METHODS FOR INFERENCE IN LARGE MULTIPLE-EQUATION MARKOV-SWITCHING MODELS

CHRISTOPHER A. SIMS, DANIEL F. WAGGONER, AND TAO ZHA

ABSTRACT. Inference for hidden Markov chain models in which the structure is a multiple-equation macroeconomic model raises a number of difficulties that are not as likely to appear in smaller models. One is likely to want to allow for many states in the Markov chain, without allowing the number of free parameters in the transition matrix to grow as the square as the number of states, but also without losing a convenient form for the posterior distribution of the transition matrix. Calculation of marginal data densities for assessing model fit is often difficult in high-dimensional models, and seems particularly difficult in these models. This paper gives a detailed explanation of methods we have found to work to overcome these difficulties. It also makes suggestions for maximizing posterior density and initiating MCMC simulations that provide some robustness against the complex shape of the likelihood in these models. These difficulties and remedies are likely to be useful generally for Bayesian inference in large time series models. The paper includes some discussion of model specification issues that apply particularly to structural VAR's with Markov-switching structure.

I. INTRODUCTION

This paper extends the methods of Hamilton (1989), Chib (1996), and Kim and Nelson (1999) to multiple equation models. In such large models, a variety of modeling choices, not needed in smaller models, are required to control dimensionality. We provide suggestions for ways to keep these models tractable. Some of the suggestions are specific to structural VAR's, but some apply more generally.

Date: 19 October 2006.

Key words and phrases. Volatility, coefficient changes, discontinuous shifts, Lucas critique, independent Markov processes.

We thank Tim Cogley, John Geweke, Michel Juillard, Ulrich Mueller, and Frank Schorfheide for helpful discussions and comments. Eric Wang provided excellent research assistance in computation on the Linux operating system. We acknowledge the technical support on parallel and grid computation from Computing College of Georgia Institute of Technology. The views expressed herein do not necessarily reflect those of the Federal Reserve Bank of Atlanta or the Federal Reserve System.

The first part of the paper considers a large class of restrictions on the parameters in the transition matrix. 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 interface that is straightforward for researchers to automate potentially complex restrictions by simply expressing them in a convenient matrix form. A number of examples are employed to illustrate how such an interface matrix can be formed.

The second part of the paper describes a general structural VAR Markov-switching framework that allows four key elements: (1) simultaneity, (2) over-identifying restrictions on both contemporaneous coefficients and lag structure, (3) switching among regimes for the residual covariance matrix independently from switching among regimes for equation coefficients and (4) switching among regimes for coefficients in one structural equation (e.g., monetary policy) independently from switching among those for coefficients in other equations. Our framework is particularly useful in addressing questions related to the current debate on whether monetary policy and the private sector's behavior have significantly changed in recent history, and indeed most of the methods described here were either applied in Sims and Zha (2006) or are extensions of methods that were applied in that paper.[1]

When one evaluates marginal data densities using the Modified Harmonic Means (MHM) method, a typical choice of a weighting function is a Gaussian density function 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 very poor approximation. We propose a more general weighting function that aims at dealing with the non-Gaussian shape of the posterior distribution. We show that our new weighting function works well for the high-dimensional models studied by this paper.

The rest of the paper is organized as follows.

Section II develops a method for estimating Markov-switching models with a certain class of linear restrictions on transition matrices. This class includes restrictions that apply when there are independently evolving states, as well as other forms of restriction that are likely to prove useful in applications.

[1]For the debate on monetary history, consult Cogley and Sargent (2002), Canova and Gambetti (2004), Beyer and Farmer (2004), Cogley and Sargent (2005), Primiceri (2005), and Sims and Zha (2006).

Section III develops tools for estimation and inference of both identified and unrestricted switching vector autoregression (VAR) models with transition matrices satisfying restrictions in this class.

In Section IV, we describe a block-wise optimization method for estimating these models. The method proves, in this application, to be much more computationally efficient than the expectation-maximization (EM) algorithm, which has been widely used in similar, but smaller, models.

In Section V, we show that the usual implementation of the Modified Harmonic Mean method (MHM) for calculating marginal data densities runs into severe difficulties in these models, and we suggest a variation on the MHM method that works much better.

A three-variable VAR application to the post-war US data is presented in Section VI.

And Section VII concludes.

II. Markov-Switching Model

II.1. **Distributional assumptions.** Let $(Y_t, Z_t, \theta, Q, S_t)$ be a collection of random variables where

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

and H is a finite set with h elements and is usually taken to be the set $\{1, \cdots, h\}$. The vector y_t contains the endogenous variables and the vector z_t contains the exogenous variables. Our analysis, however, encompass the 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

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

We shall follow the convention that if u and v are random vectors for which a density function exists, $p(u,v)$ denotes the density function. The marginal and conditional density functions are expressed as

$$p(v) = \int p(u,v)\,du,$$

and

$$p(u \mid v) = \frac{p(u,v)}{\int p(u,v)\,du}.$$

We assume that $p(u,v)$ is integrable. Hence, $p(u \mid v)$ and $p(v)$ will exist for almost all v. The objects θ and Q are parameters, Y_t and Z_t are observed data, and S_t can be considered either a sequence of unobserved 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[2] 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.

$$p(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, S_{t-1}) = q_{s_t, s_{t-1}}$$

for $t > 0$.

Condition 2.

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

for $t > 0$.

Condition 3.

$$p(z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, S_t) = p(z_t \mid Y_{t-1}, Z_{t-1}).$$

Condition 1 states formally that the sequence S_t evolves according to an exogenous Markov process with the transition matrix Q. Condition 2 is needed for obtaining a standard posterior density function of Q conditional on S_T.[3] Condition 3 ensures that z_t is an exogenous variable.

[2]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 $(\mathbb{R}^m)^t$.

