Journal of Econometrics 146 (2008) 255–274

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Methods for inference in large multiple-equation Markov-switching models

Christopher A. Sims [a], Daniel F. Waggoner [b], Tao Zha [b,c,*]

[a] Princeton University, United States
[b] Federal Reserve Bank of Atlanta, United States
[c] Emory University, United States

ARTICLE INFO

Article history:
Available online 9 September 2008

JEL classification:
C32
C52
E4

Keywords:
Density overlap
New MHM
Incremental and discontinuous changes
Composite Markov process
Integrated-out likelihood

ABSTRACT

Inference for multiple-equation Markov-chain models raises a number of difficulties that are unlikely to appear in smaller models. Our framework allows for many regimes in the transition matrix, without letting the number of free parameters grow as the square as the number of regimes, but also without losing a convenient form for the posterior distribution. Calculation of marginal data densities is difficult in these high-dimensional models. This paper gives methods to overcome these difficulties, and explains why existing methods are unreliable. It makes suggestions for maximizing posterior density and initiating MCMC simulations that provide robustness against the complex likelihood shape.

1. Introduction

This paper extends the methods of Hamilton (1989), Chib (1996), and Kim and Nelson (1999) to large Markov-switching multiple-equation models. In such large models, a variety of modelling choices, not needed in single-equation models, are required to control dimensionality. We provide a general framework for keeping these models tractable, develop a new procedure to implement the modified harmonic means (MHM) method for achieving accuracy of the estimated marginal data density, and supply a general-purpose software package to make estimation and inference of large Markov-switching models computationally feasible.

This paper considers a large class of restrictions on the parameters in the transition matrix. Under certain conditions, this class maintains a standard posterior density form for the free parameters in the transition matrix. Although one could directly derive and code up the posterior density function case by case, we propose a general-purpose interface that is straightforward for researchers to automate potentially complex restrictions by simply expressing them in convenient matrix form. We show how such an interface matrix can be formed in the context of a variety of models and discuss how our framework is related to these models.

In the macroeconomic literature, there are two potential problems with Bayesian analysis. First, if the Markov-Chain Monte-Carlo (MCMC) algorithm begins with an arbitrary starting point without searching for the maximum likelihood estimate (MLE) or the posterior estimate at the peak of the posterior density function, this starting point may turn out to be in the very low probability region and the posterior draws simulated from the MCMC algorithm may get stuck in this region. We develop an efficient blockwise optimization method designed to find the MLE or the posterior mode for a complicated or large dynamic model.

The second problem is associated with the standard MHM method in which the variance of the weighting function can be arbitrarily scaled. Changes in the variance of the weighting function can cause the estimate of the marginal data density to fluctuate drastically; the standard approach of truncating the tail of the weighting distribution doest not, in general, prevent this fluctuation. Our new way of implementing the MHM method is designed to deal with this uncertainty explicitly. We show that such uncertainty can be significantly reduced by explicitly calculating the degree of overlap between the weighting function and the posterior density.

When one evaluates the marginal data density using the standard MHM method, a typical choice of the weighting function is a Gaussian density constructed from the first two sample moments of the posterior distribution. If the posterior distribution is very non-Gaussian, however, such a weighting function can be a poor approximation. We propose a more general weighting function that aims at dealing with the non-Gaussian shape of the

* Corresponding address: Federal Reserve Bank of Atlanta, 1000 Peachtree Street, N.E., Atlanta, GA 30309, United States. Tel.: +1 404 498 8353; fax: +1 404 498 8956.
E-mail address: tzha@earthlink.net (T. Zha).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.08.023

posterior distribution. This kind of weighting function includes a Gaussian density as a special case and proves to work well for high-dimensional models such as vector autoregressions (VARs).

The rest of the paper is organized as follows.

Section 2 proposes a general framework for large Markov-switching models with a variety of restrictions imposed on the transition matrix. Section 3 shows how our framework encompasses a large set of regime-switching models and discusses the advantage of our framework in the context of existing time-varying models.

Section 4 presents the prior and discusses the importance of an informative prior and the issues related to training samples and Bayesian information criterion (BIC). Under some regularity conditions, the likelihood function and the posterior distribution are derived in Section 5.

In Section 6, we proposes a blockwise optimization method for finding the posterior mode. The method proves, in large or complicated Markov-switching models, to be computationally more efficient than the Monte Carlo expectation-maximization (EM) algorithm, which has been widely used in similar, but smaller, models.

In Section 7, we construct a variation on the MHM method that works much better, offer a practical way of gauging how much the weighting function and the posterior density overlap, and discuss the importance of being able to evaluate the overall likelihood.

Section 8 applies our general framework to Markov-switching VAR models. We illustrate how our computer software makes it feasible to fit a large set of empirical models to the post-war US data. We use an empirical model to show that the standard MHM runs into severe difficulties and as a result may lead to an erroneous estimate of the marginal data density.

And Section 9 concludes.

2. General Markov-switching framework

2.1. Distributional assumptions

Let $(Y_t, Z_t, \theta, Q, S_t)$ be a collection of random variables where

$$Y_t = (y_1, \ldots, y_t) \in \left(\mathbb{R}^n\right)^t,$$
$$Z_t = (z_1, \ldots, z_t) \in \left(\mathbb{R}^m\right)^t,$$
$$\theta = (\theta_i)_{i\in H} \in \left(\mathbb{R}^r\right)^h,$$
$$Q = \left(q_{i,j}\right)_{(i,j)\in H\times H} \in \mathbb{R}^{h^2},$$
$$S_t = (s_0, \ldots, s_t) \in H^{t+1},$$
$$S_{t+1}^T = (s_{t+1}, \ldots, s_T) \in H^{T-t},$$

and H is a finite set with h elements and is usually taken to be the set $\{1, \ldots, h\}$. The object y_t represents an $n \times 1$ vector of endogenous variables and z_t represents an m vector of exogenous variables. Thus, our analysis encompasses a special case in which there are no exogenous variables. The matrix Q is a Markov transition matrix and $q_{i,j}$ is the probability that s_t is equal to i given that s_{t-1} is equal to j. The matrix Q is restricted to satisfy

$$q_{i,j} \geq 0 \quad \text{and} \quad \sum_{i\in H} q_{i,j} = 1.$$

For $1 \leq j \leq h$, let q_j be the jth column of Q and q be an h^2-dimensional column vector stacking these q_j's. The objects θ and q are vectors of parameters, Y_t and Z_t are observed data, and S_t can be considered as either a sequence of latent variables or a vector of nuisance parameters. We assume that $(Y_t, Z_t, \theta, q, S_t)$ has a joint density function $p(Y_t, Z_t, \theta, q, S_t)$, where we use the Lebesgue measure[1] on $(\mathbb{R}^n)^t \times (\mathbb{R}^m)^t \times (\mathbb{R}^r)^h \times \mathbb{R}^{h^2}$ and the counting measure on H^{t+1}. This density satisfies the following conditions.

Condition 1.

$$p(s_t \mid Y_{t-1}, Z_{t-1}, \theta, q, S_{t-1}) = q_{s_t, s_{t-1}}, \quad \text{for } t > 0.$$

Condition 2.

$$p(z_t \mid Y_{t-1}, Z_{t-1}, \theta, q, S_t) = p(z_t \mid Y_{t-1}, Z_{t-1}), \quad \text{for } t > 0.$$

Condition 3.

$$p(y_t \mid Y_{t-1}, Z_t, \theta, q, S_t) = p\left(y_t \mid Y_{t-1}, Z_t, \theta_{s_t}, q, s_t\right), \quad \text{for } t > 0.$$

Condition 1 states formally that the sequence S_t evolves according to an exogenous Markov process with the transition matrix Q. Condition 2 states that z_t is a predetermined variable. Condition 3 asserts that the model for y_t conditional on the past depends only on the current value of the state, not on lagged values of it. This condition is needed to make possible the backward recursion discussed in Section 5; this condition also makes it feasible to integrate out all the regimes S_T for obtaining the likelihood $p(Y_T \mid Z_T, \theta, q)$.

2.2. Restrictions on Q

An important part of our general framework is to encompass a wide range of restrictions on Q, while maintaining the standard form of its posterior probability density function. Suppose Q is unrestricted and the following condition is satisfied.

Condition 4.

$$p(y_t \mid Y_{t-1}, Z_t, \theta, q, s_t) = p(y_t \mid Y_{t-1}, Z_t, \theta, s_t).$$

Then the density of q_j conditional on (Y_T, Z_T, θ, S_T) is of the Dirichlet form, if the prior on q_j is of the Dirichlet form and the initial distribution on s_0 does not depend on q. All the restrictions on q studied in this paper preserve the Dirichlet form of the posterior distribution of q conditional on other parameters of the model, when Condition 4 holds. In Section 5 we will discuss the situation in which Condition 4 does not hold.

For $1 \leq j \leq v$, let w_j be a d_j-dimensional vector, where v may be greater than h (although it is less than or equal to h in most applications) and the elements of w_j are non-negative and sum to one. Let w be a d-dimensional column vector obtained by stacking w_j's, where $d = \sum_{j=1}^{v} d_j$. The restrictions on q are represented by

$$q = Mw, \tag{1}$$

where M is an $h^2 \times d$ matrix such that

$$M = \begin{bmatrix} M_{1,1} & \cdots & M_{1,v} \\ \vdots & \ddots & \vdots \\ M_{h,1} & \cdots & M_{h,v} \end{bmatrix}.$$

Denote an $h \times 1$ vector of ones by 1_h. The submatrix $M_{i,j}$ is an $h \times d_j$ matrix and satisfies the following two conditions:

Condition 5. For each (i, j), all the elements of $M_{i,j}$ are non-negative and $1_h' M_{i,j} = \lambda_{i,j} 1_{d_j}'$, where $\lambda_{i,j}$ is the sum of the elements in any column of $M_{i,j}$.

Condition 6. Each row of M has at most one non-zero element.

Condition 5 is necessary to ensure that the elements of q_j are positive and sum to one. Condition 6 ensures that the likelihood as a function of w_j has the Dirichlet density form. It follows from these conditions that one may assume without loss of generality

[1] Instead of the Lebesgue measure, any sigma-finite measure on $\mathbb{R}^n$ and $\mathbb{R}^m$ can be used as long as the product measure is used on $(\mathbb{R}^n)^t$ and $(\mathbb{R}^m)^t$.

that $d_j \leq h$ and $d \leq h^2$. Our class of restrictions on Q encompasses a wide range of models discussed in the literature.

Working directly on the transition matrix Q that satisfies the restrictions specified by (1), without explicitly constructing the transformation matrix M in the manner of Conditions 5 and 6, is conceptually feasible but practically difficult. In particular, if restrictions are complicated and the researcher does not wish to derive and code up the posterior density of free elements in the transition matrix each time when a new application is studied, the setup represented by (1) provides an efficient way to automate the handling of different kinds of restrictions in one convenient, general framework and to eliminate potential mathematical and programming errors that may occur for each new application. When the researcher chooses to use our computer program, moreover, the general-purpose interface matrix M in (1) as one of inputs for the program becomes very handy and easy to implement.[2] From (1) it is clear that q or Q is known once w is given. For the rest of the paper, therefore, we will focus on the free parameter vector w.

In the next section we illustrate how to construct the transformation matrix M for a wide class of regime-switching models. Most of the examples are used to show how to keep the number of free parameters in the transition matrix from growing too fast as the number of regimes increases.

