\Sigma}$.

5. Sample $\underline{\Phi}, \underline{\Omega} \mid D^{\Phi}, D^{\Sigma}, \Phi, \Sigma$

6. Sample $\underline{S}|D^{\Phi}, D^{\Sigma}, \underline{\nu}, \Phi, \Sigma$.

7. Sample $\underline{\nu}|D^{\Phi}, D^{\Sigma}, \underline{S}, \Phi, \Sigma$.

The priors are set the same as in the hierarchical structural break model except Σ^*, which is set the same as the non-hierarchical structural break model.

Finally, a special case of this model is when breaks are restricted to the regression parameters but the covariance matrix is fixed or breaks in the covariance matrix but regression parameters do not break. This case be achieved by setting $\pi_{\Sigma} = 0$ or $\pi_{\Phi} = 0$ and dropping some of the sampling steps above. For example, for only breaks in the regression parameters, $\pi_{\Sigma} = 0$, steps 2 and 4 are dropped and minor adjustments made to steps 5-7.

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 $\Psi = \{D, \Theta, \pi, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu}\}$ denote the posterior draws from the hierarchical structural break model. Then the predictive density one-period ahead conditional on Ψ is:

$$p(y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}, \Psi) = \sum_{d_{T+1}=1}^{T+1} p(y_{T+1}|d_{T+1}, x_{1,T+1}, Y_{1,T}, \Psi) p(d_{T+1}|Y_{1,T}, X_{1,T}, \Psi). \quad (28)$$

This is computed as in (14). The value of $d_{T+1} = 1$ indicates 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:

$$p(y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}) \approx \frac{1}{R} \sum_{i=1}^{R} p(y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}, \Psi^{(i)}). \quad (29)$$

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 the hierarchical structural break model given the data $Y_{1,T}$ is:

$$LML = \sum_{t=1}^{T} \log(p(y_t|x_t, Y_{1,t-1}, X_{1,t-1})). \quad (30)$$

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 (30) 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.

In order to measure long-run forecasting ability, we also report long-run predictive likelihoods in the applications. An h-period ahead density forecast is obtained through simulations. At each step out-of-sample, a break is permitted following the model structure. Specifically, for one draw of the posterior sample $\Psi^{(i)}$ conditional on $Y_{1,T}$, the last period's

time-varying parameter can be drawn as $\theta^{(i)}$ from the posterior sampler. The next period duration $d_{T+1}^{(i)}$ has probability $\pi^{(i)}$ to take value 1, which means that $\theta_{T+1}^{(i)}$ is drawn from the hierarchical distribution; otherwise $\theta_{T+1}^{(i)} = \theta_T^{(i)}$. Then, conditional on $\theta_T^{(i)}$, $y_{T+1}^{(i)}$ is randomly drawn from the VAR framework. Then, we simulate forward $d_{T+2}^{(i)}$, $\theta_{T+2}^{(i)}$ and $y_{T+2}^{(i)}$ conditional on $d_{T+1}^{(i)}$, $\theta_{T+1}^{(i)}$, $y_{T+1}^{(i)}$ and $\Psi^{(i)}$. This procedure is repeated until we obtain $d_{T+h}^{(i)}$ and $\theta_{T+h}^{(i)}$. Plugging in the observable y_{T+h}, the h-period ahead predictive likelihood is evaluated as:

$$p(y_{T+h}|x_{T+1}, Y_{1,T}, X_{1,T}) \approx \frac{1}{R}\sum_{i=1}^{R} p(y_{T+h}|Y_{1,T}, X_{1,T}, \theta_{T+h}^{(i)}, y_{T+1}^{(i)}, ..., y_{T+h-1}^{(i)}). \tag{31}$$

In our applications, we report predictive likelihoods when $h = 3, 6, 12$. We calculate the sum of the logarithm of long-run predictive likelihoods over the same sample period as the log-predictive likelihoods to measure the models' long-run density forecast ability.

Finally, the predictive mean can be computed based on (28) which integrates over all possible break points:

$$E[y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}, \Psi] = \sum_{d_{T+1}=1}^{T+1} E[y_{T+1}|d_{T+1}, x_{T+1}, Y_{1,T}, X_{1,T}, \Psi]p(d_{T+1}|Y_{1,T}, X_{1,T}, \Psi). \tag{32}$$

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 Ψ and weighted by the probability of duration d_{T+1}. The final estimate, with parameter uncertainty accounted for, is:

$$E[y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}] \approx \frac{1}{R}\sum_{i=1}^{R} E[y_{T+1}|x_{T+1}, Y_{1,T}, X_{1,T}, \Psi^{(i)}]. \tag{33}$$

The h-period ahead predictive mean is computed for $h = 3, 6, 12$ through simulation similar to that of the long run predictive likelihood. After obtaining a large sample of $y_{T+h}^{(i)}$, the h-period predictive mean is calculated as:

$$E(y_{T+h}|x_{T+1}, Y_{1,T}, X_{1,T}) \approx \frac{1}{R}\sum_{i=1}^{R} y_{T+h}^{(i)}. \tag{34}$$

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 (SB-VAR) model, the hierarchical structural change (H-SB-VAR) model and the independent hierarchical structural break model.

We further estimate the univariate hierarchical structural break model (H-SB-U) of Maheu and Song (2014) for each variable. Specifically, denote $y_{i,t}$ as the ith variable in vector y_t, it is modeled as

$$y_{i,t} = \phi_{i,t}' x_t + e_{i,t}, \quad e_{i,t} \sim N(0, \sigma_{i,t}^2). \tag{35}$$

Each variable is modeled as an AR-X subject to breaks and includes lags of the dependent variable as well as lags of the other variables so that the conditional mean of $y_{i,t}$ is identical to that appearing in the VAR models. The parameter $\theta_{i,t} = (\phi_{i,t}, \sigma_{i,t}^2)$ has a dynamic equation similar to (11) but with an hierarchical prior following Maheu and Song (2014). The marginal (predictive) likelihood for y_t is simply the product of the marginal (predictive) likelihood for $y_{i,t}$ for all i. The univariate approach treats structural change independently for each series. This may provide a flexible way to model structural changes. The cost is that the covariance structure of the innovations is completely ignored.

In order to account for potential heteroskedasticity, we also estimated a scalar vector diagonal GARCH model in the VAR setting, denoted by VAR-VDGARCH. Specifically:

$$y_t = \Phi' x_t + e_t, \quad e_t \sim N(0, H_t) \tag{36}$$

$$H_t = CC' + a \odot e_{t-1} e_{t-1}' + b \odot H_{t-1}, \tag{37}$$

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 a and b are scalars a and C is a lower triangular matrix, which is estimated.

The prior of Φ is the same as that of the linear model when Σ 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.