[3]This tractable result no longer holds for most regime-switching rational expectations models (Farmer, Waggoner, and Zha, 2006). In that case, the Metropolis algorithm may be used instead. We thank Tim Cogley for bringing our attention to this point.

II.2. **Propositions.** From Conditions 1 - 3, one can prove the following propositions (the proofs can be found in Hamilton (1989), Chib (1996), and Kim and Nelson (1999)). These propositions are used throughout the rest of this paper.

Proposition 1.

$$p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right) = \sum_{s_{t-1} \in H} q_{s_t, s_{t-1}} p\left(s_{t-1} \mid Y_{t-1}, Z_{t-1}, \theta, Q\right)$$

for $t > 0$.

Proposition 2.

$$p\left(s_t \mid Y_t, Z_t, \theta, Q\right) = \frac{p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right)}{\sum_{s_{t-1} \in H} p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right)}$$

for $t > 0$.

Proposition 3.

$$p\left(s_t \mid Y_t, Z_t, \theta, Q, s_{t+1}\right) = p\left(s_t \mid Y_T, Z_T, \theta, Q, S_{t+1}^T\right)$$

for $0 \leq t < T$.

Proposition 4.

$$p\left(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, S_T\right) = \left(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, S_t\right)$$

for $0 < t \leq T$.

II.3. **Restrictions on Q.** An important part of this paper is to consider a wide range of restrictions on Q while maintaining the standard form of its posterior probability density function. We first consider a general class of linear restrictions. This class includes exclusion restrictions, fixing certain transition probabilities at a known non-zero constant, and keeping certain transition probabilities proportional to one another. The second class of restrictions is nonlinear and involves a tensor product of transition matrices to allow for independent Markov processes.

II.3.1. *Linear restrictions on Q.* For $1 \leq j \leq h$, let q_j be the j^{th} column of Q and let q be an h^2-dimensional column vector stacking these q_j's. If Q is unrestricted, the likelihood as a function of q_j is proportional to a Dirichlet density. The same is true of the posterior if the prior on q_j is of Dirichlet and the initial distribution on s_0 does not depend on q. We shall consider linear restrictions on q that preserve this property.

For $1 \leq j \leq v$, let w_j be a d_j-dimensional vector, where v may be greater or less than h and the elements of w_j are non-negative and sum to one. Let w be a d-dimensional column vector stacking w_j's, where $d = \sum_{j=1}^{v} d_j$. We describe the linear restrictions on q by

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

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

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

$M_{i,j}$ is an $h \times d_j$ matrix and satisfies the following two conditions.

Condition 4. Let $\lambda_{i,j,r}$ be the sum of the elements in any column of $M_{i,j}$, where the column is indexed by $r \in \{1, \ldots, d_j\}$. Then, $\sum_{j=1}^{v} \lambda_{i,j,r} = 1$.

Condition 5. All the elements of M are non-negative and each row of M has at most one non-zero element.

Condition 4 is necessary to ensure that the elements of q_j sum to one. Condition 5 ensures that the elements of q_j are positive and 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 that $d_j \leq h$ and $d \leq h^2$. Our class of restrictions on Q encompasses most examples discussed in the literature.

Clearly one could work 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 4 and 5. However, if restrictions are complicated and a researcher does not want to derive and code up the posterior density of the free elements in the transition matrix each time when a new application is studied, the setup (1) provides a way to automate the handling of different kinds of restrictions in one convenient framework. Furthermore, when

the researcher chooses to use our computer program, the general-purpose interface matrix M in (1) as one of inputs for the program becomes very handy and easy to implement.[4] In the following we illustrate how to construct the transformation matrix M for a number of useful examples. Some 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 states increases.

Example 1. Sims (1999) discusses a structural break with an irreversible regime change. In a two-state case where the second state is absorbing or irreversible, we have

$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where $v = 2$, $d_1 = 2$, and $d_2 = 1$. In general, exclusion restrictions of the form $q_{i,j} = 0$ require that the $(h(j-1)+i)^{\text{th}}$ row of M_j be zero.

Example 2. A symmetric jumping among states considered by Sims (2001) introduces a parsimonious parameterization of Q to avoid over-parameterization. The transition matrix studied by Sims (2001) has the following form

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

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

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

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

Example 3. Consider a three-state example where the third state is irreversible. A transition to this absorbing state occurs only from the second state and the transition probability is

[4]The software is available at http://home.earthlink.net/ tzha02/ProgramCode/programCode.html.

$1/4$. It follows that the transition matrix is of the form

$$\begin{bmatrix} q_{1,1} & q_{1,2} & 0 \\ q_{2,1} & q_{2,2} & 0 \\ 0 & 1/4 & 1 \end{bmatrix}.$$

This example is used to show how to implement exclusion restrictions and, more generally, how to handle the case in which some of the transition probabilities are known. To put these restriction in the matrix form of M, let $v = 3$, $d_1 = 2$, $d_2 = 2$, and $d_3 = 1$. The 9×5 matrix M is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 0 & 0 \\ 0 & 0 & 0 & 3/4 & 0 \\ 0 & 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Example 4. This example pertains to incremental changes in the model parameters (Cogley and Sargent, 2005).[5] This kind of parameter drift can be approximated arbitrarily well by expanding the number of states while containing the elements of Q in a much smaller dimension. 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. One way to achieve this task is to concentrate weight on the diagonal of Q (Zha, In press). Specifically, one can express incremental increases and discontinuous jumps