3. A class of regime-switching models

In this section we show that the framework presented in the last section is flexible enough to encompass a variety of regime-switching models by representing various types of restrictions on the transition matrix in our analytical framework, which also serves as an interface for the user of our software package. We discuss these models in comparison with other time-varying models studied in the literature and explicate why we prefer our framework.

3.1. Structural breaks

By splitting the sample into two subsamples, Clarida et al. (2000) and Lubik and Schorfheide (2004) find that US monetary policy has switched regime, once for all, since early 1980. One can improve their sample-splitting method by modelling the structural break in our Markov-switching framework in which there is a probability that monetary policy in the first part of the sample switches to an irreversible regime in the second part of the sample, while other parts of the economy remain constant.[3] The transition matrix for this regime shift can be written as

$$\begin{bmatrix} q_{1,1} & 0 \\ q_{2,1} & 1 \end{bmatrix}.$$

Using the form (1), the restrictions can be expressed as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad M_{2,2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and $M_{i,j} = 0$ for $i \neq j$, where $v = 2$, $d_1 = 2$, and $d_2 = 1$.

In one of the structural models in Sims and Zha (2006), they find that monetary policy regime that has prevailed in the 1990s and 2000s was also dominant in most of the 1960s and in some parts of the 1970s. They treat this regime to be recurrent. If one believes that this policy regime would last indefinitely from now on (the belief we do not share), how does one specify such regime changes in our framework? In this situation, we have two policy regimes: hawkish and dovish in response to inflation. The hawkish regime is recurrent in the first part of the sample and then become irreversible in the second part of the sample. We can write the transition matrix as

$$Q = \begin{bmatrix} q_{1,1} & q_{1,2} & 0 \\ q_{2,1} & q_{2,2} & 0 \\ 0 & q_{3,2} & 1 \end{bmatrix},$$

with a further restriction that the parameters in the policy equation in the first regime is the same as those in the third regime. Thus, the first and third regime represent the hawkish policy, while the second regime represents the dovish policy. In our framework, the restrictions on Q can be expressed as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad M_{2,2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad M_{3,3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and $M_{i,j} = 0$ for $i \neq j$, where $v = 3$, $d_1 = 2$, $d_2 = 3$, and $d_3 = 1$.

In the literature on reduced-form statistical models, structural breaks are sometimes modeled as multiple change points where s_t can either remain at the current regime or switch to the next higher value (Chib, 1998). Because s_t is not allowed to switch back to the previous lower value, the changing-point model precludes the case of recurrent regimes as previously discussed. This one-step ahead transition matrix is represented as

$$Q = \begin{bmatrix} q_{1,1} & 0 & \cdots & 0 & 0 \\ q_{2,1} & q_{2,2} & \cdots & 0 & 0 \\ 0 & q_{3,2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \vdots & q_{h-1,h-1} & 0 \\ 0 & 0 & \vdots & q_{h,h-1} & 1 \end{bmatrix}.$$

In our framework, these exclusion restrictions imposed on Q can be expressed as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad M_{2,2} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \ldots,$$

$$M_{h-1,h-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad M_{h,h} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

and $M_{i,j} = 0$ for $i \neq j$, where $v = h$, $d_1 = \cdots = d_{h-1} = 2$, and $d_h = 1$.

This specification does not impose that all the regimes allowed for in the transition matrix actually occur in the sample. Since we use MCMC simulation of the posterior for inference, and draw sequences of regimes in the simulation, the number of regimes that actually occur in the sample is a posteriori uncertain, and we can easily tabulate its posterior distribution. This is the approach of Chopin and Pelgrin (2004), though they propose a specific approach to parameterization of the model that does not match what we suggest here. Note that one can tabulate a distribution of the number of regimes that actually occurred in the sample from

[2] The software is available at http://home.earthlink.net/tzha02/ProgramCode/programCode.html. In Appendix C, we illustrate a concrete example of how to use the interface with our software program.

[3] Our method also improves on the approach of Beyer and Farmer (2004), who do not treat breaks stochastically.

the MCMC simulations with any specification of Q, though when the regimes are few and recurrent the distribution is likely to be nearly degenerate.[4]

Koop and Potter (in press) have another approach to change points, not directly based on a hidden Markov chain. They observe that models that postulate a distribution of a fixed number of multiple change points across the sample can imply implausibly high probabilities of change at the beginning or end of the sample, conditional on the number that have occurred in the rest of the sample. This problem does not arise in Markov-switching setups like ours. They also suggest that it may sometimes be natural to have the probability of a switch depend on the time since the last shift, which a Markov-switching framework like ours cannot implement.

3.2. Incremental and discontinuous shifts

The parameter-drift and stochastic-volatility model studied by Cogley and Sargent (2005) captures continuously incremental changes in the model parameters.[5] Such incremental changes can be approximated arbitrarily well by expanding the number of regimes (Tauchen, 1986). Our approach has advantage over that of Cogley and Sargent (2005) because it allows for occasional discontinuous shifts in regime as well as frequent, incremental changes in parameters, while keeping the number of free parameters in the transition matrix in a much smaller dimension.[6] One way to achieve this objective is to concentrate weight on the diagonal of Q (Zha, 2008). Specifically, one can express incremental changes and discontinuous jumps among h regimes as

$$Q = \begin{bmatrix} \pi_1 & \beta_2\alpha_2(1-\pi_2) & \dots & \beta_h\alpha_h^{h-1}(1-\pi_h) \\ \beta_1\alpha_1(1-\pi_1) & \pi_2 & \dots & \beta_h\alpha_h^{h-2}(1-\pi_h) \\ \beta_1\alpha_1^2(1-\pi_1) & \beta_2\alpha_2(1-\pi_2) & \dots & \beta_h\alpha_h^{h-3}(1-\pi_h) \\ \dots & \dots & \dots & \dots \\ \beta_1\alpha_1^{h-1}(1-\pi_1) & \beta_2\alpha_2^{h-2}(1-\pi_2) & \dots & \pi_h \end{bmatrix},$$

where the free parameter π_i is to be estimated and the hyperparameters $0 < \alpha_i < 1$ and β_i are taken as given. The restrictions can be written as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & \beta_1\alpha_1 \\ 0 & \beta_1\alpha_1^2 \\ \dots & \dots \\ 0 & \beta_1\alpha_1^{h-1} \end{bmatrix}, \quad M_{2,2} = \begin{bmatrix} 0 & \beta_2\alpha_2 \\ 1 & 0 \\ 0 & \beta_2\alpha_2 \\ \dots & \dots \\ 0 & \beta_2\alpha_2^{h-2} \end{bmatrix}, \dots,$$

$$M_{n+1,n+1} = \begin{bmatrix} 0 & \beta_h\alpha_h^{h-1} \\ 0 & \beta_h\alpha_h^{h-2} \\ 0 & \beta_h\alpha_h^{h-3} \\ \dots & \dots \\ 1 & 0 \end{bmatrix},$$

where the value of α_i controls the speed of decay and the value of β_i is so chosen that elements in the second column of $M_{i,i}$ sum to 1. Note that $v = h, d_1 = \cdots = d_h = 2$, and $M_{i,j} = 0$ for $i \neq j$.

The above example shows that we can reduce a large number of elements in the transition matrix to a handful of free parameters whose dimension is equal to the number of regimes. The empirical results of Cogley and Sargent (2005) show that the dimension of parameters that change significantly can be extremely small. In our framework, the class of restrictions specified in (1) enables us to keep the number of free parameters fixed while expanding the number of regimes. Consider an $h \times h$ transition matrix Q in the form of

$$\begin{bmatrix} a & b/2 & \cdots & 0 & 0 \\ b & a & \ddots & \vdots & \vdots \\ 0 & b/2 & \ddots & b/2 & 0 \\ \vdots & \vdots & \ddots & a & b \\ 0 & 0 & \cdots & b/2 & a \end{bmatrix},$$

where $a + b = 1$. This restricted transition matrix implies that when we are in regime j, the probability of moving to regime $j-1$ or $j+1$ is symmetric and independent of j. Let $v = 1$ and $d_1 = 2$. We can express this restriction as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}, \quad M_{h,1} = \begin{bmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and for $1 < i < h$, the $h \times 2$ matrix $M_{i,1}$ is zero except for a submatrix centered at the ith row that has the form

$$\begin{bmatrix} 0 & 1/2 \\ 1 & 0 \\ 0 & 1/2 \end{bmatrix}.$$

In general, our framework is flexible enough to handle more elaborate cases where the jumping probabilities are not symmetric or independent or where the regime jumps to nearby (but not adjacent) regimes.

In practice, it is sometimes found that the data do not favor a large number of regimes for dynamic macroeconomic models. When the number of regimes is small, we follow Sims's (2001) approach to parsimonious parametrization of Q by introducing symmetric jumping among adjacent regimes. In the case of four regimes, for example, the transition matrix is restricted as

$$Q = \begin{bmatrix} \pi_1 & (1-\pi_2)/2 & 0 & 0 \\ 1-\pi_1 & \pi_2 & (1-\pi_3)/2 & 0 \\ 0 & (1-\pi_2)/2 & \pi_3 & 1-\pi_4 \\ 0 & 0 & (1-\pi_3)/2 & \pi_4 \end{bmatrix} \tag{2}$$

where π_1, π_2, π_3, and π_4 are free parameters to be estimated. These restrictions can be expressed as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad M_{2,2} = \begin{bmatrix} 0 & 1/2 \\ 1 & 0 \\ 0 & 1/2 \\ 0 & 0 \end{bmatrix},$$

$$M_{3,3} = \begin{bmatrix} 0 & 0 \\ 0 & 1/2 \\ 1 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad M_{4,4} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and $M_{i,j} = 0$ for $i \neq j$, where $v = 4$ and $d_1 = d_2 = d_3 = d_4 = 2$. Our experience indicates that the data tends to favor this restricted transition matrix than the unrestricted version.

[4] It is worth noting that when estimating the number of regimes that have occurred in the sample, we are recognizing that there may be "collapsed regimes" – regimes that do not occur – in a particular MCMC draw. This is the terminology of Scott (2002), who points out that in a model where collapsed regimes are likely it will be important to use a hierarchical prior and substantive prior restrictions to avoid pathologies in the MCMC sampler that collapsed regimes can produce.

[5] See also Sims (1993), Cogley and Sargent (2002), Stock and Watson (2003), Canova and Gambetti (2004) and Primiceri (2005).

[6] Discontinuous shifts have been found to be an important source of conditional heteroskedasticity observed in many macroeconomic time series (Kim and Nelson,1999, Chapter 6; Hamilton, 1988).

3.3. Other models

In this subsection we further illustrate the flexibility of our framework by applying it to other models that are often used in macroeconomics.