The last model is the time-varying parameter vector autoregression with stochastic volatility model (TVP-VAR-SV) from Cogley et al. (2010). It is a popular approach in empirical macroeconomics with dynamic flexibility and controls for heteroskedasticity. It takes the form,

$$y_t = \Phi_t' x_t + e_{yt}, \quad e_{yt} \sim N(0, B_y^{-1} H_{yt} B_y^{-1}) \tag{38}$$

$$\Lambda_t = \Lambda_{t-1} + e_{st}, \quad e_{st} \sim N(0, B_s^{-1} H_{st} B_s^{-1}) \tag{39}$$

where $\Lambda_t = \text{vec}(\Phi_t')$, e_{yt} and e_{st} are independent, B_y and B_s are lower triangular. H_{yt} and H_{st} are both diagonal with entries in which log-volatility follow an independent random walk without drift,

$$\log \sigma_{j,t}^2 = \log \sigma_{j,t-1}^2 + u_{j,t}, \quad u_{j,t} \sim NID(0, \eta_j^2), \tag{40}$$

where j indexes the volatility process in the measurement or state equation. This model is estimated using the R package bvarsv (Krueger 2015).

Table 1 shows the log-marginal likelihoods, log-predictive likelihoods and long-run log-predictive likelihoods from various models. The last 120 observations are used to compute the predictive likelihoods and long-run predictive likelihoods. We let all horizon forecasts start from the same period, so the h-period ahead forecast has only $121-h$ observations. We estimate each model using up to 4 lags and report the results that have the largest predictive likelihood. The TVP-VAR-SV does not have marginal likelihoods because it requires a training sample for prior elicitation. The univariate structural break model (H-SB-U) does not have long-run density forecasts, because the model is incomplete.

There are significant improvements from adding GARCH or SV dynamics and moving to the structural break specifications. The best models are the nonhierarhical structural change model (SB-VAR) with 2 lags and the hierarchical structural change specification (H-SB-VAR) with 3 lags in terms of the predictive likelihood and the univariate structural change model with 1 lag in terms of the marginal likelihood. The log-Bayes factors for these model versus the VAR version is more than 22. In addition, the log-predictive likelihoods of the best models are at least 18 larger than the VAR-VDGARCH model. The structural change models has a log-predictive Bayes factors of at least 10 against the TVP-VAR-SV model. The hierarchical structural change model dominates all other models strongly for any long-run density forecast.[7]

We plot the cumulative log-predictive Bayes factors from period 1 to T for the structural break model, the hierarchical version and the VAR-VDGARCH model by using the VAR(3) model as the benchmark in Figure 1. The red line is the difference between the VAR-VD-GARCH and VAR(3). The green line represents the difference between the SB-VAR(3) and VAR(3) and the blue line represents the difference between the H-SB-VAR(3) and VAR(3). The figure shows that the VAR-VDGARCH gains in predictive power over time compared to the benchmark VAR because it incorporates heteroskedasticity. The structural change models are still the best in general over time. The figure provides evidence of structural changes. For instance, at the financial crisis in 2009, we observe a jump-up in the green and blue lines at that period indicating the value of structural break in density forecasts.

Table 2 shows the root mean square forecast errors (RMSFE) for the predictive mean. The last 30 years (120 observations) are used for out-of-sample forecasts. The results are mixed. The standard VAR produces good point forecasts and when it is improved upon it comes from one the hierarchical structural change models or in one case from the VAR-VDGARCH. There is a clear benefit in moving to the hierarchical prior for the structural break models (H-SB-VAR vs SB-VAR). The H-SB-VAR model is always competitive and has the smallest RMSFE in 5 of 8 cases.

Figure 2 plots the posterior probability of structural change, $P(d_t = 1|Y_{1,T})$, over time for the H-SB-VAR(1) model. It shows a strong pattern of parameter uncertainty. 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 hierarchical structural change model using the full sample. The top panel is the oil price change

[7]The GARCH model is the most competitive model of the others with a log-predictive likelihood only around 4 points lower than the H-SB-VAR model.

and the bottom is the real GDP growth. The model that gives the best out-of-sample density forecasts also gives good in-sample fit.

Figure 4 shows the standard deviations and the correlation coefficients of the error term over time for H-SB-VAR. 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 coincides with the financial crisis. The correlation changes over time, however the deviations from zero, both positive and negative, 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: the 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), and the federal funds rate.[8] 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. In the following we report results for models with a lag length that has the largest log-predictive likelihood. For all models a lag length from 1 to 4 is considered. The exception is for the H-SB-VAR specification in which lag lengths of 1 and 2 where estimated to reduce computation time.

Based on the H-SB-VAR model 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, which is in line with the great moderation. However, 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 multivariate structural break model supports more than 5 regimes.

Figures 5 and 6 show the smoothed break probability over time for the SB-VAR and the H-SB-VAR models, respectively. The hierarchical model finds more than the two regimes identified by the non-hierarchical model. The SB-VAR has a relatively uninformative prior while the hierarchical model (H-SB-VAR) shrinks the prior towards something more plausible and hence breaks are needed to capture significant changes in the data. This results in improved forecasts that we discuss below.

Defining a break as the posterior break probability $p(d_t = 1|I_T) > 0.5$, the H-SB-VAR 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-

[8] This is the same set of variables used in Carriero et al. (2015).

points. This finding is consistent with Koop and Potter (2007) in their univariate analysis of U.S. GDP growth and inflation data.

Figure 7 plots the posterior mean of $\phi_{1,t}^{(ii)}$ over time from the H-SB-VAR model. It represents the average effect of the first lag of the variable on itself. The unemployment rate (UR) and the federal fund rate (FFR) are very persistent for most of the time, while the rest of the variables are mean reverting. Figure 8 plots the posterior mean of the volatility $\sigma_t^{(i)}$ of the innovations over time. All variables except Core PCE have a trend of decreasing volatility over time. The displayed parameter changes 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.

Following a referee's suggestion we compute the inflation gap persistence measure $R_{t,h}^2$, from Cogley et al. (2010).[9] When h increases, a rapid decreasing of $R_{t,h}^2$ towards zero means that most of the variation will be explained by the most recent shocks. This scenario implies weak persistence. If the $R_{t,h}^2$ converges to zero slowly, it shows strong persistence. A detailed description can be found in Cogley et al. (2010). Figure 9 shows $R_{t,h}^2$ over time based on the multivariate hierarchical structural break model for horizon $h = 1, 3, 6, 12$. We found the rate of decrease of persistence is quite fast in general. At $h = 1$, the persistence is already very low (around 0.2). And then it quickly drops to around 0.05 at $h = 3$. The model predicts the inflation persistence measured by the change of $R_{t,h}^2$ over h is quite low. Our results are not the same as Cogley et al. (2010), who found time-varying inflation persistence. Because the density forecast do not favour the H-SB-VAR model, as discussed below, our results should be interpreted with caution.

The same set of comparison models as in the previous application are used in out-of-sample forecasts. Table 4 shows the predictive likelihoods and the root mean square forecast errors for the last 10 years (120 observations) 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 structural break models significantly improve upon the VAR with log-predictive Bayes factors that range from 12.8 to 200. Note that adding the hierarchical prior to the SB-VAR model moves the log-predictive likelihood from -1695.6 to -1551.8, a gain of 143.8.

The univariate hierarchical structural break specification (H-SB-U) has the best density forecasts.[10] The log-predictive Bayes factor is 91.5 against the best multivariate hierarchical version and 143.8 against the non-hierarchical specification. Learning through the hierarchical structure is clearly important to improving density forecasts whether it be the multivariate or univariate model. The dominance of the univariate structural change specification indicates that a more flexible change-point model which allows for asynchronous parameter change, may work better in the multivariate setting. Separate break processes

[9] See the Appendix for the derivations of this measure.