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

among $n+1$ states as

$$Q = \begin{bmatrix} \pi_1 & \beta_2\alpha_2(1-\pi_2) & \dots & \beta_{n+1}\alpha_{n+1}^{n}(1-\pi_{n+1}) \\ \beta_1\alpha_1(1-\pi_1) & \pi_2 & \dots & \beta_{n+1}\alpha_{n+1}^{n-1}(1-\pi_{n+1}) \\ \beta_1\alpha_1^2(1-\pi_1) & \beta_2\alpha_2(1-\pi_2) & \dots & \beta_{n+1}\alpha_{n+1}^{n-2}(1-\pi_{n+1}) \\ \dots & \dots & \dots & \dots \\ \beta_1\alpha_1^{n}(1-\pi_1) & \beta_2\alpha_2^{n-1}(1-\pi_2) & \dots & \pi_{n+1} \end{bmatrix},$$

where π_i is a free parameter and $0 < \alpha_i < 1$ is taken as a given. The restrictions can be written as

$$M_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & \beta_1\alpha_1 \\ 0 & \beta_1\alpha_1^2 \\ \dots & \dots \\ 0 & \beta_1\alpha_1^{n} \end{bmatrix}, M_{2,2} = \begin{bmatrix} 0 & \beta_2\alpha_2 \\ 1 & 0 \\ 0 & \beta_2\alpha_2 \\ \dots & \dots \\ 0 & \beta_2\alpha_2^{n-1} \end{bmatrix}, \dots, M_{n+1,n+1} = \begin{bmatrix} 0 & \beta_{n+1}\alpha_{n+1}^{n} \\ 0 & \beta_{n+1}\alpha_{n+1}^{n-1} \\ 0 & \beta_{n+1}\alpha_{n+1}^{n-2} \\ \dots & \dots \\ 1 & 0 \end{bmatrix},$$

where the values of α_i and β_i must be such that elements in each column of $M_{i,i}$ sum up to 1. Note that $v = n+1$, $d_1 = \cdots = d_{n+1} = 2$, and $M_{i,j} = 0$ for $i \neq j$.

Example 5. The above example shows that we can reduce a large number of elements in the transition matrix to free parameters whose dimension is equal to the number of states. The class of linear restrictions specified in (1) enables us to reduce a number of free parameter even further. Consider an $h \times h$ transition matrix Q in the form of

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

This restricted transition matrix implies that when we are in state j, the probability of moving to state $j-1$ or $j+1$ is symmetric and independent of j. Let $v = 1$ and $d_1 = 2$. We

can express this restriction as

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

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

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

In general, our setup is flexible enough to handle more elaborate cases where the jumping probabilities are not symmetric or independent or where a variable jumps from a state to nearby (but not adjacent) states.

Example 6. The original approach of Hamilton (1989) makes it explicit for the model parameters to depend on not only the current state but also the previous state. This dependence on the past state can be easily modelled in our framework. Suppose the original state variable, denoted by s_t^o, takes on two values and has the transition matrix $P = \left(p_{i,j}\right)$. Let the composite state variable, $s_t = \{s_t^o, s_{t-1}^o\}$, consist of a pair of current and previous states. There will be four possibilities for s_t and the overall transition matrix Q must be of the form

		(s_{t-1}, s_{t-2})			
		$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
(s_t, s_{t-1})	$(1,1)$	$p_{1,1}$	$p_{1,1}$	0	0
	$(1,2)$	0	0	$p_{1,2}$	$p_{1,2}$
	$(2,1)$	$p_{2,1}$	$p_{2,1}$	0	0
	$(2,2)$	0	0	$p_{2,2}$	$p_{2,2}$

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

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

II.3.2. *Independent Markov processes.* We now consider the case in which there are κ independent Markov processes. Let $h = \prod_{k=1}^{\kappa} h^k$, $H = \prod_{k=1}^{\kappa} H^k$ where $H^k = \left\{1, \cdots, h^k\right\}$, and $s_t = \left(s_t^1, \cdots, s_t^\kappa\right)$ where $s_t^k \in H^k$. The transition matrix Q is therefore restricted to the form

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

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

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

The tensor product representation of Q implies that if $i = \left(i^1, \cdots, i^\kappa\right) \in H$ and $j = \left(j^1, \cdots, j^\kappa\right) \in H$, then $q_{i,j} = \prod_{k=1}^{\kappa} q_{i^k,j^k}^k$. Conditional on Q, the composite (overall) Markov process s_t consists of κ independent Markov processes s_t^k. If Q were not restricted to this tensor product representation, then it would contain $\left(\prod_{k=1}^{\kappa} h^k\right)\left(\prod_{k=1}^{\kappa} h^k - 1\right)$ parameters. With this nonlinear restriction, there are only $\sum_{k=1}^{\kappa} h^k\left(h^k - 1\right)$ parameters – a substantial reduction.

One can combine the two types of restrictions by imposing the linear 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 nonnegative 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 $d^k = \sum_{j=1}^{v^k} d_j^k$, and M^k be a $\left(h^k\right)^2 \times d^k$ matrix satisfying Conditions 4 and 5. It follows from Section II.3.1 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 which of the independent Markov state variables is under consideration. It is important to remember, however, that all of the results apply to a product of independent Markov state variables by simply adding the superscript k in appropriate places.

II.4. **Prior.** In this section we describe the prior on all the model parameters. We begin with the prior on Q if Q is unrestricted. 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.