3.3.1. Time-dependent transition probabilities

In this subsection we demonstrate the flexibility of our framework by showing an example of using the transition matrix to capture some threshold features. Consider a three-regime example where the third regime is irreversible. A transition to this absorbing regime is time-dependent and the transition probability at time t is $\iota\{f(Y_{t-1}, Z_{t-1}) > c\}$ where c is a real constant, $f()$ is a function, and $\iota\{\}$ is an indicator function that returns to 1 if the statement in the curly brackets is true and 0 otherwise. The transition probability matrix from s_{t-1} to s_t takes the form

$$Q_{t-1} = \begin{bmatrix} q_{1,1} & q_{1,2}\,(1 - \iota\{f(Y_{t-1}, Z_{t-1}) > c\}) & 0 \\ q_{2,1} & q_{2,2}\,(1 - \iota\{f(Y_{t-1}, Z_{t-1}) > c\}) & 0 \\ 0 & \iota\{f(Y_{t-1}, Z_{t-1}) > c\} & 1 \end{bmatrix}.$$

To put these restrictions in the matrix form M_{t-1}, let $v = 3, d_1 = 2$, $d_2 = 2$, and $d_3 = 1$. The 9×5 matrix M_{t-1} is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & (1 - \iota\{f(Y_{t-1}, Z_{t-1}) > c\}) & 0 & 0 \\ 0 & 0 & 0 & (1 - \iota\{f(Y_{t-1}, Z_{t-1}) > c\}) & 0 \\ 0 & 0 & 0 & 0 & \iota\{f(Y_{t-1}, Z_{t-1}) > c\} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As will be discussed in Section 5, the MCMC algorithm valid for the constant transition matrix can be carried over to the case where transition probabilities are functions of Y_{t-1} and Z_{t-1}.

3.3.2. Cross-time composite Markov process

The original approach of Hamilton (1989) makes it explicit for the model parameters to depend on not only the current regime but also the previous regime. Such a historical dependence on regimes can also arise from regime-switching DSGE models (Liu et al., 2008). To see how to deal with this feature in our framework, consider the original regime variable, denoted by s_t^o, takes on two values and has the transition matrix $Q^o = \left(q_{i,j}^o\right)$. Let the composite Markov process, $s_t = \{s_t^o, s_{t-1}^o\}$, consist of a pair of current and previous regimes. There are four possibilities for s_t and the overall transition matrix Q for s_t must be of the form

		(s_{t-1}, s_{t-2})			
		(1, 1)	(1, 2)	(2, 1)	(2, 2)
(s_t, s_{t-1})	(1, 1)	$q_{1,1}^o$	$q_{1,1}^o$	0	0
	(1, 2)	0	0	$q_{1,2}^o$	$q_{1,2}^o$
	(2, 1)	$q_{2,1}^o$	$q_{2,1}^o$	0	0
	(2, 2)	0	0	$q_{2,2}^o$	$q_{2,2}^o$.

To express this restricted Q in the form of (1), we have $v = 2$, $d_1 = d_2 = 2, M_{1,2} = M_{2,2} = M_{3,1} = M_{4,1} = 0$,

$$M_{1,1} = M_{2,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad M_{3,2} = M_{4,2} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

This same technique applies to a switching model that depends on the composite regime $s_t = \{s_t^o, \ldots, s_{t-k}^o\}$ for any integer $k \geq 0$.

3.3.3. Independent Markov processes

In macroeconomic applications, a Markov process governing one equation (such as monetary policy) may be independent of a Markov process controlling another equation (such as fiscal policy); a Markov process governing shock processes may be independent of a Markov process dictating coefficients in the model. In general, we consider τ independent Markov processes such that $h = \prod_{k=1}^{\tau} h^k$ and $H = \prod_{k=1}^{\tau} H^k$, where $H^k = \left\{1, \ldots, h^k\right\}$, $s_t = \left(s_t^1, \ldots, s_t^\tau\right)$, and $s_t^k \in H^k$. The transition matrix Q is therefore restricted to the form

$$Q = Q^1 \otimes \cdots \otimes Q^\tau$$

where $Q^k = \left(q_{i,j}^k\right)$ is an $h^k \times h^k$ matrix such that

$$q_{i,j}^k \geq 0 \quad \text{and} \quad \sum_{i \in H^k} q_{i,j}^k = 1.$$

The tensor product representation of Q implies that if $i = \left(i^1, \ldots, i^\tau\right) \in H$ and $j = \left(j^1, \ldots, j^\tau\right) \in H$, then $q_{i,j} = \prod_{k=1}^{\tau} q_{i^k,j^k}^k$. Conditional on Q, the composite Markov process s_t consists of τ independent Markov processes s_t^k. If Q were not restricted to this tensor product representation, it would contain $\left(\prod_{k=1}^{\tau} h^k\right)\left(\prod_{k=1}^{\tau} h^k - 1\right)$ parameters. With these non-linear restrictions, there are only $\sum_{k=1}^{\tau} h^k\left(h^k - 1\right)$ parameters—a substantial reduction of the number of parameters.

One can, moreover, combine this type of restriction with restrictions on each Q^k individually. Specifically, we let

- q^k be the $\left(h^k\right)^2$-dimensional vector obtained by stacking the columns of Q^k,
- w_j^k be a d_j^k-dimensional vector whose elements are non-negative and sum to one for $1 \leq j \leq v^k$,
- w^k be the the d^k-dimensional vector obtained by stacking the w_j^k, where $d^k = \sum_{j=1}^{v^k} d_j^k$,
- M^k be a $\left(h^k\right)^2 \times d^k$ matrix satisfying Conditions 5 and 6.

It follows from Section 2.2 that Q^k can restricted by requiring

$$q^k = M^k w^k.$$

In the remainder of this paper, we simplify the notation by suppressing the superscript denoting the particular independent Markov regime variable that is under consideration. It is important to remember, however, that all of the results apply to a product of independent Markov regime variables by simply putting the superscript k back in appropriate places.

3.3.4. Correlated Markov processes

It is often argued that business cycle turning points contain leading and coincident components (Stock and Watson, 2002). In a large panel data set, some time series may be grouped as leading indicators and others as coincident indicators. Using the multiple-equation Markov-switching framework, we can identify one regime governing the parameters in the leading-indicator equations and another regime governing those in the coincident-indicator equations (Kim and Nelson (1999, Chapter 5), and Kaufmann (2007)). Denote the leading regime variable by s_{lt} and the coincident regime variable by s_{ct}; and consider the composite regime variable $s_t = \{s_{lt}, s_{ct}\}$, where $s_{lt} = \{1, 2\}$ and $s_{ct} = \{1, 2\}$. Since the regime variable s_{lt} leads the regime variable s_{ct}, the transition matrix Q for the composite regime variable s_t is

		(s_{lt-1}, s_{ct-1})			
		(1, 1)	(1, 2)	(2, 1)	(2, 2)
(s_{lt}, s_{ct})	(1, 1)	π_1	$1 - \pi_2$	0	0
	(1, 2)	0	π_2	0	$1 - \pi_4$
	(2, 1)	$1 - \pi_1$	0	π_3	0
	(2, 2)	0	0	$1 - \pi_3$	π_4.

To represent the restrictions imposed on Q in the form of (1), we have $v = 4, d_1 = d_2 = d_3 = d_4 = 2$,

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad M_{2,2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$M_{3,3} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad M_{4,4} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

and $M_{i,j} = 0$ for $i \neq j$. From Q one can derive the transition probabilities for s_{lt} and s_{ct}. Let $\bar{q}_1$, $\bar{q}_2$, $\bar{q}_3$, and $\bar{q}_4$ be the ergodic probabilities for the composite regime variable s_t and assume that $\bar{q}_1 + \bar{q}_2 \neq 0$, $\bar{q}_3 + \bar{q}_4 \neq 0$, $\bar{q}_1 + \bar{q}_3 \neq 0$, and $\bar{q}_2 + \bar{q}_4 \neq 0$. It follows that

$$\Pr\{s_{lt} = 1 \mid s_{lt-1} = 1\} = \frac{\bar{q}_1\pi_1 + \bar{q}_2}{\bar{q}_1 + \bar{q}_2},$$

$$\Pr\{s_{lt} = 2 \mid s_{lt-1} = 2\} = \frac{\bar{q}_3 + \bar{q}_4\pi_4}{\bar{q}_3 + \bar{q}_4},$$

$$\Pr\{s_{ct} = 1 \mid s_{ct-1} = 1\} = \frac{\bar{q}_1 + \bar{q}_3\pi_3}{\bar{q}_1 + \bar{q}_3},$$

$$\Pr\{s_{ct} = 2 \mid s_{ct-1} = 2\} = \frac{\bar{q}_2\pi_2 + \bar{q}_4}{\bar{q}_2 + \bar{q}_4}.$$

In general, for any composite regime $s_t = \left(s_t^1, \ldots, s_t^\tau\right)$ and the associated transition matrix Q, one can restrict Q such that the marginal probability of s_t^k and the probability of s_t^k conditional on other types of regimes satisfy certain properties.

4. The prior

In this section we describe a way to set the prior on all the model parameters. We begin with the case where Q is unrestricted, as this case is commonly considered in the literature. For $1 \leq i, j \leq h$, let $\alpha_{i,j}$ be a positive number. The prior on Q is of the Dirichlet form

$$p(Q) = \prod_{j \in H} \left[\left(\frac{\Gamma\left(\sum_{i \in H} \alpha_{i,j}\right)}{\prod_{i \in H} \Gamma\left(\alpha_{i,j}\right)} \right) \times \prod_{i \in H} \left(q_{i,j}\right)^{\alpha_{i,j}-1} \right], \tag{3}$$

where $\Gamma(\cdot)$ denotes the standard gamma function. In such an unrestricted case if our a priori beliefs about the persistence of regimes are symmetric across regimes and we do not expect regimes to be ordered by "distance" from one another, it is natural to form the prior to reflect prior beliefs about the persistence of the regimes, making the probabilities of jumping to other regimes identical. This is also what Sims and Zha (2006) did. In the application below we set $\alpha_{i,j} = 1$ for $i \neq j$ and let $p_{j,\mathrm{dur}} = Eq_{j,j}$ be the expected value of the probability of staying in the same regime j. We have

$$p_{j,\mathrm{dur}} = Eq_{j,j} = \frac{\alpha_{j,j}}{\sum_i \alpha_{i,j}} = \frac{\alpha_{j,j}}{\alpha_{j,j} + (h-1)}.$$

It follows that

$$\alpha_{j,j} = \frac{p_{j,\mathrm{dur}}(h-1)}{1 - p_{j,\mathrm{dur}}}. \tag{4}$$

By setting $\alpha_{i,j} = 1$ for $i \neq j$, we insure that the prior density is bounded away from zero as $q_{j,j} \to 1$.

Sims and Zha (2006) instead chose both a mean and a prior standard deviation for $q_{j,j}$, and chose the prior standard deviation tight enough to make $\alpha_{i,j} > 1$ for $i \neq j$. If we set $\alpha_{i,j} = \bar{\alpha}_j$ for all $i \neq j$, the implied marginal prior for $q_{j,j}$ is a Beta$(\alpha_{j,j}, (h-1)\bar{\alpha}_j)$ pdf. One could plot these one-variable pdf's and choose $\bar{\alpha}_j$ (holding the mean $\alpha_{j,j}/(\alpha_{j,j} + (h-1)\bar{\alpha}_j)$ fixed) to make this pdf look reasonable. With a large number of regimes, this can easily lead to choosing $\bar{\alpha}_j < 1$, implying that the marginal density on the off-diagonal $q_{i,j}$'s is unbounded at zero. There is no simple answer to these conflicting criteria. In models with large numbers of regimes, imposing more structure on Q is attractive in part because symmetric priors would look unreasonable without careful parameterizations.