[10] Recall that this is essentially a univariate version of the H-SB-VAR model found in Maheu and Song (2014) applied to each data series individually. See equation (35) of the previous section.

appear to be more important than modeling contemporaneous correlations in the innovations. Similarly, the TVP-VAR-SV model is superior to the multivariate structural change models. Although the TVP-VAR-SV model has a separate evolution for each time-varying parameter and stochastic volatility for the conditional covariance matrix the application of the univariate structural break model to each series is substantially better with a log-Bayes factor of 32.7 in its favour.

Figure 10 displays the difference in cumulative log-predictive likelihoods for several models against the benchmark VAR(4) model. The H-SB-U and TVP-VAR-SV are often close in the out-of-sample period but the performance of the former significantly improves around the financial crisis.

The remaining columns of the Table 4 report the root mean square forecast error of the predictive mean forecasts. The univariate structural change models often are the most accurate (4 out of 7 cases) but it can have poor performance when it is not. The VAR has the best point forecasts on Core PCE.

To see why the H-SB-U performs better than the multivariate structural change models, Figure 11 displays the posterior probability of a break for each series. These results are very different than Figure 6 in which all parameters must break at the same time. For the univariate models, except for Unemployment and Housing Starts all series display evidence of structural breaks but the frequency of breaks and the timing is very different. These results underscore the important of allowing for independent breaks in each data series and the impact this can have on forecasting.

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 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. The model is extended to independent breaks in regression coefficients and the volatility parameters. Two applications show the usefulness of our model to multivariate time series. We show that allowing for flexible independent structural break processes for each series can be very important in forecasting.

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, $\Phi_{.j}$, is the regression coefficients for the jth equation and has a multivariate normal distribution

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

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

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

The density function is

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

B Sample from a matrix Gaussian

For $\Phi|\Sigma \sim MN(\underline{\Phi}, \Sigma \otimes \underline{\Omega})$, 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.

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

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

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

   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.

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

   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.

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

   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 VAR-VDGARCH Prior

1. The prior of the regression coefficients Φ 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 θ_1 and θ_2 to satisfy $a = 0.2$ and $b = 0.7$. The prior of θ_1 and θ_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 triagular matrix). This vector is set as the prior mean.

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

F MCMC for independent break chains

1. Sample $D^{\Phi}, \Phi \mid D^{\Sigma}, \Sigma, \pi_{\Phi}, \pi_{\Sigma}, \underline{\Phi}, \underline{\Omega}, \underline{S}, \underline{\nu}, Y_{1,T}, X_{1,T}$.

   From (9), the conditional density of y_t is

   $$y_t \mid \cdot \sim N(\Phi_t' x_t, \Sigma_t)$$

   If there is no break at time t, the conditional posterior $\Phi_t \mid d_{t-1}^{\Phi} = d, Y_{1,t-1}, \cdot$ is

$$\begin{aligned}
&p(\Phi_t \mid d_{t-1}^{\Phi} = d, \cdot) \\
\propto &etr\left\{-\frac{1}{2}\left[vec(\Phi_t) - vec(\underline{\Phi})\right]'\left(\Sigma^* \otimes \Omega\right)^{-1}\left[vec(\Phi_t) - vec(\underline{\Phi})\right]\right\} \\
&\times etr\left\{-\frac{1}{2}\sum_{i=1}^{d}(y_{t-i} - \Phi_t' x_{t-i})'\Sigma_{t-i}^{-1}(y_{t-i} - \Phi_t' x_{t-i})\right\}
\end{aligned}$$

   Let's focus on the terms inside the etr operator and discard the multiplier $-\frac{1}{2}$ for notational simplicity.

$$\begin{aligned}
&\left[vec(\Phi_t) - vec(\underline{\Phi})\right]'\left(\Sigma^* \otimes \Omega\right)^{-1}\left[vec(\Phi_t) - vec(\underline{\Phi})\right] + \sum_{i=1}^{d}(y_{t-i} - \Phi_t' x_{t-i})'\Sigma_{t-i}^{-1}(y_{t-i} - \Phi_t' x_{t-i}) \\
=&vec(\Phi_t)'\left(\Sigma^* \otimes \Omega\right)^{-1} vec(\Phi_t) - 2vec(\underline{\Phi})'\left(\Sigma^* \otimes \Omega\right)^{-1} vec(\Phi_t) + const \\
&+ \sum_{i=1}^{d} vec(\Phi_t)'\left(\Sigma_{t-i}^{-1} \otimes x_{t-i}x_{t-i}'\right) vec(\Phi_t) - 2\sum_{i=1}^{d}\left((y_{t-i}'\Sigma_{t-i}^{-1}) \otimes x_{t-i}'\right) vec(\Phi_t) + const \\
=&vec(\Phi_t)'\left((\Sigma^*)^{-1} \otimes \Omega^{-1} + \sum_{i=1}^{d}\left(\Sigma_{t-i}^{-1} \otimes x_{t-i}x_{t-i}'\right)\right) vec(\Phi_t)
\end{aligned}$$

$$-2\left[vec(\underline{\Phi})'\left((\Sigma^*)^{-1}\otimes\Omega^{-1}\right)+\sum_{i=1}^{d}\left((y'_{t-i}\Sigma_{t-i}^{-1})\otimes x'_{t-i}\right)\right]vec(\Phi_t)$$

Hence

$$vec(\Phi_t)\mid D_{t-1}^{\Phi}=d,\cdot\sim N(\overline{m},\overline{H}^{-1}),$$

where $\overline{H}=(\Sigma^*)^{-1}\otimes\Omega^{-1}+\sum_{i=1}^{d}\left(\Sigma_{t-i}^{-1}\otimes x_{t-i}x'_{t-i}\right)$ and $\overline{m}=\overline{H}^{-1}\Bigg[\left((\Sigma^*)^{-1}\otimes\Omega^{-1}\right)vec(\underline{\Phi})+\sum_{i=1}^{d}\left((\Sigma_{t-i}^{-1}y_{t-i})\otimes x_{t-i}\right)\Bigg]$

Suppose that if there is no structural change, the conditional data density of y_t is

$$y_t\mid\Phi_t,d_t^{\Phi}=d+1,\cdot\sim N((I_N\otimes x'_t)vec(\Phi_t),\Sigma_t)$$

Integrating out Φ_t, we have

$$y_t\mid d_t^{\Phi}=d+1,\cdot\sim N((I_N\otimes x'_t)\overline{m},(I_N\otimes x'_t)\overline{H}^{-1}(I_N\otimes x_t)+\Sigma_t)$$

Now we can draw D^{Φ} first and then draw $\Phi\mid D^{\Phi}$ by applying the same technique in this paper.

2. Sample $D^{\Sigma},\Sigma\mid D^{\Phi},\Phi,\pi_{\Phi},\pi_{\Sigma},\underline{\Phi},\underline{\Omega},\underline{S},\underline{\nu},Y_{1,T},X_{1,T}$

   If there is no structural change ar time t, the conditional posterior $\Sigma_t\mid d_{t-1}^{\Sigma}=d,Y_{1,t-1}$ is

   $$\Sigma_t\mid d_{t-1}^{\Sigma}=d,Y_{1,t-1},\cdot\sim IW(\overline{S},\overline{\nu}),$$

   where $\overline{\nu}=\underline{\nu}+d$ and $\overline{S}=\underline{S}+\sum_{i=1}^{d}(y_{t-i}-\Phi'_{t-i}x_{t-i})(y_{t-i}-\Phi'_{t-i}x_{t-i})'$.