We now consider the restricted transition matrix Q as discussed in Section II.3.1. Denote $w_j = \left[w_{1,j}, \cdots, w_{d_j,j}\right]'$. The prior on w_j is of the Dirichlet form

$$\frac{\Gamma\left(\sum_{i=1}^{d_j} \beta_{i,j}\right)}{\prod_{i=1}^{d_j} \Gamma\left(\beta_{i,j}\right)} \prod_{i=1}^{d_j} \left(w_{i,j}\right)^{\beta_{i,,j}-1} \tag{4}$$

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

The joint prior density for θ, Q, S_T is

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

By Condition 1, $p(s_t \mid \theta, Q, S_{t-1}) = q_{s_t, s_{t-1}}$. We assume that the prior on θ is independent of the prior on Q and that $p(s_0 \mid \theta, Q) = \frac{1}{h}$ for every $s_0 \in H$.[6] The resulting prior has the following form

$$p(\theta, Q, S_T) = \frac{p(\theta)\, p(Q)}{h} \prod_{t=1}^{T} q_{s_t, s_{t-1}}. \tag{5}$$

II.5. **Likelihood.** Using Proposition 4 and Conditions 2 and 3, one can show that the joint density of Y_T and Z_T conditional on θ and Q is

$$p(Y_T, Z_T \mid \theta, Q) = \prod_{t=1}^{T} p(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q)$$

Note

$$\begin{aligned} p(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q) &= \sum_{s_t \in H} p(y_t, z_t, s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q) \\ &= \sum_{s_t \in H} p(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, s_t)\, p(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q) \end{aligned}$$

[6] The conventional assumption for $p(s_0 \mid \theta, Q)$ is the ergodic distribution of Q, if it exists. This convention, however, makes the conditional posterior distribution of Q an unknown and complicated one.

and

$$
\begin{aligned}
&p\left(y_t, z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, s_t\right) \\
&\qquad = p\left(y_t \mid Y_{t-1}, Z_t, \theta, Q, s_t\right) p\left(z_t \mid Y_{t-1}, Z_{t-1}, \theta, Q, s_t\right) \\
&\qquad = p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(z_t \mid Z_{t-1}\right),
\end{aligned}
$$

it follows that

$$
\begin{aligned}
p\left(Y_T, Z_T \mid \theta, Q\right) &= \prod_{t=1}^{T} p\left(z_t \mid Z_{t-1}\right) \prod_{t=1}^{T} \left[\sum_{s_t \in H} p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right) \right] \\
&= p\left(Z_T\right) \prod_{t=1}^{T} \left[\sum_{s_t \in H} p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right) \right]
\end{aligned}
$$

Conditional on the vector of exogenous variables Z_t, the likelihood of Y_T is

$$
p\left(Y_T \mid Z_T, \theta, Q\right) = \prod_{t=1}^{T} \left[\sum_{s_t \in H} p\left(y_t \mid Y_{t-1}, Z_t, \theta, s_t\right) p\left(s_t \mid Y_{t-1}, Z_{t-1}, \theta, Q\right) \right] \tag{6}
$$

This likelihood can be evaluated recursively, using Propositions 1 and 2.

II.6. **Posterior distribution.** By the Bayes rule, it follows from (5) and (6) that the posterior distribution of (θ, Q) is

$$
p(\theta, Q \mid Y_T, Z_T) \propto p(\theta, Q) p(Y_T \mid Z_T, \theta, Q). \tag{7}
$$

The posterior density $p(\theta, Q \mid Y_T, Z_T)$ is unknown and complicated; the MCMC simulation directly from this distribution can be inefficient and problematic. One can, however, use the idea of Gibbs sampling to obtain the empirical joint posterior density $p(\theta, Q, S_T \mid Y_T, Z_T)$ by sampling alternately from the following conditional posterior distributions:

$$
\begin{aligned}
&p(S_T \mid Y_T, Z_T, \theta, Q), \\
&p(Q \mid Y_T, Z_T, S_T, \theta), \\
&p(\theta \mid Y_T, Z_T, Q, S_T).
\end{aligned}
$$

Simulation from the conditional posterior density $p(\theta \mid Y_T, Z_T, Q, S_T)$ is model-dependent, which we will discuss in Section III. In this section we study the first two conditional posterior distributions.

II.6.1. *Conditional posterior distribution of S_T.* The distribution of S_T conditional on Y_T, Z_T, θ, and Q is

$$\begin{aligned} p(S_T \mid Y_T, Z_T, \theta, Q) &= p(s_T \mid Y_T, Z_T, \theta, Q)\, p\left(S_{T-1} \mid Y_T, Z_T, \theta, Q, S_T^T\right) \\ &= p(s_T \mid Y_T, Z_T, \theta, Q) \prod_{t=0}^{T-1} p\left(s_t \mid Y_T, Z_T, \theta, Q, S_{t+1}^T\right) \end{aligned}$$

where $S_{t+1}^T = \{s_{t+1}, \cdots, s_T\}$. From Proposition 3,

$$\begin{aligned} p\left(s_t \mid Y_T, Z_T, \theta, Q, S_{t+1}^T\right) &= p(s_t \mid Y_t, Z_t, \theta, Q, s_{t+1}) \\ &= \frac{p(s_t, s_{t+1} \mid Y_t, Z_t, \theta, Q)}{p(s_{t+1} \mid Y_t, Z_t, \theta, Q)} \\ &= \frac{p(s_{t+1} \mid Y_t, Z_t, \theta, Q, s_t)\, p(s_t \mid Y_t, Z_t, \theta, Q)}{p(s_{t+1} \mid Y_t, Z_t, \theta, Q)} \\ &= \frac{q_{s_{t+1}, s_t} p(s_t \mid Y_t, Z_t, \theta, Q)}{p(s_{t+1} \mid Y_t, Z_t, \theta, Q)} \end{aligned}$$

The conditional density $p\left(s_t \mid Y_T, Z_T, \theta, Q, S_{t+1}^T\right)$ is straightforward to evaluate according to Propositions 1 and 2, . Starting with s_T and working backward, we can easily sample S_T from the posterior conditional on Y_T, Z_T, θ, Q by using the following fact

$$\begin{aligned} p(s_t \mid Y_T, Z_T, \theta, Q) &= \sum_{s_{t+1} \in H} p(s_t, s_{t+1} \mid Y_T, Z_T, \theta, Q) \\ &= \sum_{s_{t+1} \in H} p(s_t \mid Y_T, Z_T, \theta, Q, s_{t+1})\, p(s_{t+1} \mid Y_T, Z_T, \theta, Q) \\ &= \sum_{s_{t+1} \in H} p(s_t \mid Y_t, Z_t, \theta, Q, s_{t+1})\, p(s_{t+1} \mid Y_T, Z_T, \theta, Q). \end{aligned}$$

Note that this density can also be evaluated recursively.