In our empirical section (Section 8.7), where quarterly data are used, for example, we set $p_{j,\mathrm{dur}} = 0.85$, implying a prior belief that the average duration of staying in the same regime is between six and seven quarters. In the two-regime case, it follows from (4) that, with the convention that $\bar{\alpha}_j = 1$,

$$\alpha_{j,j} = 5.666667, \qquad \alpha_{i,j} = 1 \quad \text{for } i \neq j. \tag{5}$$

In the four-regime case, it follows from (4) that

$$\alpha_{j,j} = 17, \qquad \alpha_{i,j} = 1 \quad \text{for } i \neq j. \tag{6}$$

When the dimension of Q increases, one must prevent the number of free parameters in Q from growing too fast by restricting the transition matrix Q as in Section 2.2. How does one specify the prior for the restricted Q? Denote $w_j = \left[w_{1,j}, \ldots, w_{d_j,j}\right]'$. The prior on w_j is of the Dirichlet form

$$\frac{\Gamma\left(\sum_{i=1}^{d_j} \beta_{i,j}\right)}{\prod_{i=1}^{d_j} \Gamma\left(\beta_{i,j}\right)} \prod_{i=1}^{d_j} \left(w_{i,j}\right)^{\beta_{i,j}-1} \tag{7}$$

where $\beta_{i,j} > 0$. The prior on Q can be recovered via (1).

In the restricted case, it is important that the value of the hyperparameter $\beta_{j,j}$ be chosen according to the prior on w_j, not on the prior of the unrestricted parameter vector q_j. To see this point, consider the four-regime case with the the transition matrix restricted as in (2). Take as an example the first two columns of this Q and express the restrictions on q_1 and q_2 in the form of $q_j = M_j w_j$:

$$q_{1,1} = w_{1,1}, \qquad q_{2,1} = w_{2,1}, \qquad q_{3,1} = 0, \qquad q_{4,1} = 0,$$
$$q_{2,2} = w_{2,2}, \qquad q_{1,2} = \frac{1}{2}w_{1,2}, \qquad q_{3,2} = \frac{1}{2}w_{1,2}, \qquad q_{4,2} = 0.$$

If we take as given the values of $\alpha_{i,j}$ specified in (4) (as supplied by the user who is accustomed to working on an unrestricted transition matrix) and transform them to $\beta_{i,j}$ as

$$\beta_{i,j} = 1 + \sum_{\{(r,s):M_{r,j}(s,i)>0\}} \left(\alpha_{r,s} - 1\right),$$

we have

$$\beta_{1,1} = \alpha_{1,1}, \qquad \beta_{2,1} = \alpha_{2,1} = 1,$$
$$\beta_{2,2} = \alpha_{2,2}, \qquad \beta_{1,2} = \alpha_{1,2} = 1.$$

According to the Dirichlet prior (7) directly imposed on the free parameters w, we have

$$Ew_{1,1} = \frac{\beta_{1,1}}{\beta_{1,1} + \beta_{2,1}}, \qquad Ew_{2,1} = \frac{\beta_{2,1}}{\beta_{1,1} + \beta_{2,1}},$$
$$Ew_{2,2} = \frac{\beta_{2,2}}{\beta_{2,2} + \beta_{1,2}}, \qquad Ew_{1,2} = \frac{\beta_{1,2}}{\beta_{2,2} + \beta_{1,2}}.$$

If we were to use the values specified in (6) for the unrestricted Q, we would have $Eq_{j,j} = Ew_{j,j} = 0.94$, implying a prior belief that the average duration of staying in the same regime is about 17 quarters, much longer than the prior belief when Q is unrestricted. This is not the prior we have originally intended to specify. For

this reason, we insist that the prior be specified directly on $w_{i,j}$ to maintain the same prior belief on the average duration, no matter we work on an unrestricted or restricted transition matrix. In our four-regime case, we let the hyperparameters $\beta_{j,j} = 5.666667$ (not 17 as (6)) and $\beta_{i,j} = 1.0$ for $i \neq j$. It follows that $p_{j,\text{dur}} = 0.85$, the same prior duration as we have intended.

The joint prior density for θ, w, S_T is

$$p(\theta, w, S_T) = p(\theta, w)\, p(s_0 \mid \theta, w) \prod_{t=1}^{T} p(s_t \mid \theta, w, S_{t-1}).$$

By Condition 1, $p(s_t \mid \theta, w, S_{t-1}) = q_{s_t, s_{t-1}}$. We assume that the prior on θ is independent of the prior on w and that $p(s_0 \mid \theta, w) = \frac{1}{h}$ for every $s_0 \in H$. A common alternative assumption for $p(s_0 \mid \theta, Q)$ makes it the ergodic distribution of Q, if the ergodic distribution exists. This convention, however, makes the conditional posterior distribution of Q an unknown and complicated one. With our specification of equiprobable initial regimes, the resulting prior has the following form

$$p(\theta, w, S_T) = \frac{p(\theta)\, p(w)}{h} \prod_{t=1}^{T} q_{s_t, s_{t-1}}. \tag{8}$$

The simplicity of the posterior can be preserved with other choices of initial distribution for the regimes that do not make it a complicated function of the transition probabilities. For example, one could normalize the regimes by insisting that regime 1 prevails with probability one at the initial date in the sample, as long as this is a non-absorbing regime.

The prescriptions we give here for formulating priors should be taken as practical suggestions, not definitive rules. In models like these, with large numbers of parameters, results can easily be sensitive to the prior, and priors set in conventional ways can easily turn out to have unexpected and unintended implications. Results, especially for the marginal data density calculations used in model comparisons, should be checked for robustness against variations in the prior.

Most applications of these methods will not be direct input to a single decision, but instead will be in the nature of scientific reporting. The task of inference is therefore to characterize likelihood shape to a wide range of potential readers, not to assess a unique best prior for decision making. The prior should be kept as simple and understandable as possible. It should reflect prior beliefs likely to be common across the study's readership, not in general the beliefs of the researcher preparing the study.

Because results are likely to be sensitive to choice of the prior, and because a careful choice of these highly multivariate priors is a complex task, various shortcuts are sometimes used in practice. The BIC criterion, for example, allows model comparison without assessing any prior. In large samples, it will (under regularity conditions) point to the same model as best as does any calculation of posterior odds based on a prior (though it will not, even in large samples, provide an accurate quantitative approximation to the posterior odds). However, it amounts to using a conventional prior, and precisely because in large models the implications of priors are hard to assess, the chance that the BIC is implicitly using a prior with bizarre implications is greater the larger and more complex the model.

Another common shortcut is the "training sample prior". This is equivalent to using the likelihood function itself as the posterior, i.e. to using a flat prior, when we use the training sample for inference on parameters. When we use it for model comparison, it scales the likelihood so that the integral of the likelihood over the training sample is one (so it functions as a "prior"). This undercuts the tendency of Bayesian model comparison to penalize large models. Indeed, if a fixed proportion of the sample is used as a training sample, Bayesian posterior odds will not converge in large samples to the smaller of two nested models when the more restricted model is the truth. This is one reflection of a general point: training sample priors make large models improve relative to small models. Especially when comparing complex models, therefore, training sample priors should be treated as no more than a temporary, expedient shortcut.

5. Likelihood and posterior distribution

The key step in evaluating the overall likelihood function $p(Y_T \mid Z_t, \theta, w)$ for Markov-switching models is to obtain the conditional likelihood function at time t:

$$p(y_t \mid Y_{t-1}, Z_t, \theta, w, s_t). \tag{9}$$

Our framework applies to any model if one can write down the conditional likelihood (9). As long as h, the number of values s_t takes, is finite and does not grow with t, the regime variable s_t can be a complicated composite Markov process, as discussed in Section 2.2. In their nonlinear dynamic general equilibrium model, for example, Sargent et al. (2006) show that one of the most difficult tasks is to derive this conditional likelihood function. Another example pertains to Markov-switching state-space models in which Kim and Nelson (1999) show how to obtain the conditional likelihood function (9) that can be approximated well without radically increasing the number of regimes h, so that Condition 3 holds approximately.

Given (9) and conditional on the vector of exogenous variables Z_t, the likelihood of Y_T is[7]

$$p(Y_T \mid Z_T, \theta, w) = \prod_{t=1}^{T} \left[\sum_{s_t \in H} p(y_t \mid Y_{t-1}, Z_t, \theta, w, s_t)\, p(s_t \mid Y_{t-1}, Z_{t-1}, \theta, w) \right]. \tag{10}$$

This likelihood can be evaluated recursively by updating $p(s_t \mid Y_{t-1}, Z_{t-1}, \theta, w)$ according to Propositions 1 and 2 in Appendix A.

By the Bayes rule it follows from (8) and (10) that the posterior distribution of (θ, w) is

$$p(\theta, w \mid Y_T, Z_T) \propto p(\theta, w) p(Y_T \mid Z_T, \theta, w). \tag{11}$$

The posterior density $p(\theta, w \mid Y_T, Z_T)$ is not of standard form, making it impossible to sample directly from this probability distribution. One can, however, use the idea of Gibbs sampling to obtain the empirical joint posterior density $p(\theta, w, S_T \mid Y_T, Z_T)$ by sampling alternately from the following conditional posterior distributions:

$$p(S_T \mid Y_T, Z_T, \theta, w),$$
$$p(w \mid Y_T, Z_T, S_T, \theta),$$
$$p(\theta \mid Y_T, Z_T, w, S_T).$$

Simulation from the conditional posterior density $p(\theta \mid Y_T, Z_T, w, S_T)$ is model-dependent, which we will discuss in the context of VARs in Section 8.

To simulate draws of S_T from $p(S_T \mid Y_T, Z_T, \theta, w)$, we begin with a draw from $p(s_T \mid Y_T, Z_T, \theta, w)$ obtained from the aforementioned forward recursion and work recursively backward to draw s_{T-1}, $s_{T-2}, \ldots, s_0$ according to

$$p(s_t \mid Y_T, Z_T, \theta, w) = \sum_{s_{t+1} \in H} p(s_t \mid Y_t, Z_t, \theta, w, s_{t+1})\, p(s_{t+1} \mid Y_T, Z_T, \theta, w),$$

[7] See Appendix B for details.

where

$$p\left(s_t \mid Y_t, Z_t, \theta, w, s_{t+1}\right) = \frac{q_{s_{t+1},s_t}\, p\left(s_t \mid Y_t, Z_t, \theta, w\right)}{p\left(s_{t+1} \mid Y_t, Z_t, \theta, w\right)}.$$

This result follows from Proposition 3 in Appendix A (see Appendix B for details of the derivation).

The conditional posterior density of w derives directly from the conditional posterior density of the free parameters w_j.[8] It follows from Conditions 1 and 4 and the prior (7) that

$$p\left(w_j \mid Y_T, Z_T, \theta, S_T\right) \propto \prod_{i=1}^{d_j} \left(w_{i,j}\right)^{n_{i,j}+\beta_{i,j}-1}, \tag{12}$$

where $n_{i,j}$ is the number of transitions from $s_{t-1} = r$ to $s_t = s$ for $M_{r,j}(s, i) > 0$ and $M_{r,j}(s, i)$ is the sth-row and ith-column element of the submatrix $M_{r,j}$.

Although Condition 4 is valid for most reduced-form Markov-switching models, it does not hold for forward-looking models such as regime-switching rational expectations models (Farmer et al., 2006).[9] In such a case, however, the Dirichlet density derived as though this condition were true is still valuable because it can be used to form a basis for the proposal density in the Metropolis–Hastings algorithm used for sampling from the true conditional posterior distribution of w.