   The conditional data density of y_t is

   $$y_t\mid\Sigma_t,d_t^{\Phi}=d+1,\cdot\sim N(\Phi'_tx_t,\Sigma_t)$$

   Integrating out Σ_t, we have

   $$y_t\mid d_t^{\Phi}=d+1\sim t\left(\Phi'_tx_t,\frac{\overline{S}}{\overline{v}+1-N},\overline{v}+1-N\right),$$

   with variance $\overline{S}/(\overline{v}-N-1)$.

   Now we can draw D^{Σ} first and then draw $\Sigma\mid D^{\Sigma}$ by applying the same technique in this paper.

3. Sample $\pi_{\Phi}\mid D^{\Phi}$.

   Assume that the prior $\pi_{\Phi}\sim B(\pi_a^{\Phi},\pi_b^{\Phi})$. Similar to the nonhierarchical model, draw

   $$\pi_{\Phi}\mid D^{\Phi}\sim B(\pi_a^{\Phi}+K_{\Phi}-1,\pi_b^{\Phi}+T-K_{\Phi}),$$

   where K_{Φ} is the total number of change-points of Φ.

4. Sample $\pi_\Sigma \mid D^\Sigma$.

   Assume that the prior $\pi_\Sigma \sim B(\pi_a^\Sigma, \pi_b^\Sigma)$. Similar to the nonhierarchical model, draw

   $$\pi_\Sigma \mid D^\Sigma \sim B(\pi_a^\Sigma + K_\Sigma - 1, \pi_b^\Sigma + T - K_\Sigma),$$

   where K_Σ is the total number of change-points of Σ.

5. Sample $\underline{\Phi}, \underline{\Omega} \mid D^\Phi, D^\Sigma, \Phi, \Sigma$

   The prior is matrix normal and inverse-Wishart.

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

   Suppose there are K regimes and each regime has distinct value Φ_i^*. The conditional posterior $\underline{\mathbf{\Phi}}, \underline{\mathbf{\Omega}} | \{\Phi_i^*\}_{i=1}^K$ is

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

   with

   $$\begin{aligned} \Omega_1 &= \Omega_0 + \sum_{i=1}^K \Phi_i (\Sigma^*)^{-1} \Phi_i' + M_0 \Lambda_0^{-1} M_0' - M_1 \Lambda_1^{-1} M_1' \\ \omega_1 &= \omega_0 + KN \\ M_1 &= (M_0 \Lambda_0^{-1} + \sum_{i=1}^K \Phi_i (\Sigma^*)^{-1}) \Lambda_1 \\ \Lambda_1 &= (\Lambda_0^{-1} + K(\Sigma^*)^{-1})^{-1} \end{aligned}$$

6. Sample $\underline{S}|D^\Phi, D^\Sigma, \underline{\nu}, \Phi, \Sigma$. Same as the hierarchical model.

7. Sample $\underline{\nu}|D^\Phi, D^\Sigma, \underline{S}, \Phi, \Sigma$. Same as the hierarchical model.

G Inflation Persistence

We compute the inflation gap persistence from Cogley et al. (2010) because it may help to understand the effectiveness of the central bank on monetary policy. The multivariate structural break model is not directly applicable to the inflation gap persistence measure of Cogley et al. (2010) because it requires conditional stationarity. To apply the formulae from Cogley et al. (2010), we difference the variables that have unit roots and fit them to the structural break model. Such transformation does not guarantee that the autoregressive coefficients conform to stationarity; but our empirical results show that all the posterior draws imply stationarity. Hence, the formulae in Cogley et al. (2010) are applicable.

Briefly, we start from Equation (9) and focus on its VAR representation by using notations similar to Cogley et al. (2010):

$$z_t = \mu_t + A_t z_{t-1} + e_t,\ e_t \sim N(0, \Sigma_t), \tag{41}$$

where z_t includes current and lagged values of y_t. The vector μ_t is the intercept and A_t represents the autoregressive coefficients.[11]

Conditional on A_t, assuming stationarity and no structural change afterwards, the conditional mean of z_t is denoted as

$$\overline{z}_t = (I - A_t)^{-1}\mu_t \tag{42}$$

The inflation gap can be written as

$$(z_t - \overline{z}_t) = A_t(z_{t-1} - \overline{z}_t) + e_t \tag{43}$$

A h-period ahead forecast of the gap has variance

$$\text{var}_t(\hat{z}_{t+h}) = \sum_{j=0}^{h-1} A_t^j \text{var}(e_t)(A_t^j)' = \sum_{j=0}^{h-1} A_t^j \Sigma_t (A_t^j)' \tag{44}$$

If no previous shocks are observed, the forecast variance of the gap is equivalent to the case when $h \to \infty$:

$$\text{var}(\hat{z}_{t+h}) = \sum_{j=0}^{\infty} A_t^j \Sigma_t (A_t^j)' \tag{45}$$

Cogley et al. (2010) proposed an R^2 type statistic to measure inflation gap persistence. Specifically,

$$R^2_{t,h} = 1 - \frac{\text{var}_t(e_\pi \hat{z}_{t+h})}{\text{var}(e_\pi \hat{z}_{t+h})} = 1 - \frac{e_\pi \left[\sum_{j=0}^{h-1} A_t^j \Sigma_t (A_t^j)'\right] e_\pi'}{e_\pi \left[\sum_{j=0}^{\infty} A_t^j \Sigma_t (A_t^j)'\right] e_\pi'}, \tag{46}$$

where e_π is a selection vector for inflation gap. When h increases, a rapid decreasing of $R^2_{t,h}$ towards zero means that most of the variation will be explained by the most recent shocks. This scenario implies weak persistence. If the $R^2_{t,h}$ converges to zero slowly, it shows strong persistence. A detailed description can be found in Cogley et al. (2010).

[11] We illustrate with SB-VAR(1) as an example. Equation (9) can be written as

$$y_t = \mu_t + A_t y_{t-1} + e_t; \quad e_t \sim N(0, \Sigma_t),$$

with $\Phi_t' = (\mu_t, A_t)$. We draw Φ_t from the posterior without imposing stationarity restrictions. Hence, no modification of the original code is needed. For each random draw of Φ_t, we find the corresponding matrix A_t and check if it implied stationarity (the modulus of all eigen values are less than 1). If stationarity is satisfied, apply Equation (46). Otherwise, discard it. In out application, all posteriors are consistent with stationarity.

References

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

Baumeister, Christiane and Kilian, Lutz. Forecasting the real price of oil in a changing world: A forecast combination approach. *Journal of Business & Economic Statistics*, 33 (3):338–351, 2015.

Baumeister, Christiane and Peersman, Gert. The role of time-varying price elasticities in accounting for volatility changes in the crude oil market. *Journal of Applied Econometrics*, 28(7):1087–1109, 2013.