II.6.2. *Conditional posterior distribution of Q_k.* The conditional posterior density of Q derives directly from the conditional posterior density of the free parameters w_j.[7] It follows from Condition 1 and the prior (4) that

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

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

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

III. STRUCTURAL VAR MODELS

The methodology developed thus far has been used by Rubio-Ramírez, Waggoner, and Zha (2006) and Sims and Zha (2006) to study a class of simultaneous-equation multivariate dynamic models that are commonly used for policy analysis. In this section, we develop and detail the econometric methods specific to these types of models.

III.1. **Likelihood.** We consider a class of models of the following form:

$$y_t'A(s_t) = \sum_{i=1}^{\rho} y_{t-i}'A_i(s_t) + z_t'C(s_t) + \varepsilon_t'\Xi^{-1}(s_t), \text{ for } 1 \leq t \leq T, \tag{8}$$

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;
- For $1 \leq k \leq h$, $A(k)$ is an invertible $n \times n$ matrix and $A_i(k)$ is an $n \times n$ matrix;
- For $1 \leq k \leq h$, $C(k)$ is an $m \times n$ matrix;
- For $1 \leq k \leq h$, $\Xi(k)$ is an $n \times n$ diagonal matrix.

For the rest of the paper we take the initial conditions $y_0, \cdots, y_{1-\rho}$ as given. Let

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

Then (8) can be written in the compact form:

$$y_t'A(s_t) = x_t'F(s_t) + \varepsilon_t'\Xi^{-1}(s_t), \text{ for } 1 \leq t \leq T \tag{9}$$

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

$$A = \big(A(1),\ldots,A(h)\big),\ F = \big(F(1),\ldots,F(h)\big),\ \Xi = \big(\Xi(1),\ldots,\Xi(h)\big),$$

$$\theta = \big(A,F,\Xi\big),$$

$$\underset{t\times n}{Y_t} = \begin{bmatrix} y_1' \\ \vdots \\ y_t' \end{bmatrix},\ \underset{t\times k}{Z_t} = \begin{bmatrix} z_1' \\ \vdots \\ z_t' \end{bmatrix},\ \underset{(t+1)\times 1}{S_t} = \begin{bmatrix} s_0 \\ \vdots \\ s_t \end{bmatrix}.$$

We assume that

$$p\big(\varepsilon_t \,|\, Y_{t-1}, Z_t, S_T, \theta, Q\big) = \text{normal}\,(\varepsilon_t \,|\, \mathbf{0}, I_n),$$

where $\mathbf{0}$ denotes a vector or matrix of zeros, I_n denotes the $n \times n$ identity matrix, and $\text{normal}\,(x \,|\, \mu, \Sigma)$ denotes the multivariate normal distribution of x with mean μ and variance Σ.[8] This assumption is equivalent to

$$p\big(y_t \,|\, Y_{t-1}, Z_t, S_t, \theta, Q\big) = \text{normal}\big(y_t \,|\, \mu_t\,(s_t), \Sigma\,(s_t)\big) \tag{10}$$

where

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

and

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

Let $a_j\,(k)$ be the j^{th} column of $A\,(k)$, $f_j\,(k)$ be the j^{th} column of $F\,(k)$, and $\xi_j\,(k)$ be the j^{th} diagonal element of $\Xi\,(k)$. Define

$$\underset{n^2\times 1}{a\,(k)} = \begin{bmatrix} a_1\,(k) \\ \vdots \\ a_n\,(k) \end{bmatrix},\ \underset{(pn+m)n\times 1}{f\,(k)} = \begin{bmatrix} f_1\,(k) \\ \vdots \\ f_n\,(k) \end{bmatrix}, \text{ and } \underset{n\times 1}{\xi\,(k)} = \begin{bmatrix} \xi_1\,(k) \\ \vdots \\ \xi_n\,(k) \end{bmatrix}$$

[8] The matrix Σ must be symmetric and non-negative semi-definite.

It follows from (10) that

$$p\left(y_t \,|\, Y_{t-1}, Z_t, S_t, \theta, Q\right) = \left|\Sigma\left(s_t\right)\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(y_t - \mu\left(s_t\right)\right)' \Sigma^{-1}\left(s_t\right)\left(y_t - \mu\left(s_t\right)\right)\right)$$

$$= \left|A\left(s_t\right)\Xi\left(s_t\right)\right| \exp\left(-\frac{1}{2}\left(y_t' A\left(s_t\right) - x_t' F\left(s_t\right)\right)\Xi^2\left(s_t\right)\left(A'\left(s_t\right) y_t - F'\left(s_t\right) x_t\right)\right)$$

$$= \left|A\left(s_t\right)\right| \prod_{j=1}^{n} \left|\xi_j\left(s_t\right)\right| \exp\left(-\frac{\xi_j^2\left(s_t\right)}{2}\left(y_t' a_j\left(s_t\right) - x_t' f_j\left(s_t\right)\right)^2\right).$$

We consider the case where the state variable $s_t = [s_{1t}\, s_{2t}]$ is a composite one such that either $s_{1t} = s_{2t}$ or s_{1t} and s_{2t} are independent random variables. The analytical results for more complicated cases will follow directly. We let a_j and f_j depend on s_{1t} and ξ_j depend on s_{2t}. Thus, the conditional likelihood function $p\left(y_t \,|\, Y_{t-1}, Z_t, S_t, \theta, Q\right)$ is equal to

$$\left|A\left(s_{1t}\right)\right| \prod_{j=1}^{n} \left|\xi_j\left(s_{2t}\right)\right| \exp\left(-\frac{\xi_j^2\left(s_{2t}\right)}{2}\left(y_t' a_j\left(s_{1t}\right) - x_t' f_j\left(s_{1t}\right)\right)^2\right). \tag{11}$$

Given (11), the likelihood of Y_T can be formed by following (6).