We now discuss the situation in which q is no longer constant as in the example of Section 3.3.1. Let the transition probability from $s_{t-1} = j$ to $s_t = i$ be $q_{i,j}\left(Y_{t-1}, Z_{t-1}, w\right)$, where $q_{i,j}(\,)$ is a general function and w is a vector of free parameters. It can be shown that the forward-recursion evaluation of the likelihood $p(\theta, w \mid Y_T, Z_T)$ and the backward-recursion algorithm of drawing S_T from $p(S_T \mid Y_T, Z_T, \theta, w)$ continue to be valid, as long as q_{s_{t+1},s_t} is replaced by $q_{s_{t+1},s_t}\left(Y_{t-1}, Z_{t-1}, w\right)$ (see Appendices A and B for details). In general, however, the posterior density $p\left(w \mid Y_T, Z_T, \theta, S_T\right)$ is not of the Dirichlet form. In this case, the Metropolis–Hastings algorithm can be used to sample from $p\left(w \mid Y_T, Z_T, \theta, S_T\right)$ and the Dirichlet density (12) can be used as a basis for the proposal density.

6. Blockwise optimization algorithm

In spite of the complexity inherent in Markov-switching multiple-equation models, it is important to find the posterior estimate at the posterior mode or MLE for several reasons. When the shape of the posterior density tends to be very non-Gaussian, as it is often the case for Markov-switching multiple-equation models, the posterior mean may have a very low probability and cannot represent the most likely scenario for the model. The posterior mode, on the other hand, always represents the most likely point, regardless of how non-Gaussin the posterior distribution is. Using a point near the posterior mode as a starting point for the MCMC algorithm, one can ensure that an unreasonably long sequence of posterior draws do not get stuck in the low probability region. The posterior mode can be used as a reference point in normalization to help avoid distorting the statistical inferences likely to be produced by inappropriate normalization (Waggoner and Zha, 2003b). And the likelihood value conditional on the posterior estimate helps detect obvious errors in computing marginal data densities for posterior odds ratios.

Hamilton (1994) proposes an EM algorithm to find the posterior estimate or MLE for a simple Markov-switching model, where the E-step can be completed analytically. For multivariate dynamic models, however, the E-step in general has no analytical form. Chib (1996) proposes a Monte Carlo EM (MCEM) algorithm in which the evaluation of the E-step of the EM algorithm is approximated by Monte Carlo simulations from the posterior distribution. For large regime-switching multiple-equation models, the MCEM algorithm can be very costly and sometimes take a couple of weeks to obtain an estimate that is close to the peak of the likelihood, as shown in Sims and Zha (2006). For parameter-drifting or stochastic volatility models (Cogley and Sargent, 2005; Justiniano and Primiceri, 2008), it is even infeasible to integrate out all latent variables numerically.[10]

Because the cost of the numerical integration of S_T can be substantial, there is no need to use the MCEM algorithm as long as the posterior density $p(\theta, w \mid Y_T, Z_T)$ given by (11) is available for evaluation. When the number of parameters is small, one may obtain the posterior estimate of θ by simply finding the value of θ that maximizes the posterior density. Sims (2001) uses this approach for his single-equation model. But for a system of multivariate dynamic equations, the number of model parameters may be too large for a straight maximization routine to be reliable.

We propose a different algorithm that is designed to work for both small and large models. We first break the parameters (θ, w) into two blocks of parameters θ and w. In practice, this separation proves critical because the conditional posterior density of θ differs substantially from that of w. If the dimension of θ is large, as in the multivariate dynamic models considered in Section 8.7, we recommend to break θ further into several sub-blocks. Given an initial guess of the values of the parameters, one can use a standard hill-climbing quasi-Newton optimization routine to find the value of each block of parameters that maximizes the posterior density while holding other blocks of parameters fixed at the previous values. Iterate this algorithm through blocks until it approximately converges. For each iteration, we recommend to employ a *constrained* optimization method to check whether there are boundary solutions associated with w or other model parameters. While this blockwise approach will at first increase likelihood more efficiently than a quasi-Newton method applied directly to the complete parameter vector, blockwise methods can be very inefficient at the final stages of convergence. When the blockwise iterations have converged or nearly converged, additional direct quasi-Newton steps on the full parameter vector should be undertaken, with BFGS (Broyden–Fletcher–Goldfarb–Shanno) updates of the full Hessian. These additional steps in our experience substantially improve the likelihood value. In Section 8.7, we show in an example that this algorithm is more efficient than the MCEM algorithm.

7. New implementation of the MHM method

Estimating the marginal data density is an important task when one compares a large set of different models for the purpose of

[8] To be consistent with Section 4, we suppress the superscript k that indicates a particular Markov process under study.

[9] We thank Tim Cogley for drawing our attention to this point.

[10] Consider the simple one-dimensional model $y_t = a_t y_{t-1} + \sigma_t \epsilon_t$, where ϵ_t is normally distributed with mean zero and variance ξ^2. The drifting parameter a_t and the volatility parameter σ_t are treated as latent state variables following the stochastic processes

$$a_t = (1-\rho_a)a + \rho_a a_{t-1} + \nu_{a,t}, \qquad \log\sigma_t = (1-\rho_\sigma)\log\sigma + \rho_\sigma \log\sigma_{t-1} + \nu_{\sigma,t},$$

where $\nu_{a,t}$ is normally distributed with mean zero and variances ξ_a^2 and $\nu_{\sigma,t}$ is normally distributed with mean zero and variances ξ_σ^2. One could form the *conditional* likelihood

$$p\left(Y_T \mid \xi, \xi_a, \xi_\sigma, a, \rho_a, \sigma, \rho_\sigma, a_1, \ldots, a_T, \sigma_1, \ldots, \sigma_T\right).$$

To find the peak of the likelihood $p\left(Y_T \mid \xi, \xi_a, \xi_\sigma, a, \rho_a, \sigma, \rho_\sigma\right)$ itself, however, one must numerically integrate out all the latent variables $a_1, \ldots, a_T, \sigma_1, \ldots, \sigma_T$. This task is computationally infeasible even for a moderate sample size T.

selecting the one that best fits to the data or for the purpose of averaging a subset of models. For many macroeconomic models, the modified harmonic mean (MHM) method of Gelfand and Dey (1994) has been a widely used method for computing the marginal data density. In this section we discuss the potential problem with this method when the posterior distribution is very non-Gaussian and propose a new way of implementing the MHM method to remedy this problem. For notational clarity, we restrict ourselves to the constant-parameter case, treat θ as a collection of all the free parameters in the model, and omit the exogenous variables Z_T and the transition probabilities w. At the end of this section, we discuss how to handle the Markov-switching models.

We begin by denoting the likelihood function by $p(Y_T \mid \theta)$ and the prior density by $p(\theta)$, both of which must have proper probability densities instead of their kernels. Given these two objects, the marginal data density is defined as

$$p(Y_T) = \int p(Y_T \mid \theta)p(\theta)\mathrm{d}\theta. \tag{13}$$

The MHM method used to approximate (13) numerically is based on the following theorem

$$p(Y_T)^{-1} = \int_{\Theta} \frac{h(\theta)}{p(Y_T \mid \theta)p(\theta)} p(\theta \mid Y_T)\mathrm{d}\theta, \tag{14}$$

where Θ is the support of the posterior probability density and $h(\theta)$, often called a *weighting* function, is any probability density whose support is contained in Θ. Denote

$$m(\theta) = \frac{h(\theta)}{p(Y_T \mid \theta)p(\theta)}.$$

A numerical evaluation of the integral on the right hand side of (14) can be accomplished in principle through the Monte Carlo (MC) integration

$$\hat{p}(Y_T)^{-1} = \frac{1}{N}\sum_{i=1}^{N} m(\theta^{(i)}), \tag{15}$$

where $\theta^{(i)}$ is the ith draw of θ from the posterior distribution $p(\theta \mid Y_T)$. If $m(\theta)$ is bounded above, the rate of convergence from this MC approximation is likely to be practical.

Geweke (1999) proposes an implementation with $h(\cdot)$ constructed from the posterior simulator. The sample mean $\bar{\theta}$ and sample covariance matrix $\bar{\Omega}$ can be calculated from draws of θ from the posterior simulator. The weighting function is chosen to be a truncated multivariate Gaussian density with mean $\bar{\theta}$ and covariance $\bar{\Omega}$. The tail of the Gaussian distribution is truncated to ensure that the support of the weighting function is contained in the support of posterior. Our experience suggests that this method works well for many existing DSGE and VAR models with no time variation on the parameters. When one allows time variation in the model's parameters, the posterior density tends to be non-Gaussian. The non-Gaussian phenomenon is manifested in three aspects. First, the posterior density can be very low at the sample mean, especially when the posterior density has multiple peaks. Second, a truncated Gaussian density function tends to be a poor local approximation to the non-Gaussian posterior density. Third, the likelihood can get close to zero in the interior points of the parameter space Θ.

To deal with the first two problems, we propose a more general class of distributions than the Gaussian family, center and scale these distributions differently, and truncate them in a more sophisticated manner. We begin with the easiest task, which involves the centering and scaling. Instead of centering the weight pdf at the sample mean, we center at the posterior mode $\hat{\theta}$; instead of scaling by the sample covariance matrix, we use

$$\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}\left(\theta^{(i)} - \hat{\theta}\right)\left(\theta^{(i)} - \hat{\theta}\right)'$$

where $\theta^{(i)}$ denotes the ith draw from the posterior simulator and N is the sample size. Computing the posterior mode is typically more expensive than computing the sample mean, but it greatly improves efficiency of the MHM method. Instead of a family of Gaussian distributions, we use a family of elliptical distributions. An elliptical distribution centered at $\hat{\theta}$ and scaled by $\hat{S} = \sqrt{\hat{\Omega}}$ has a density of the form

$$g(\theta) = \frac{\Gamma(k/2)}{2\pi^{k/2}\left|\det\left(\hat{S}\right)\right|}\frac{f(r)}{r^{k-1}}$$

where k is the dimension of θ, $r = \sqrt{\left(\theta - \hat{\theta}\right)'\hat{\Omega}^{-1}\left(\theta - \hat{\theta}\right)}$, and $f(\,)$ is any one-dimensional density defined on the positive reals. We note that the Gaussian distribution is a special case in the family of elliptical distributions. Since we know how to sample from the one-dimensional density $f(\,)$, making draws for an elliptical distribution is straightforward. We simply draw x from the k-dimensional standard Gaussian distribution and r from the density $f(\,)$, and define

$$\theta = \frac{r}{\|x\|}\hat{S}x + \hat{\theta}.$$

The one-dimensional density $f(\,)$ is chosen in the following way. For each draw $\theta^{(i)}$ from the posterior distribution, let

$$r^{(i)} = \sqrt{\left(\theta^{(i)} - \hat{\theta}\right)'\hat{\Omega}^{-1}\left(\theta^{(i)} - \hat{\theta}\right)}.$$

From these simulated $r^{(i)}$, we can easily form an estimate of their cumulative density function. The density $f(r)$ should be chosen so that its cumulative density closely matches the estimated one. There are many ways to accomplish this task. For instance, $f(r)$ could be chosen to be a step function such that the cumulative density is a piecewise-linear approximation to the estimated cumulative density. We chose a simpler but efficient technique. Let the density $f(r)$ have a support on $[a, b]$ and be defined as

$$f(r) = \frac{v r^{v-1}}{b^v - a^v}.$$

The hyperparameters a, b, and v are chosen as follows. Let c_1, c_{10}, and c_{90} be chosen so that one percent of the $r^{(i)}$ are less than c_1, ten percent of the $r^{(i)}$ are less than c_{10}, and ninety percent of the $r^{(i)}$ are less than c_{90}. Denote the density function $f(r)$ with $a = 0$ by $f_0(r)$. The values of b and v are so chosen that the probability of $r < c_{10}$ from $f_0(r)$ is 0.1 and the probability of $r < c_{90}$ from $f_0(r)$ is 0.9. These choices translate into

$$v = \frac{\log(1/9)}{\log(c_{10}/c_{90})}, \qquad b = \frac{c_{90}}{0.9^{1/v}}. \tag{16}$$

For the reasons elaborated below, we set the value of a to c_1 to keep $f(r)$ bounded above. With the nonzero value of a and the values of v and b specified in (16), one should note that the probability of $r < c_p$ from $f(r)$ will not be exactly p, where $p = 0.1$ or $p = 0.9$.