Bauwens, Luc, Carpantier, Jean-Franois, and Dufays, Arnaud. Autoregressive moving average infinite hidden markov-switching models. *Journal of Business & Economic Statistics*, 35(2):162–182, 2017.

Belmonte, M., Koop, G., and Korobilis, D. Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94, 2014.

Carriero, Andrea, Clark, Todd E, and Marcellino, Massimiliano. Bayesian vars: specification choices and forecast accuracy. *Journal of Applied Econometrics*, 30(1):46–73, 2015.

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.

Cogley, Timothy, Primiceri, Giorgio E, and Sargent, Thomas J. Inflation-gap persistence in the us. *American Economic Journal: Macroeconomics*, 2(1):43–69, 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.

Guo, Hui and Kliesen, Kevin L. Oil price volatility and us macroeconomic activity. *FEDERAL RESERVE BANK OF SAINT LOUIS REVIEW*, 87(6):669, 2005.

Hamilton, James D. Causes and consequences of the oil shock of 2007-08. *Brookings Papers on Economic Activity*, Spring:215–259, 2009.

Hamilton, James D. Nonlinearities and the macroeconomic effects of oil prices. *Macroeconomic Dynamics*, 15:364–378, 2011.

Herbst, Edward and Schorfheide, Frank. Sequential monte carlo sampling for dsge models. *Journal of Applied Econometrics*, 29(7):1073–1098, 2014.

Hooker, Mark A. What happened to the oil price-macroeconomy relationship? *Journal of monetary Economics*, 38(2):195–213, 1996.

Jochmann, Markus and Koop, Gary. Regime-switching cointegration. *Studies in Nonlinear Dynamics and Econometrics*, 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.

Kilian, Lutz. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. *American Economic Review*, 99(3):1053–1069, 2009.

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

Koop, Gary, Leon-Gonzalez, Roberto, and Strachan, Rodney W. Bayesian inference in a time varying cointegration model. *Journal of Econometrics*, 165(2):210–220, 2011.

Krueger, Fabian. bvarsv: Bayesian analysis of a vector autoregressive model with stochastic volatility and time-varying parameters. CRAN, https://CRAN.R-project.org/package=bvarsv, 2015.

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

Maheu, John M. and Song, Yong. A new structural break model, with an application to canadian inflation forecasting. *International Journal of Forecasting*, 30(1):144 – 160, 2014.

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.

Pettenuzzo, Davide and Timmermann, Allan. Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics*, 164(1):60 – 78, 2011.

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.

Table 1: Oil and Real GDP: Log Long-run Predictive Likelihood[a]

	log ML	Log PL			
		$h = 1$	$h = 3$	$h = 6$	$h = 12$
VAR(3)	-841.1	-637.0	-646.3	-607.8	-578.6
VAR(3)-VDGARCH	-836.0	-625.9	-626.6	-601.2	-571.2
SB-VAR(2)	-818.6	**-611.1**	-626.2	-610.4	-579.9
TVP-VAR(1)-SV	-	-623.7	-663.8	-613.7	-580.8
H-SB-U(1)	**-805.9**	-612.4	-	-	-
H-SB-VAR(3) (ind)	-833.5	-631.5	-634.3	-612.6	-578.7
H-SB-VAR(3)	-807.9	-611.6	**-622.8**	**-597.4**	**-567.9**

[a] The notation h is the forecast horizon. For each model, we only report the results for the lag that has the largest predictive likelihood. log ML denotes the log-marginal likelihood. The log-predictive likelihoods (log PL) are calculated from the last 30 years of data (120 observations). VAR-VDGARCH is the model in (36)-(37), SB-VAR is the non-hierarchical structural break model (10)-(12), TVP-VAR(1)-SV is the model in (38)-(40), H-SB-U(1) is the univariate structural break model in (35) run on each dependent variable independently, H-SB-VAR(3) is the hierarchical structural break model while (ind) denotes the same model with independent breaks in the regression parameters and the covariance matrix (21)-(25). A bold entry denotes the maximum value in the respective column.

Table 2: Oil and Real GDP: Long-run Forecasts[a]

	RMSFE			
Oil	$h=1$	$h=3$	$h=6$	$h=12$
VAR(3)(benchmark)	**15.7**	16.7	**15.5**	**15.8**
VAR(3)-VDGARCH	15.8	17.0	15.7	16.5
SB-VAR(2)	17.0	18.9	17.9	18.1
TVP-VAR(1)-SV	16.3	17.4	16.8	32.9
H-SB-U(1)	16.9	-	-	-
H-SB-VAR(3) (ind)	17.2	17.6	15.9	16.0
H-SB-VAR(3)	15.9	**15.6**	**15.5**	16.0
GDP	$h=1$	$h=3$	$h=6$	$h=12$
VAR(3)(benchmark)	0.59	0.62	**0.62**	0.64
VAR(3)-VDGARCH	0.55	**0.61**	0.64	0.66
SB-VAR(2)	0.58	0.64	0.70	0.76
TVP-VAR(1)-SV	0.59	0.62	0.65	0.81
H-SB-U(1)	0.57	-	-	-
H-SB-VAR(3) (ind)	**0.56**	0.63	0.63	0.64
H-SB-VAR(3)	0.58	**0.61**	**0.62**	**0.63**

[a] RMSFE is the root mean square forecast error based in the predictive mean from a model calculated from the last 30 years of data (120 observations). The notation h is the forecast horizon. For each model, we only report the results for the lag that has the largest predictive likelihood. See Table 1 for the list of model labels. A bold entry denotes the minimum value of the respective column.

Table 3: 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

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

	log-PL	UR	Core PCE	Nonfarm Em	Retail	Housing	IP	FFR
VAR(4) (benchmark)	-1751.8	0.144	**1.530**	1.568	14.210	8.007	8.789	0.206
SB-VAR(3)	-1695.6	0.153	2.300	1.230	18.123	7.298	10.152	0.205
TVP-VAR(2)-SV	-1584.5	0.143	1.593	1.211	14.150	7.810	**8.438**	0.171
H-SB-U(3)	**-1551.8**	**0.140**	1.931	**1.165**	16.238	**7.283**	11.076	**0.164**
VAR(3)-VDGARCH	-1701.9	0.152	1.718	1.782	**14.090**	7.779	8.732	0.201
H-SB-VAR(1) (ind)	-1739.0	0.147	3.530	1.479	21.754	7.313	11.110	0.203
H-SB-VAR(1)	-1643.3	0.152	1.667	1.772	14.246	7.368	8.931	0.190

The last 10 years (120 observations) of data are used in forecasting. The second column is the log-predictive likelihood (log PL). The remaining columns report the root mean square forecast error (RMSFE) for each series. We only report the results for the lag that has the largest log-predictive likelihood (log PL). See Table 1 for the list of model labels. A bold entry in the second column denotes the maximum value while elsewhere it denotes a minimum value of the respective column.