III.2. **A priori restrictions.**

III.2.1. *Restrictions on time variation.* If we let all parameters vary across states, 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 state. Given the post-war macroeconomic data, however, it is not uncommon to have some states lasting for only a few years and thus the number of associated 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

$$\underset{m\times n}{F(s_{1t})} = \underset{m\times n}{G(s_{1t})} + \underset{m\times n}{\bar{S}}\,\underset{n\times n}{A(s_{1t})}. \tag{12}$$

where

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

We let G be a collection of all $G(k)$ for $k = 1, \ldots, h_1$. If we place a prior distribution on $G(s_{1t})$ that 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 j^{th} 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 states. 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 III.2.2. We express $g_j(k)$ for $k = 1, \ldots, h_1$ in the form

$$\operatorname{diag}\left(\begin{bmatrix} g_j(1)' & \ldots & g_j(h_1)' \end{bmatrix}'\right) = \operatorname{diag}\left(\begin{bmatrix} g'_{\delta_j(1)} & \ldots & g'_{\delta_j(h_1)} \end{bmatrix}'\right) \operatorname{diag}\left(g_{\psi_j}\right), \tag{13}$$

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 k^{th} sub-vector corresponds to the parameters in the k^{th} state.

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

$$a_j(s_{1t})\,\xi_j(s_{2t}),\, g_{ij,\ell}(s_{1t})\,\xi_j(s_{2t}),\, c_j(s_{1t})\,\xi_j(s_{2t}) =$$

$$\begin{cases} a_j,\, g_{ij,\ell},\, c_j & \text{Case I} \\ a_j\,\xi_j(s_{2t}),\, g_{ij,\ell}\,\xi_j(s_{2t}),\, c_j\,\xi_j(s_{2t}) & \text{Case II} \\ a_j(s_{1t})\,\xi_j(s_{2t}),\, g_{\psi_{ij,\ell}}\, g_{\delta_{ij}(s_{1t})}\,\xi_j(s_{2t}),\, c_j(s_{1t})\,\xi_j(s_{2t}) & \text{Case III} \end{cases}, \tag{14}$$

where $g_{ij,\ell}(s_{1t})$ is the element of $g_j(s_{1t})$ for the i^{th} variable at the ℓ^{th} lag and $c_j(s_{1t})$ is a vector of parameters corresponding to the exogenous variable z_t in equation j. The parameter $g_{\psi_{ij,\ell}}$ is the element of g_{ψ_j} for the i^{th} variable at the ℓ^{th} lag in any state; it is constant across states. The parameter $g_{\delta_{ij}(s_{1t})}$ is the element of $g_{\delta_j(s_{1t})}$ for the i^{th} variable in state s_{1t} at any lag. Thus, when the state s_{1t} changes, $g_{\delta_{ij}(s_{1t})}$ changes with variables but does not vary across lags. The variability across variables when the sate changes is necessary

to allow long run (policy) responses to vary over time, 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 restricted to be independent of state (i.e., constant across time).

In this setup, we include $c_j(k)$ in the stacked column vector g_{ψ_j}. In principle, one could include the time-varying parameter $c_j(k)$ as part of the time-varying component vector $g_{\delta_j(k)}$. With the normalization $c_j(1) = 1$, however, the likelihood function for $c_j(k)$ where $k \geq 2$ is so ill-behaved that our Gibbs sampler fails to work. Moreover, our reparameterization 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. It is important to note that the other elements of g_{ψ_j} are restricted to be invariant to state.

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 time-varying disturbance variances only. In this case, $\xi_j(s_{2t})$ measures volatility for the j^{th} structural equation. Case III represents a structural equation with time-varying coefficients.[9]

There are many applications that derive directly from various combinations of Case II and Case III for different equations. Some combinations, for example, enable one to distinguish regime shifts in policy behavior from their effects on private sector behavior — the practical lesson of the Lucas critique. The model with Case II for all equations suggests no structural break for both policy and the private sector; the model with Case II for the policy equation and Case III for all other equations hypothesizes that the policy rule is stable and structural breaks originate from the private sector. Both of these models, while consistent with rational expectations, take the view that the Lucas critique is unimportant in practice. On the other hand, the model with Case III for all equations is most consistent with the Lucas critique and if found to have a superior fit to the data, suggests that extrapolating the

[9]The reduced-form equation for Case III, however, has both time-varying coefficients and heteroscedastic disturbances. This feature reinforces the point that one should work directly on the structural form, not the reduced-form, of the model.

effects of policy changes from linear approximations may be misleading.[10] The model with Case III for the policy equation and Case II for all other equations is an unconventional but quite interesting hypothesis. It is unconventional because it contradicts many theoretical examples delivered by rational expectations. Yet it implies that the Lucas critique may be practically unimportant because, despite regime shifts in policy, the private sector responds linearly to the history of policy variables.