To deal with the third problem, i.e., the likelihood tending to be zero in the interior points of the parameter space, we propose a method to truncate the elliptical distribution. Let U be a positive number and Θ_U be the region defined by

$$\Theta_U = \{\theta : m(\theta) < U\}.$$

The weighting function $h(\theta)$ is chosen to be an elliptical density function truncated so that its support is Θ_U. If q_U is the probability that draws from the elliptical distribution lies in Θ_U, then $h(\theta)$ is given by

$$h(\theta) = \frac{\chi_{\Theta_U}(\theta)}{q_U} g(\theta),$$

where $\chi_A(\theta)$ is an indicator function that returns one if θ falls in the set A and zero otherwise. The value of q_U can be estimated from random draws from the elliptical density $g(\theta)$. Since draws from the elliptical distribution are i.i.d., the estimate of q_U has a binomial distribution and its accuracy can be readily obtained. The lower the truncation value of U, the larger the effective sample size of a sequence of $m(\theta^{(i)})$, but the less accurate the value of $\hat{q}_U$. Therefore, there is a balance between having a low cut-off value of U and having a reasonable estimate of q_U.

Because we chose a nonzero value of a for $f(r)$, the weight function $h(\theta)$ is effectively bounded above. Thus, the upper bound truncation on $m(\theta)$ can be easily implemented by a lower bound truncation on the posterior density kernel itself. Specifically, Let L be a positive number and Θ_L be the region defined by

$$\Theta_L = \{\theta : p(Y_T \mid \theta)p(\theta) > L\}.$$

The weighting function $h(\theta)$ is chosen to be a truncated elliptical density such that its support is Θ_L. If q_L is the probability that random draws from the elliptical distribution lies in Θ_L, $h(\theta)$ is given by

$$h(\theta) = \frac{\chi_{\Theta_L}(\theta)}{q_L} g(\theta).$$

Our computational experience indicates that a good choice of L is a value such that 90% of draws from the posterior distribution lie in Θ_L.

To implement our new MHM method in practice, we denote the kernel of the posterior probability density by

$$k(\theta|Y_T) = p(Y_T \mid \theta)p(\theta).$$

The procedure for implementing our new MHM method is as follows.

(1) Simulate a sequence of posterior draws $\theta^{(i)}$ and record the minimum and maximum values of $k(\theta|Y_T)$, denoted by $k_{\min}$ and $k_{\max}$ respectively. Let $k_{\min} < L < k_{\max}$.
(2) Simulate i.i.d. draws of θ from $g(\theta)$ and compute the proportion of these draws that belong to Θ_L. This proportion, denoted by $\hat{q}_L$, is the estimate of q_L. The estimate $\hat{q}_L$ has a binomial distribution and its accuracy depends on the number of i.i.d. draws from $g(\theta)$. If $\hat{q}_L < 1.0\text{e}{-}06$, this estimate is unreliable because three or four standard deviations will include the value zero. As a rule of thumb, we keep $\hat{q}_L \geq 1.0\text{e}{-}05$.
(3) For each value of L, estimate the marginal data density according to (15).

Alternatively, our procedure can be implemented by selecting a good value of the upper bound U imposed on $m(\theta)$. Denote the minimum and maximum values of $m(\theta)$ sampled from the posterior distribution by $m_{\min}$ and $m_{\max}$. For each value of $m_{\min} < U < m_{\max}$, compute an estimate of q_U and then obtain an estimate of the marginal data density accordingly.

The importance of choosing the weighting function $h(\theta)$ as close as possible to the possibly non-Gaussian posterior kernel is illustrated in the empirical exercises in Section 8.7. Equally important is our procedure of eliminating the extremely high values of $m(\theta^{(i)}, w^{(i)})$ drawn from the posterior distribution. The development in both areas is crucial to achieving accuracy of the estimated marginal data density. The success depends on how much the weighting function $h(\theta)$ overlaps with the posterior kernel $k(\theta)$.

The computation of q_L provides a practical mechanism to gauge how much of the overlap $h(\theta)$ and $k(\theta)$ share. For any sequence of posterior draws $\theta^{(i)}$, if we increase the variance of $h(\theta)$, the value of $m(\theta^{(i)})$ tends to decrease and thus the estimated marginal data density will be artificially blown up, no matter whether the tail of the distribution represented by $h(\theta)$ is truncated or not. If we discipline the way of changing the variance of $h(\theta)$ by insisting that q_L be computed, we will find that the estimate of q_L goes to zero as the variance of $h(\theta)$ becomes too large or too small. If we cannot estimate q_L, it means that the two densities represented by $h(\theta)$ and $k(\theta)$ have very little overlap and thus the estimated marginal data density is likely to be misleading.[11] In Section 8.7, we use empirical results to illustrate this important point.

We have thus far discussed our new MHM method using the constant-parameter model. For Markov-switching models, the only difference is the treatment of the transition matrix Q in which w_j for $j = 1, \ldots, h$ is a vector of free parameters as discussed in Section 4. The transition matrix parameters w_j's are treated separately from θ because a Dirichlet density as the weighting function for w_j, instead of a truncated power density, is a better approximation to the posterior density of w_j.

In linear Gaussian state-space models, Gerlach et al. (2000) and Giordani and Kohn (2008) develop an efficient MCMC algorithm for sampling s_t conditional on S_1^{t-1}, S_{t+1}^T, Y_T, Z_T, θ, and w. Their approach does not require that Condition 3 should hold, and they show it can be adapted to cases where s_t is not Markov. Using Gerlach et al. (2000) and Giordani and Kohn's (2008) algorithm, one can evaluate the conditional likelihood function $p(Y_T \mid S_T, \theta, w)$. In this paper, we consider a more restrictive class of models, but as a result are able to integrate S_T out of the posterior density analytically and to draw S_T from its conditional posterior distribution jointly, instead of one s_t at a time. The ability to integrate S_T out analytically is crucial to finding the values of θ and w at the posterior peak. Moreover, in evaluating the marginal data density, since the posterior distribution of S_T is non-Gaussian and the dimension of S_T tends to be extremely large, it is more difficult to form a good joint weighting function for S_T, θ, and w than to form one, as in our setup, for θ and w alone. In practice, one may, like Justiniano and Primiceri (2008), use as a weighting function

$$h(\theta, w, S_T) = h(\theta)\, p(w, S_T), \tag{17}$$

where $p(S_T, w)$ is the prior density of S_T and w, as discussed in Section 4. In Section 8.7, we show that because $p(S_T, w)$ can be a poor approximation to the posterior distribution of S_T and w, the estimate of the marginal data density using this weighting function tends to be unreliable.

Kim and Nelson (1999) show how, in a state space model with switching, to get a reliable estimate of $p(Y_T \mid \theta, w)$ by making Condition 3 hold approximately. In the context of Markov-switching state-space model, Schorfheide (2005) shows that the estimate of $p(Y_T \mid \theta, w)$ can be accurately obtained without expanding the dimension of s_t drastically. For other models where Condition 3 holds, such as VARs and the economic model of Sargent et al. (2006), the likelihood $p(Y_T \mid \theta, w)$ can be evaluated exactly. With the evaluation of $p(Y_T \mid \theta, w)$ readily available, one can

[11] This new method, by truncating $m(\theta, w)$ to ensure reasonable overlap between the weighting function and the posterior density, is related to the bridge sampling technique developed by Meng and Wong (1996). Similar to our method, the degree of overlap between the two densities in bridge sampling is crucial to how well that technique works. For bridge sampling, a good weighting function will certainly help achieve an accurate estimate of MDD for a given model.

use the methods developed in the previous section and in this section to find the estimate of θ and w at the posterior mode and compute the marginal data density for each model. These methods are designed to avoid the potential problems associated with the non-Gaussian posterior distribution. The computer software package developed for this paper, together with the parallel and grid computing tools developed by Ramachandran et al. (2007), makes it computationally feasible to estimate and compare a large set of models.

8. Application

In this section we apply the general framework to structural VARs with Markov switching and fit a set of models to the US data. The empirical results are used to illustrate the difficulties encountered by the standard MHM method and the remedies provided by our new method.

Sims and Zha (2006) use a class of structural Markov-switching VARs to study whether and how US monetary policy has changed but leave econometric details to an unpublished manuscript (Sims and Zha, 2004). In this section, we give a complete description of the prior, the likelihood, and the posterior distribution so that researchers can use these results in their specific application.

8.1. Structural VARs

If Markov-switching VARs were put in the state-space form, the existing methods such as those in Kim and Nelson (1999) and Gerlach et al. (2000) could be applied. This approach, however, is inefficient because the state vector is unnecessary and the Kalman filtering can be avoided in the case of VARs. We work directly on VARs without any state vector.

Consider a class of models of the following form:

$$y_t' A(s_t) = \sum_{i=1}^{\rho} y_{t-i}' A_i(s_t) + z_t' C(s_t) + \varepsilon_t' \Xi^{-1}(s_t), \quad \text{for } 1 \le t \le T, \tag{18}$$

where

- ρ is a lag length;
- y_t is an n-dimensional column vector of endogenous variables at time t;
- z_t is an m-dimensional column vector of exogenous and deterministic variables at time t;
- ε_t is an n-dimensional column vector of unobserved random shocks at time t;
- $A(k)$ is an invertible $n \times n$ matrix and $A_i(k)$ is an $n \times n$ matrix for $1 \le k \le h$;
- $C(k)$ is an $m \times n$ matrix for $1 \le k \le h$;
- $\Xi(k)$ is an $n \times n$ diagonal matrix for $1 \le k \le h$.