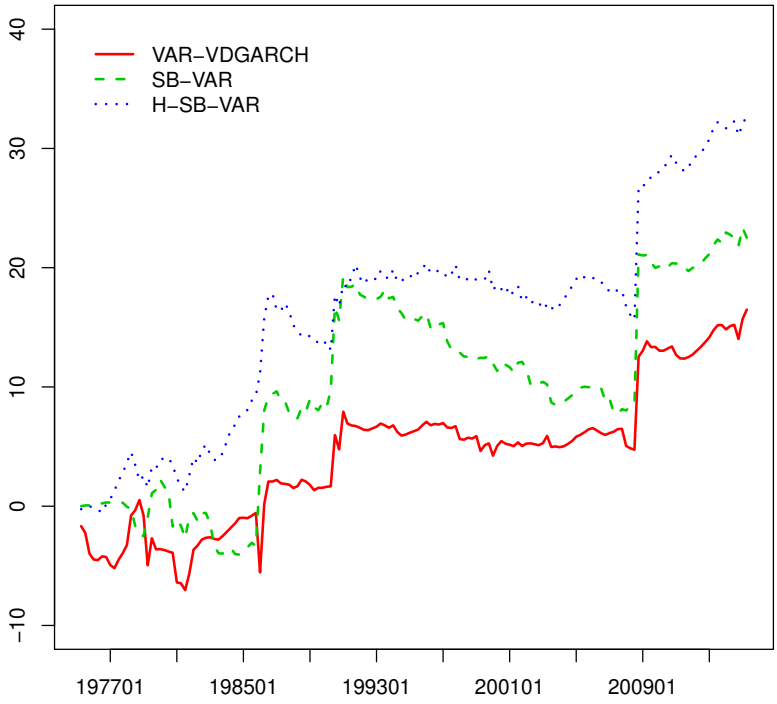

Figure 1: Oil and GDP: Difference in cumulative log-predictive likelihoods against the VAR(3) model

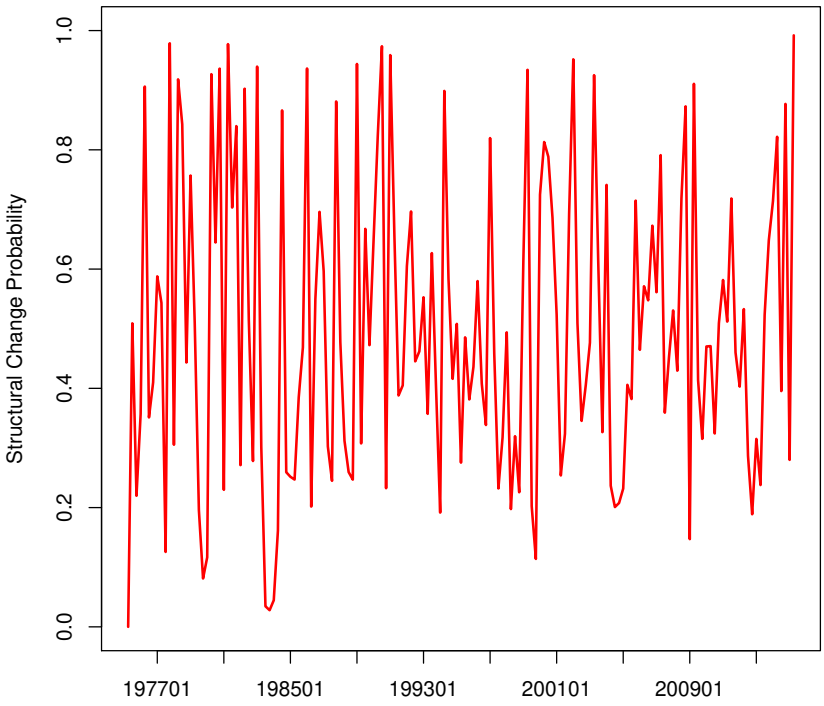

Figure 2: Oil and GDP: Posterior probability of structural change, H-SB-VAR, $P(d_t = 1|Y_{1,T})$.

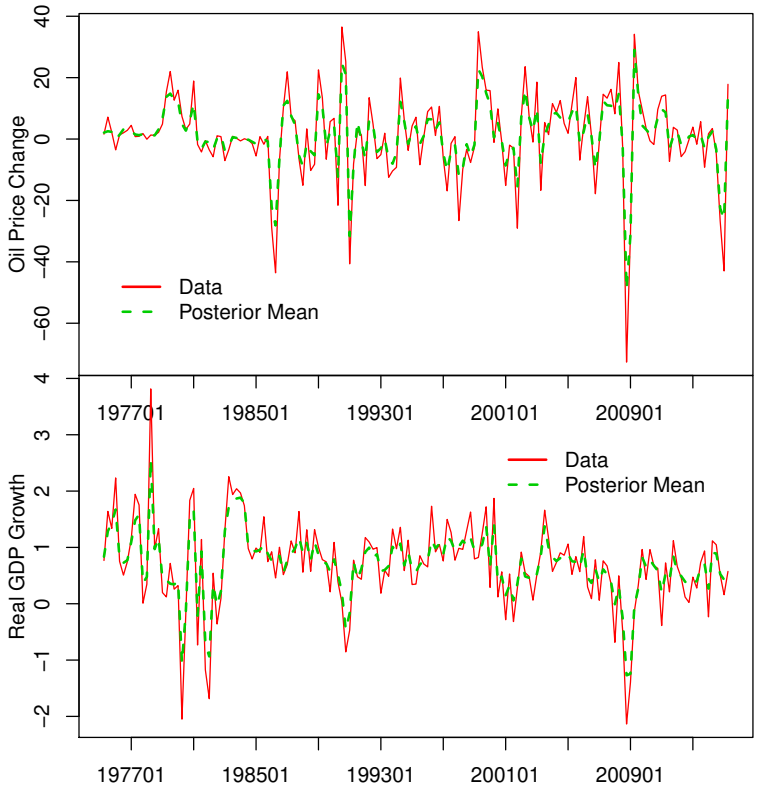

Figure 3: Oil and GDP: Posterior Mean from H-SB-VAR

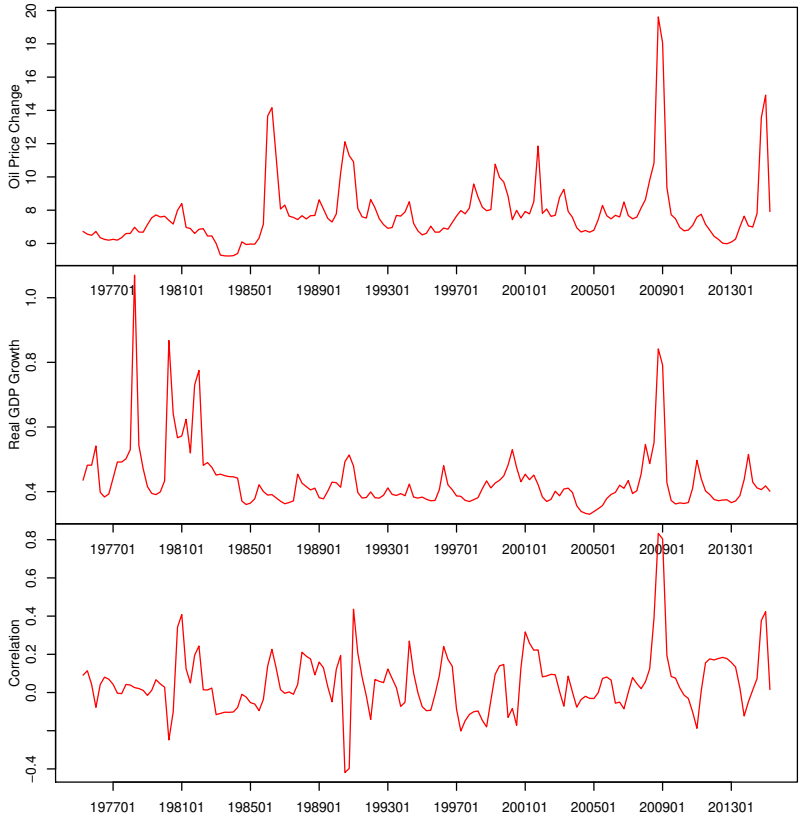

Figure 4: Oil and GDP: Posterior mean of standard deviation and correlation from error terms from H-SB-VAR