III.2.2. *Identifying restrictions.* It is well known that the model (9) would be unidentified without further identifying restrictions. We follow the identified VAR literature and apply linear restrictions on A and F in the form of

$$\mathfrak{R}_j \begin{bmatrix} a_j \\ f_j \end{bmatrix} = 0, \tag{15}$$

where $\mathfrak{R}_j$ is an $(n+np+m)\times(n+np+m)$ and is not of full rank. Appendix A shows that the above restrictions are equivalsent to the existence of an $n \times r_{b,j}$ matrix U_j with orthonormal columns, a $(pn+m)\times r_{g,j}$ matrix V_j with orthonormal columns, and a $(pn+m)\times n$ matrix $\hat{W}_j$ with $V_j'\hat{W}_j = 0$ such that

$$a_j(k) = U_j b_j(k), \tag{16}$$

$$f_j(k) = V_j g_j(k) - \hat{W}_j U_j b_j(k). \tag{17}$$

The $r_{b,j}\times 1$ vector $b_j(k)$ and the $r_{g,j}\times 1$ vector $g_j(k)$ are free parameters to be estimated. If we replace $\hat{W}_j$ in (17) with $W_j = \hat{W}_j + V_j\tilde{W}_j$ for any $r_{g,j}\times n$ matrix $\tilde{W}_j$, the underlying linear restrictions (15) will still hold, although $V_j'W_j \neq 0$ in general. For $\bar{S}$ defined in (12), one can show that there exists $\tilde{W}_j$ such that $W_j = \bar{S}$ where

$$\tilde{W}_j = V_j'\left(\bar{S} - \hat{W}_j\right).$$

It follows from (11), (16), and (17) that

$$p\left(y_t \,|\, Y_{t-1}, Z_t, S_t, \theta, Q\right) = \\ |A(s_{1t})| \left[\prod_{j=1}^{n} \left|\xi_j(s_{2t})\right| \exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(\left(y_t' + x_t'W_j\right)U_j b_j(s_{1t}) - x_t'V_j g_j(s_{1t})\right)^2\right)\right]. \tag{18}$$

[10] Theoretical arguments for this view can be found in Cooley, LeRoy, and Raymon (1984), Sims (1987), and more recently Leeper and Zha (2003).

In addition to the time-variation restrictions (14), the lag 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 $\mathbb{R}^{r_{g,j}}$

$$g_{\delta_j(k)} = \Delta_j \delta_j(k) + \bar{\delta}_j \tag{19}$$

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

$$g_{\psi_j} = \Psi_j \psi_j, \tag{20}$$

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 free parameters to be estimated, while the other vectors and matrices on the right hand sides of (19) and (20) are given by 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 (14), $r_{\delta,j}$ 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 or matrix 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 states are the same. Thus, the dimension $r_{\psi,j}$ is no greater than $r_{g,j}$. For Case III, we set

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

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

III.2.3. *The prior.* We begin with a prior imposed directly on $a_j(k)$, g_{ψ_j}, $\delta_j(k)$, and $\xi_j^2(k)$. The prior on the free parameters $b_j(k)$ and ψ_j will then be derived from the linear restrictions (16) and (20).

In order to use the reference prior in the VAR literature, we let the prior distributions of $a_j(k)$ and g_{ψ_j} take the Gaussian form:

$$p(a_j(k)) = \text{normal}\big(a_j(k)\,|\,\mathbf{0}, \bar{\Sigma}_{a_j}\big), \tag{21}$$

$$p(g_{\psi_j}) = \text{normal}\big(g_{\psi_j}\,|\,\mathbf{0}, \tilde{\Sigma}_{g_{\psi_j}}\big), \tag{22}$$

for $k = 1,\ldots,h_1$ and $j = 1,\ldots,n$, where $\tilde{\Sigma}_{g_{\psi_j}} = I_{h_1} \otimes \tilde{\Sigma}_g$. The prior covariance matrices $\bar{\Sigma}_{a_j}$ and $\tilde{\Sigma}_g$ are the same as the 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.[11] In other words, our prior is symmetric across states, 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 also incorporate into the model 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 (9) and X_d be an $(n+1)\times m$ matrix of dummy observations on the right hand side such that

$$Y_d A(k) = X_d\,\big(G_\psi + \bar{S}A(k)\big) + \tilde{E}_d, \tag{23}$$

[11] In our setup, the state variable s_{1t} for $A(s_{1t})$ and the state variable s_{2t} for $\Xi(s_{2t})$ are independently treated. In Sims and Zha (2006), the two state 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 $x_1 = x_2 y$, 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).

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

$$p\big(\text{vec}(A(k))\big) \propto |A(k)|^{-(n+1)}$$

to correct the degrees of freedom for the overall prior of $A(k)$, it can be shown that combining the dummy prior (23) and the normal prior (21)-(22) leads to the following overall prior:[12]

$$p(a_j(k)) = \text{normal}\big(a_j(k)\,|\,\mathbf{0},\bar{\Sigma}_{a_j}\big), \tag{24}$$

$$p(g_{\psi_j}) = \text{normal}\big(g_{\psi_j}\,|\,\mathbf{0},\bar{\Sigma}_{g_{\psi_j}}\big), \tag{25}$$

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 (16) and (20), one can derive from (24) and (25) that the implied prior distribution for $b_j(k)$ and ψ_j is

$$p(b_j(k)) = \text{normal}\big(b_j(k)\,|\,\mathbf{0},\bar{\Sigma}_{b_j}\big), \tag{26}$$

$$p(\psi_j) = \text{normal}\big(\psi_j|\,\mathbf{0},\bar{\Sigma}_{\psi_j}\big), \tag{27}$$

where

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

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

The prior distribution of $\delta_j(k)$ is assumed to be normal:

$$p(\delta_j(k)) = \text{normal}\left(\delta_j(k)\,|\,\mathbf{0},\bar{\Sigma}_{\delta_j(k)}\right), \tag{28}$$

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.