The initial conditions $y_0, \ldots, y_{1-\rho}$ are taken as given. Let

$$\underset{\rho n \times 1}{x_t} = \begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-\rho} \\ z_t \end{bmatrix} \quad \text{and} \quad \underset{(\rho n+m)\times n}{F(s_t)} = \begin{bmatrix} A_1(s_t) \\ \vdots \\ A_\rho(s_t) \\ C(s_t) \end{bmatrix}.$$

Then (18) can be written in the compact form:

$$y_t' A(s_t) = x_t' F(s_t) + \varepsilon_t' \Xi^{-1}(s_t), \quad \text{for } 1 \le t \le T. \tag{19}$$

We introduce the following notation that will be used repeatedly later in this paper:

$$A = \{A(1), \ldots, A(h)\}, \qquad F = \{F(1), \ldots, F(h)\},$$
$$\Xi = \{\Xi(1), \ldots, \Xi(h)\}, \qquad \theta = \{A, F, \Xi\},$$
$$\underset{t\times n}{Y_t} = \begin{bmatrix} y_1' \\ \vdots \\ y_t' \end{bmatrix}, \qquad \underset{t\times m}{Z_t} = \begin{bmatrix} z_1' \\ \vdots \\ z_t' \end{bmatrix}, \qquad \underset{(t+1)\times 1}{S_t} = \begin{bmatrix} s_0 \\ \vdots \\ s_t \end{bmatrix}.$$

We assume that

$$p(\varepsilon_t \mid Y_{t-1}, Z_t, S_t, \theta, w) = \text{normal}(\varepsilon_t \mid 0_n, I_n),$$

where 0_n denotes an $n \times 1$ vector of zeros, I_n denotes the $n \times n$ identity matrix, and $\text{normal}(x \mid \mu, \Sigma)$ denotes the multivariate normal distribution of x with mean μ and variance Σ. This assumption is equivalent to

$$p(y_t \mid Y_{t-1}, Z_t, S_t, \theta, w) = \text{normal}(y_t \mid \mu_t(s_t), \Sigma(s_t)), \tag{20}$$

where w is a vector of free parameters in the transition matrix as discussed in Section 4,

$$\mu_t(k) = \left(F(k) A^{-1}(k)\right)' x_t,$$

and

$$\Sigma(k) = \left(A(k)\, \Xi^2(k)\, A'(k)\right)^{-1}.$$

For $1 \le k \le h$, let $a_j(k)$ be the jth column of $A(k)$, $f_j(k)$ be the jth column of $F(k)$, and $\xi_j(k)$ be the jth diagonal element of $\Xi(k)$. Define

$$\underset{n^2\times 1}{a(k)} = \begin{bmatrix} a_1(k) \\ \vdots \\ a_n(k) \end{bmatrix}, \qquad \underset{(\rho n+m)n\times 1}{f(k)} = \begin{bmatrix} f_1(k) \\ \vdots \\ f_n(k) \end{bmatrix}, \quad \text{and}$$

$$\underset{n\times 1}{\xi(k)} = \begin{bmatrix} \xi_1(k) \\ \vdots \\ \xi_n(k) \end{bmatrix}.$$

For expositional clarity, we focus on the composite Markov process $s_t = (s_{1t}\ s_{2t})$ where s_{1t} and s_{2t} are independent regime variables, although the analytical results for more complicated Markov processes can be derived similarly. We let a_j and f_j depend on s_{1t} and ξ_j depend on s_{2t}. It follows from (20) that the conditional likelihood function $p(y_t \mid Y_{t-1}, Z_t, S_t, \theta, w)$ is equal to

$$|A(s_{1t})| \prod_{j=1}^{n} \left|\xi_j(s_{2t})\right| \exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(y_t' a_j(s_{1t}) - x_t' f_j(s_{1t})\right)^2\right). \tag{21}$$

Given (21), the overall likelihood of Y_T can be formed by following (10).

8.2. Restrictions on time variation

If we let all parameters vary across regimes, the number of free parameters in the model becomes impractically high when the system of equations is large or the lag length is long. For a typical quarterly model with 5 lags and 6 endogenous variables, for example, the number of parameters in $F(s_{1t})$ is of order 180 for each regime. Given the post-war macroeconomic data, however, it is not uncommon to have some regimes lasting for only a few years and thus the number of relevant observations is far less than 180 quarters. It is therefore essential to simplify the model by restricting the degree of time variation in the model's parameters. Such a restriction entails complexity and difficulties that have not been dealt with in the simultaneous-equation literature.

To begin with, we rewrite F as

$$\underset{m\times n}{F(s_{1t})} = \underset{m\times n}{G(s_{1t})} + \underset{m\times n}{\bar{S}}\ \underset{n\times n}{A(s_{1t})}. \tag{22}$$

where

$$\bar{S} = \begin{bmatrix} \mathbf{I}_n \\ \mathbf{0}_{(m-n)\times n} \end{bmatrix}.$$

We let G be a collection of all $G(k)$ for $k = 1, \ldots, h_1$. If the prior distribution on $G(s_{1t})$ has mean zero, the specification of $\bar{S}$ is consistent with the reduced-form random walk feature implied by the existing Bayesian VAR models (Sims and Zha, 1998). This type of prior tends to imply that greater persistence (in the sense of a tighter concentration of the prior on the random walk) is associated with smaller disturbance variances. This feature is

reasonable, as it is consistent with the idea that beliefs about the unconditional variance of the data are *not* highly correlated with beliefs about the degree of persistence in the data.

Let $g_j(k)$ be the jth column of $G(k)$. The time-variation restrictions imposed on $g_j(k)$ can be generally expressed by two components, one being time varying and the other being constant across regimes. Denote the first component by the $r_{g,j} \times 1$ vector $g_{\delta_j(k)}$ and the second component by the $h_1 r_{g,j} \times 1$ vector g_{ψ_j}, where the subscripts $\delta_j(k)$ and ψ_j will be discussed further in Section 8.3. We express $g_j(k)$ for $k = 1, \ldots, h_1$ in the form

$$\begin{aligned}&\operatorname{diag}\left(\begin{bmatrix} g_j(1)' & \cdots & g_j(h_1)' \end{bmatrix}'\right)\\ &= \operatorname{diag}\left(\begin{bmatrix} g'_{\delta_j(1)} & \cdots & g'_{\delta_j(h_1)} \end{bmatrix}'\right)\operatorname{diag}\left(g_{\psi_j}\right), \end{aligned} \tag{23}$$

where $\operatorname{diag}(x)$ is the diagonal matrix with the diagonal being the column vector x. The long vector g_{ψ_j} is formed by stacking h_1 sub-vectors and the kth sub-vector corresponds to the parameters in the kth regime.

In this paper, we focus on the following three cases of restricted time variations in the jth equation for $a_j(s_{1t})$ and $g_j(s_{1t})$ where $j \in \{1, \ldots, n\}$, although our general method is capable of dealing with other time variation cases.

$$\begin{aligned}&a_j(s_{1t})\,\xi_j(s_{2t}),\ g_{ij,\ell}(s_{1t})\,\xi_j(s_{2t}),\ c_j(s_{1t})\,\xi_j(s_{2t})\\ &= \begin{cases} a_j,\ g_{ij,\ell},\ c_j & \text{Case I}\\ a_j\,\xi_j(s_{2t}),\ g_{ij,\ell}\,\xi_j(s_{2t}),\ c_j\,\xi_j(s_{2t}) & \text{Case II}\\ a_j(s_{1t})\,\xi_j(s_{2t}),\ g_{\psi_{ij,\ell}}\,g_{\delta_{ij}(s_{1t})}\,\xi_j(s_{2t}),\ c_j(s_{1t})\,\xi_j(s_{2t}) & \text{Case III,}\end{cases}\end{aligned} \tag{24}$$

where $g_{ij,\ell}(s_{1t})$ is the element of $g_j(s_{1t})$ for the ith variable at the ℓth lag and $c_j(s_{1t})$ is a vector of parameters corresponding to the exogenous variable vector z_t in equation j. The parameter $g_{\psi_{ij,\ell}}$ is the element of g_{ψ_j} for the ith variable at the ℓth lag in any regime; it is constant across regimes. The parameter $g_{\delta_{ij}(s_{1t})}$ is the element of $g_{\delta_j(s_{1t})}$ for the ith variable in regime s_{1t} at any lag. Thus, when the regime s_{1t} changes, $g_{\delta_{ij}(s_{1t})}$ changes with variables but does not vary across lags. The variability across variables when the regime changes is necessary to allow long run responses to vary across regimes, while the restriction on the time variation across lags is essential to prevent over-parameterization. The parameters a_j, $g_{ij,\ell}$, and c_j without the symbol (s_{1t}) mean that these parameters are independent of regime (i.e., constant across time).

In this setup, we include $c_j(k)$ in the stacked column vector g_{ψ_j}. This parameterization of grouping $c_j(k)$ in g_{ψ_j} preserves the prior correlations between $c_j(k)$ and the other lagged coefficients as implied by the Sims and Zha (1998) dummy-observation prior, an important part of the prior specification. Note that the other elements of g_{ψ_j} are restricted to be independent of regime.

Case I represents a traditional constant-parameter VAR equation, which has been dealt with extensively in the literature and thus will not be a focal discussion of this paper. Case II represents a structural equation with only shock variances changing regime. In this case, $\xi_j(s_{2t})$ measures the volatility of the shock process in the jth structural equation. Case III represents a structural equation with both time-varying coefficients and heteroscedastic disturbances.

8.3. Identifying restrictions

It is well known that the model (19) would be unidentified without further identifying restrictions. We follow Waggoner and Zha (2003a) and apply linear restrictions on A and F in the form of

$$\mathfrak{R}_j \begin{bmatrix} a_j \\ f_j \end{bmatrix} = 0, \tag{25}$$

where $\mathfrak{R}_j$ is an $(n + \rho n + m) \times (n + \rho n + m)$ and is not of full rank. This class of restrictions is general enough to encompass restrictions used in the VAR literature (Rudebusch and Svensson, 1999; George et al., 2008); it can be used to deal with over-parameterization. It follows from (25) that

$$a_j(k) = U_j b_j(k), \tag{26}$$

$$f_j(k) = V_j g_j(k) - W_j U_j b_j(k), \tag{27}$$

where U_j is an $n \times q_j$ matrix with orthonormal columns, V_j is a $(\rho p + m) \times r_j$ matrix with orthonormal columns, and W_j is a $(\rho p + m) \times n$ matrix (see Appendix D for details). To make (27) agreeable to the random walk form given by (22), the restrictions on the first-lag coefficient matrix A_1 must be a subset of those on the contemporaneous coefficient matrix A_0, and in this case we can take W_j to be $\bar{S}$.

From (21), (26) and (27), one can rewrite the likelihood as

$$\begin{aligned}&p(y_t \mid Y_{t-1}, Z_t, S_t, \theta, w)\\ &= |A(s_{1t})|\left[\prod_{j=1}^{n} |\xi_j(s_{2t})| \exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(\left(y_t' + x_t' W_j\right) U_j b_j(s_{1t}) - x_t' V_j g_j(s_{1t})\right)^2\right)\right].\end{aligned} \tag{28}$$

In addition to the time-variation restrictions (24), the lagged coefficient vector $g_j(k)$ for $k \in \{1, \ldots, h_1\}$ may be further restricted. Specifically, one may impose linear restrictions directly on $g_{\delta_j(k)}$ and g_{ψ_j} through the affine transformation from $\mathbb{R}^{r_{\delta,j}}$ to $\mathbb{R}^{r_{g,j}}$

$$g_{\delta_j(k)} = \Delta_j \delta_j(k) + \bar{\delta}_j \tag{29}$$

and the affine transformation from $\mathbb{R}^{r_{\psi,j}}$ to $\mathbb{R}^{h_1 r_{g,j}}$

$$g_{\psi_j} = \Psi_j \psi_j, \tag{30}$$

where Δ_j is an $r_{g,j} \times r_{\delta,j}$ matrix, Ψ_j is an $h_1 r_{g,j} \times r_{\psi,j}$ matrix, $\bar{\delta}_j$ is an $r_{g,j} \times 1$ vector, $\delta_j(k)$ is an $r_{\delta,j} \times 1$ vector, and ψ_j is an $r_{\psi,j} \times 1$ vector. The vectors $\delta_j(k)$ and ψ_j are the free parameters to be estimated, while the other vectors and matrices on the right hand sides of (29) and (30) are given by the linear restrictions. We assume without loss of generality that Δ_j and Ψ_j have orthonormal columns so that both $\Delta_j'\Delta_j$ and $\Psi_j'\Psi_j$ are identity matrices.