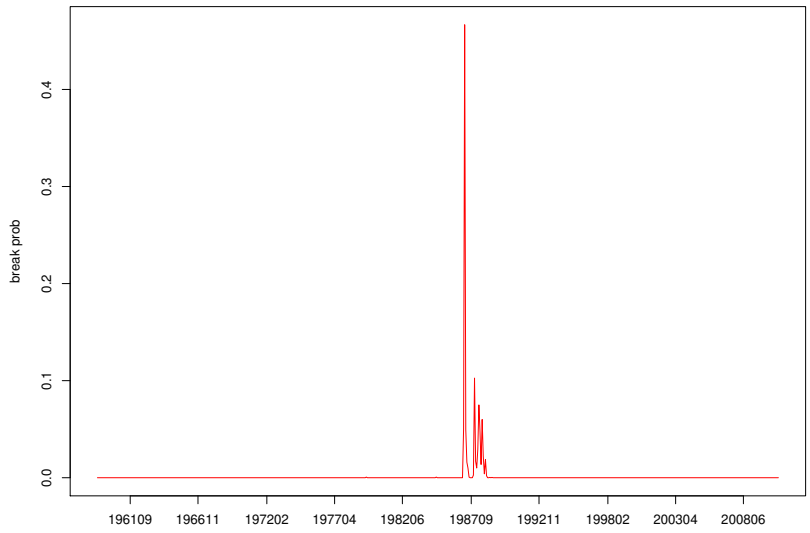

Figure 5: 7-variable SB-VAR, non-hierarchical model: break probability, $P(d_t = 1|Y_{1,T})$

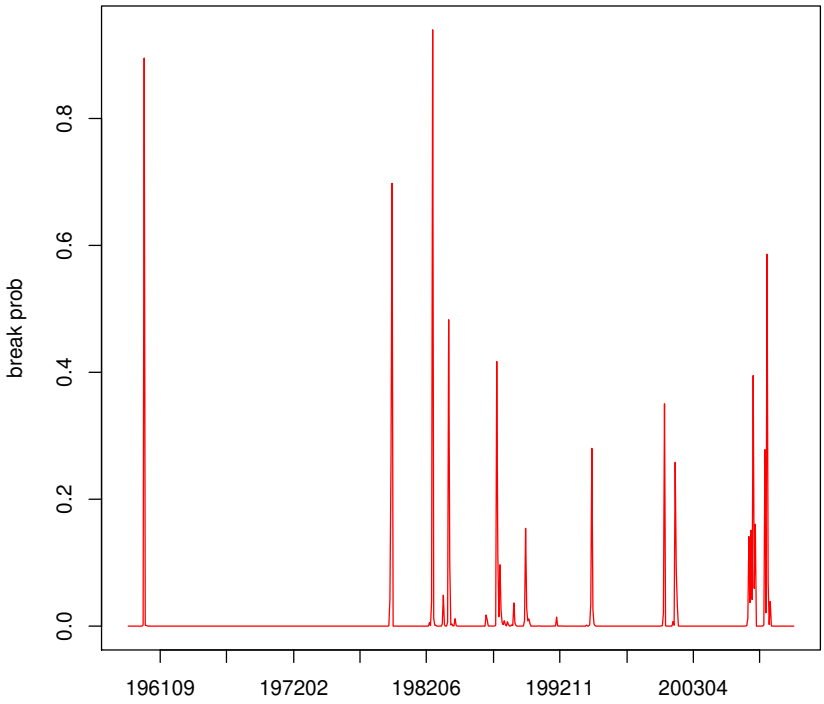

Figure 6: 7-variable H-SB-VAR(1), hierarchical model: break probability, $P(d_t = 1|Y_{1,T})$

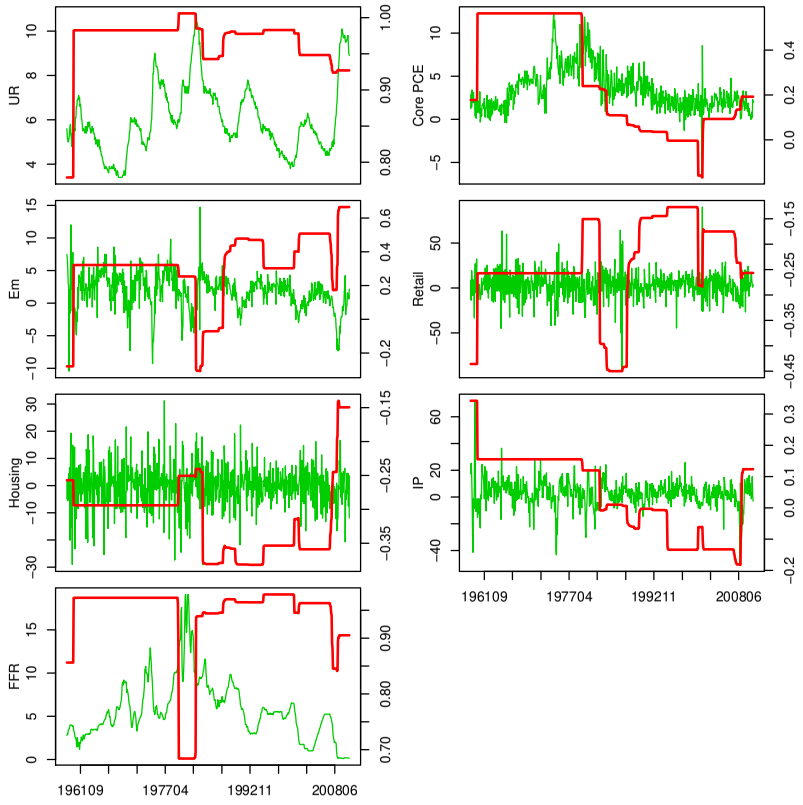

Figure 7: 7-variable H-SB-VAR(1), hierarchical model: red is the persistence parameter, green is the data

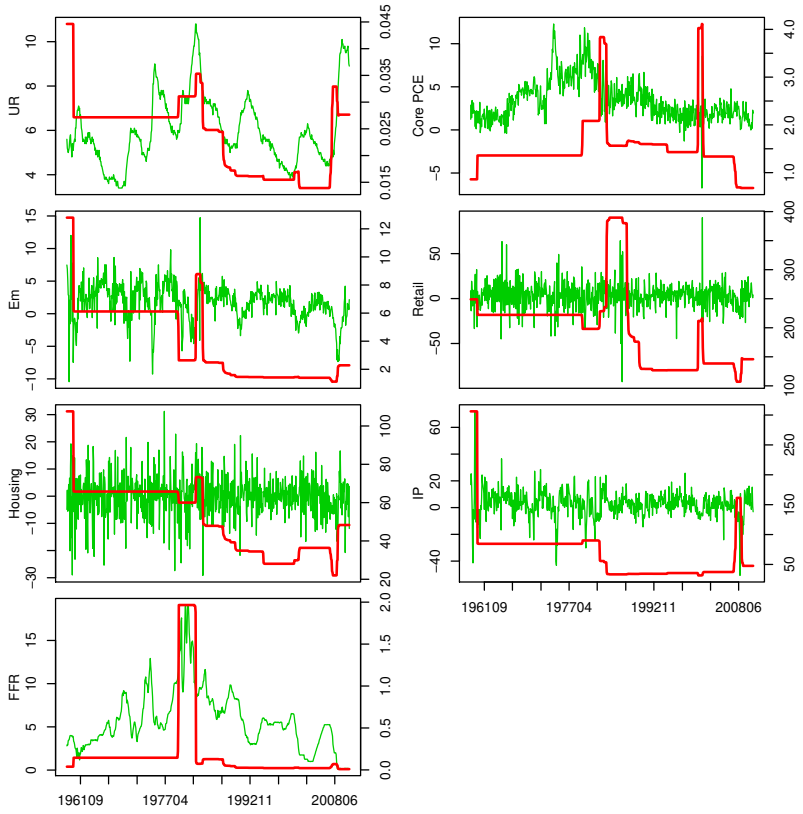

Figure 8: 7-variable H-SB-VAR(1), hierarchical model: red is the volatility, green is the data

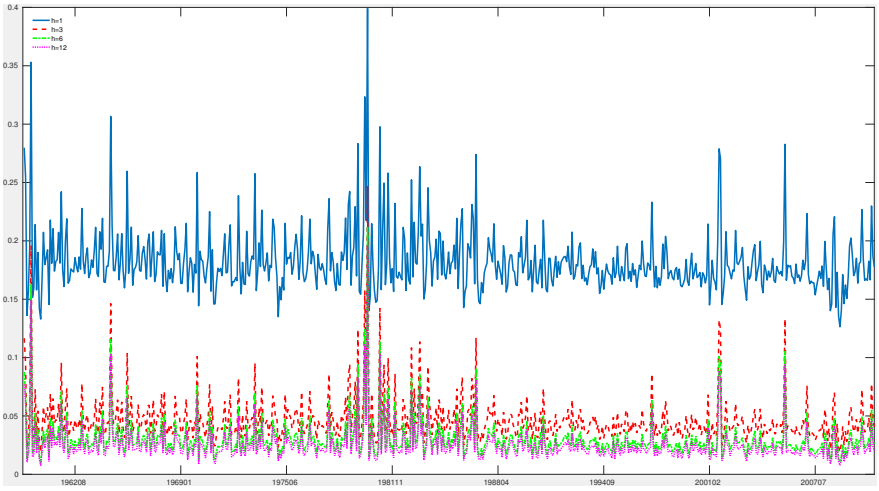

Figure 9: Inflation Persistence Measure of Cogley et al. (2010)

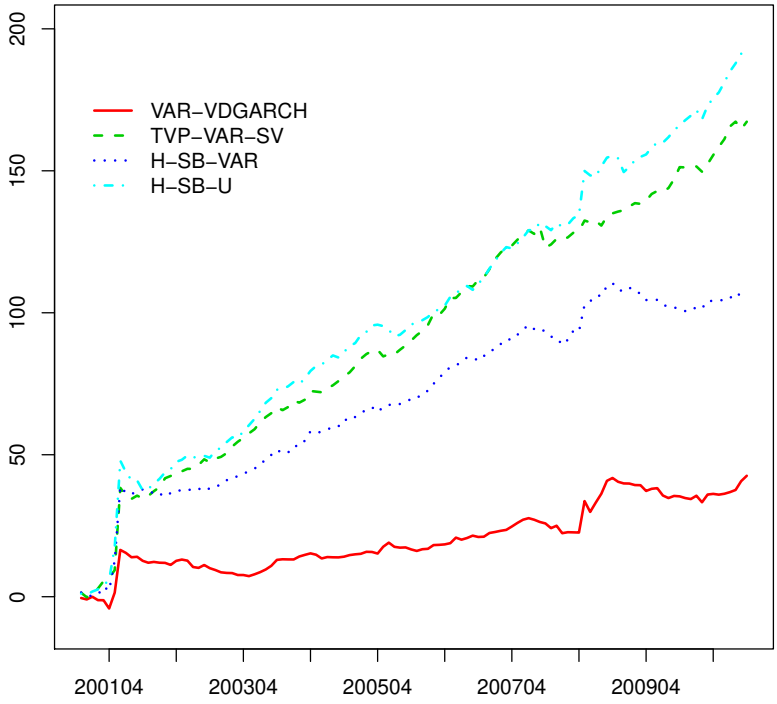

Figure 10: Difference in cumulative log-predictive likelihoods against the VAR(4) model

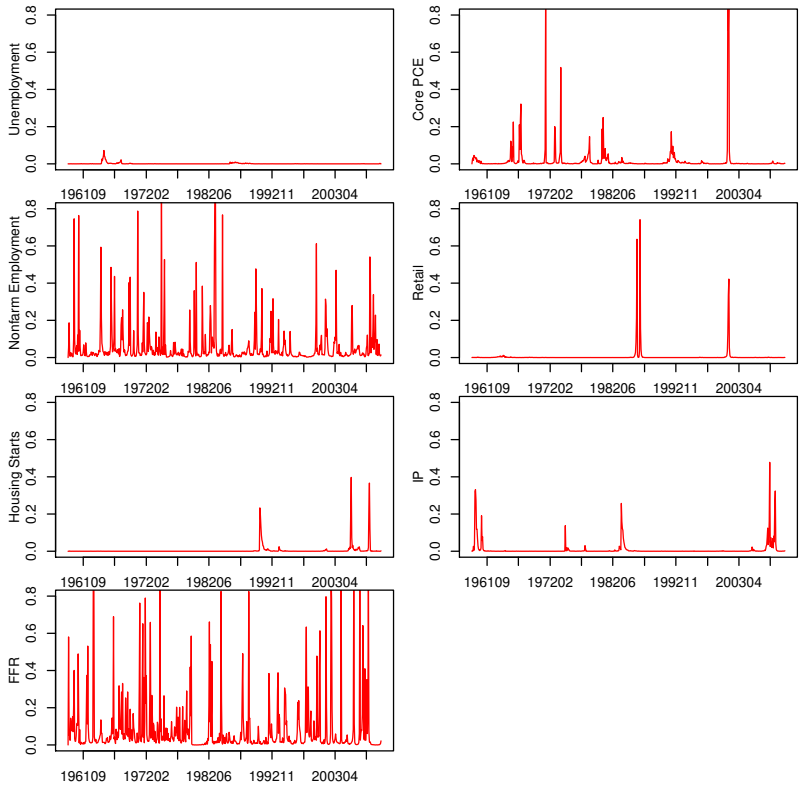

Figure 11: Posterior probability of a break from each univariate model (H-SB-U(3))