Consider the most common situation in which the constant term is the only exogenous variable. As implied by (24), $r_{\delta,j}$ is much smaller than $r_{g,j}$ so that the time varying component has a small dimension. Similarly, the dimension $r_{\psi,j}$ is much smaller than $h_1 r_{g,j}$. For Case II, we set $\Delta_j = \mathbf{0}$ and $\bar{\delta}_j = \mathbf{1}$ where $\mathbf{1}$ denotes a vector of ones. In practice, therefore, there is no free parameter vector $\delta_j(k)$ to deal with. All the sub-vectors in g_{ψ_j} that correspond to different regimes are the same. Thus, the dimension $r_{\psi,j}$ is no greater than $r_{g,j}$. For Case III, we set

$$\bar{\delta}_j = \begin{bmatrix} \mathbf{0}_{n\rho \times 1} \\ 1 \end{bmatrix},$$

where the last element corresponds to the constant term in the jth equation. The first $n\rho$ elements in the kth sub-vector of g_{ψ_j} are restricted to be the same as those elements in any other sub-vector.

8.4. The prior

We begin with the prior imposed directly on $a_j(k)$ and g_{ψ_j}. From this prior we derive the prior on the free parameters $b_j(k)$ and ψ_j, using the linear restrictions represented by (26) and (30).

The prior distributions of $a_j(k)$ and g_{ψ_j} take the Gaussian form:

$$p(a_j(k)) = \text{normal}\left(a_j(k) \mid \mathbf{0}, \bar{\Sigma}_{a_j}\right), \tag{31}$$

$$p(g_{\psi_j}) = \text{normal}\left(g_{\psi_j} \mid \mathbf{0}, \tilde{\Sigma}_{g_{\psi_j}}\right), \tag{32}$$

for $k = 1, \ldots, h_1$ and $j = 1, \ldots, n$, where $\tilde{\Sigma}_{g_{\psi_j}} = I_{h_1} \otimes \tilde{\Sigma}_g$.[12] The prior covariance matrices $\bar{\Sigma}_{a_j}$ and $\tilde{\Sigma}_g$ are the same as the

[12] The notation $\bar{\Sigma}_g$ will be introduced later after an additional prior component is incorporated.

prior covariance matrices specified by Sims and Zha (1998) for the contemporaneous and lagged coefficients in the constant-parameter VAR model. Because these prior covariance matrices are the same across k, $a_j(k)$ has exactly the same prior distribution for different values of k so that k is essentially irrelevant for this prior.[13] In other words, our prior is symmetric across regimes, for a priori knowledge of how they should differ is difficult to obtain through the prior distribution of this kind.

Following Sims and Zha (1998), we incorporate into the prior the $n+1$ "dummy observations" formed from the initial observations as an additional part of the prior. These dummy observations, used as an additional prior component, express widely-held beliefs in unit roots and cointegration in macroeconomic series and play an indispensable role in improving out-of-sample forecast performance. Let Y_d be an $(n+1)\times n$ matrix of dummy observations on the left hand side of system (19) and X_d be an $(n+1)\times m$ matrix of dummy observations on the right hand side such that

$$Y_d A(k) = X_d \left(G_\psi + \bar{S}A(k)\right) + \tilde{E}_d, \tag{33}$$

where G_ψ is a $(pn+m)\times n$ matrix formed from g_{ψ_j} and $\tilde{E}_d$ is an $(n+1)\times n$ matrix of standard normal random variables. If we add the diffuse prior

$$p\left(\mathrm{vec}(A(k))\right) \propto |A(k)|^{-(n+1)}$$

to correct for the degrees of freedom in the overall prior of $A(k)$, it can be shown that combining the dummy equations (33) and the normal prior (31) and (32) leads to the following overall prior[14]:

$$p(a_j(k)) = \text{normal}\left(a_j(k) \mid \mathbf{0}, \bar{\Sigma}_{a_j}\right), \tag{34}$$

$$p(g_{\psi_j}) = \text{normal}\left(g_{\psi_j} \mid \mathbf{0}, \bar{\Sigma}_{g_{\psi_j}}\right), \tag{35}$$

where $\bar{\Sigma}_{g_{\psi_j}} = I_{h_1} \otimes \bar{\Sigma}_g$ and

$$\bar{\Sigma}_g = \left(X_d' X_d + \tilde{\Sigma}_g^{-1}\right)^{-1}.$$

Given the linear restrictions (26) and (30), one can derive from (34) and (35) that the implied prior distribution for $b_j(k)$ and ψ_j is

$$p(b_j(k)) = \text{normal}\left(b_j(k) \mid \mathbf{0}, \bar{\Sigma}_{b_j}\right), \tag{36}$$

$$p(\psi_j) = \text{normal}\left(\psi_j \mid \mathbf{0}, \bar{\Sigma}_{\psi_j}\right), \tag{37}$$

where

$$\bar{\Sigma}_{b_j} = \left(U_j' \bar{\Sigma}_{a_j}^{-1} U_j\right)^{-1},$$

$$\bar{\Sigma}_{\psi_j} = \left(\Psi_j' \bar{\Sigma}_{g_{\psi_j}}^{-1} \Psi_j\right)^{-1}.$$

It has been shown that this kind of prior makes it possible to estimate a very large VAR (Leeper et al., 1996; Robertson and Tallman, 1999; Banbura et al., 2007). Unlike the reduced-form VARs studied by George et al. (2008), the prior imposed directly on the structural parameters A helps avoid the curse of dimensionality as the size of a VAR increases (see Sims and Zha (1998) for details).

To complete the section, we specify the prior distributions for $\delta_j(k)$ and $\xi_j^2(k)$ as follows. The prior distribution of $\delta_j(k)$ is assumed to be normal:

$$p(\delta_j(k)) = \text{normal}\left(\delta_j(k) \mid \mathbf{0}, \bar{\Sigma}_{\delta_j(k)}\right), \tag{38}$$

where $\bar{\Sigma}_{\delta_j(k)} = \sigma_\delta^2 I_{r_{\delta,j}}$ and $I_{r_{\delta,j}}$ is the $r_{\delta,j}\times r_{\delta,j}$ identity matrix. The prior distribution of $\xi_j^2(k)$ is assumed to have the gamma density function:

$$p(\xi_j^2(k)) = \text{gamma}\left(\xi_j^2(k) \mid \bar{\alpha}_j, \bar{\beta}_j\right), \tag{39}$$

where

$$\text{gamma}(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x}.$$

8.5. The posterior distribution

Given the likelihood function (28) and the prior density function (36)–(39), our objective is to obtain the conditional posterior density function $p(\theta \mid Y_T, Z_T, S_T, w)$ by sampling alternately from the following conditional posterior distributions:

$$p(b_j(k) \mid Y_T, Z_T, S_T, G, \Xi, w, b_i(k)), \tag{40}$$

$$p(\delta_j(k) \mid Y_T, Z_T, S_T, A, \Xi, w, \psi_j), \tag{41}$$

$$p(\psi_j \mid Y_T, Z_T, S_T, A, \Xi, w, \delta_j(k)), \tag{42}$$

$$p(\xi_j^2(k) \mid Y_T, Z_T, S_T, A, G, w), \tag{43}$$

where $i \neq j$ and $i = 1, \ldots, n$. The first posterior density (40) is not of any standard form. To sample from this distribution, the Metropolis–Hastings algorithm will be employed. The second and third posterior distributions represented by (41) and (42) are multivariate normal. The fourth posterior density (43) has a gamma distribution. The expressions for these posterior densities are algebraically complicated and are given in Appendix E.

8.6. Normalization

To obtain the accurate posterior distributions of θ or a function of θ such as an impulse response, one must normalize signs of structural equations. Otherwise, the posterior distribution will be symmetric with multiple modes, making statistical inference meaningless. Such normalization is also essential to achieving efficiency in evaluating the marginal data density for model comparison. We choose Waggoner and Zha's (2003b) normalization rule to determine the signs of columns of $A(k)$ and $F(k)$ for any given $k \in \{1, \ldots, h_1\}$. Since our original prior is un-normalized and symmetric around the origin, this prior density must be multiplied by 2^n when the marginal data density is estimated with MCMC draws that are normalized by the rule proposed by Waggoner and Zha (2003b).

There is scale normalization on $\delta_j(k_1)$ and $\xi_j(k_2)$. For this kind of normalization, we impose the restrictions $\delta_j(k_1) = \mathbf{1}_{r_{\delta,j}\times 1}$ and $\xi_j(k_2) = 1$ for $j \in \{1, \ldots, n\}$, $k_1 \in \{1, \ldots, h_1\}$, and $k_2 \in$

[13] In our setup, the regime variable s_{1t} for $A(s_{1t})$ and the regime variable s_{2t} for $\Xi(s_{2t})$ are independently treated. In Sims and Zha (2006), the two regime variables are the same. For the Case II model, therefore, $a_j(k)$ are restricted to be the same for all k's under the Sims and Zha setup and we denote this vector by a_j^*. This restriction implies that the prior covariance matrix for a_j^* differs from $\bar{\Sigma}_{a_j}$. To see this point, consider two standard normal random variables x_1 and x_2. With the restriction $x_1 = x_2$, one can show that

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}' = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}' x^*,$$

where x^* is normally distributed with mean 0 and variance 2. Thus, the distribution of x^* is different from that of x_1 or x_2. By analogy, $a_j(1)$ and $a_j(2)$ can be thought as x_1 and x_2; and a_j^* as x^*. For the examples we have studied, it turns out that the prior under our current setup gives a higher marginal data density with the hyperparameter values suggested by Sims and Zha (1998) and Robertson and Tallman (1999, 2001).

[14] The proof follows directly from the fact that

$$(X_d'X_d + \bar{\Sigma}_{g_{\psi_j}}^{-1})^{-1}(X_d'Y_d + \bar{\Sigma}_{g_{\psi_j}}^{-1}\bar{S}) = \bar{S},$$

$$Y_d'Y_d + \bar{\Sigma}_{a_j}^{-1} + \bar{S}'\bar{\Sigma}_{g_{\psi_j}}^{-1}\bar{S} - \overline{\Sigma}_{0j}^{-1} = \bar{\Sigma}_{a_j}^{-1},$$

where

$$\overline{\Sigma}_{0j}^{-1} = (Y_d'X_d + \bar{S}'\bar{\Sigma}_{g_{\psi_j}}^{-1})(X_d'X_d + \bar{\Sigma}_{g_{\psi_j}}^{-1})^{-1}(X_d'Y_d + \bar{\Sigma}_{g_{\psi_j}}^{-1}\bar{S}).$$