[12]The proof follows directly from the fact (Sims and Zha, 1998) 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

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

The prior distribution of $\xi_j^2(k)$ is assumed to have the gamma density function:

$$p(\xi_j^2) = \gamma\left(\xi_j^2 \,|\, \bar{\alpha}_j, \bar{\beta}_j\right), \tag{29}$$

where

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

III.3. **The posterior distribution.** Given the likelihood function (18) and the prior density function (26)-(29), our objective is to obtain the conditional posterior density function $p(\theta\,|\,Y_T, Z_T, S_T, Q)$ by sampling alternately from the following conditional posterior distributions:

$$p(b_j(k)\,|\,Y_T, Z_T, S_T, G, \Xi, Q, b_i(k)),$$
$$p(\delta_j(k)\,|\,Y_T, Z_T, S_T, A, \Xi, Q, \psi_j),$$
$$p(\psi_j\,|\,Y_T, Z_T, S_T, A, \Xi, Q, \delta_j(k)),$$
$$p(\xi_j^2(k)\,|\,Y_T, Z_T, S_T, A, G, Q),$$

where $i \neq j$ and $i = 1,\ldots,n$. We now discuss each of these four conditional density functions.

III.3.1. *Conditional posterior density of $b_j(k)$.* Combining the likelihood (18) and the prior (26) implies that the posterior density of $b_j(k)$, conditional on S_T, G, Ξ, Q, and $b_i(k)$ for $i \neq j$, is proportional to

$$\exp\left(-\frac{1}{2}b_j'(k)\bar{\Sigma}_{b_j}^{-1}b_j(k)\right)\prod_{t\in\{t:\,s_{1t}=k\}}\left[|A(k)|\exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(y_t'a_j(k)-x_t'f_j(k)\right)^2\right)\right],$$

for $k = 1,\ldots,h_1$. It is important to note that both $a_j(k)$ and $f_j(k)$ are affine functions of $b_j(k)$. To evaluate the above density kernel more efficiently, we sometimes use the following functional form:

$$\exp\left(-\frac{1}{2}b_j'(k)\bar{\Sigma}_{b_j}^{-1}b_j(k)\right)|A(k)|^{T_{1,k}}\times$$
$$\prod_{t\in\{t:\,s_{1t}=k\}}\left[\exp\left(-\frac{1}{2}\left(a_j'(k)\Sigma_{yy,k}a_j(k)-2f_j'(k)\Sigma_{xy,k}a_j(k)+f_j'(k)\Sigma_{xx,k}f_j(k)\right)\right)\right].$$

where $T_{1,k}$ is the number of t's such that $s_{1t} = k$,

$$\Sigma_{yy,k} = \sum_{t \in \{t: s_{1t}=k\}} \xi_j^2(s_{2t})\, y_t y_t',$$

$$\Sigma_{xy,k} = \sum_{t \in \{t: s_{1t}=k\}} \xi_j^2(s_{2t})\, x_t y_t',$$

$$\Sigma_{xx,k} = \sum_{t \in \{t: s_{1t}=k\}} \xi_j^2(s_{2t})\, x_t x_t'.$$

Unlike the constant-parameter simultaneous-equation VAR models studied by Waggoner and Zha (2003a), the above conditional posterior density of $b_j(k)$ is nonstandard. We thus use a Metropolis algorithm with the following proposal density for the transition from $b_j(k)$ to $b_j^\star(k)$

$$p\left(b_j^\star(k) \mid b_j(k), Y_T, Z_T, S_T, b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_n, G, \Xi, Q\right) = \text{normal}\left(b_j^\star(k) \mid \underset{r_{b,j}\times 1}{\mathbf{0}},\ \kappa_{b_j(k)} \Sigma_{b_j(k)}\right) \quad (30)$$

where $b_j^\star(k)$ is a proposal draw, $\kappa_{b_j(k)}$ is a scale factor that can be adjusted to keep the acceptance ratio optimal (e.g., between 25% and 40%), and

$$\Sigma_{b_j(k)}^{-1} = \bar{\Sigma}_{b_j(k)}^{-1} + U_j' \left(\Sigma_{yy,k} + W_j' \Sigma_{xy,k} + \Sigma_{xy,k}' W_j + W_j' \Sigma_{xx,k} W_j\right) U_j.$$

III.3.2. *Conditional posterior densities of $\delta_j(k)$ and ψ_j.* As discussed in Section III.2.2, the long vector g_{ψ_j} is stacked from h_1 sub-vectors. It can be seen from (20) that the restriction matrix Ψ_j can be formed from h_1 corresponding sub-matrices. If we denote

$$g_{\psi_j} = \begin{bmatrix} g_{\psi_{j,1}} \\ \cdots \\ g_{\psi_{j,k}} \\ \cdots \\ g_{\psi_{j,h_1}} \end{bmatrix}, \ \Psi_j = \begin{bmatrix} \Psi_{j,1} \\ \cdots \\ \Psi_{j,k} \\ \cdots \\ \Psi_{j,h_1} \end{bmatrix},$$

we have

$$g_{\psi_{j,k}} = \Psi_{j,k} \psi_j. \quad (31)$$

From the conditional likelihood (18), the prior distribution (28), and the restriction (19), one can obtain the posterior density kernel of $\delta_j(k)$ conditional on S_T, A, Ξ, Q, and ψ_j as

$$\prod_{k=1}^{h_1} \exp\left(-\frac{1}{2}\delta_j(k)'\bar{\Sigma}_{\delta_j(k)}^{-1}\delta_j(k)\right) \times \prod_{t\in\{t:\, s_{1t}=k\}} \exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(\left(y_t' + x_t' W_j\right) U_j b_j(k) - x_t' V_j \mathrm{diag}\left(g_{\psi_{j,k}}\right)\left(\Delta_j \delta_j(k) + \bar{\delta}_j\right)\right)^2\right).$$

Rearranging the terms in the above equation leads to

$$p\left(\delta_j(k) \mid Y_T, Z_T, S_T, A, \Xi, Q, \psi_j\right) = \mathrm{normal}\left(\delta_j(k) \mid \tilde{\mu}_{\delta_j(k)}, \tilde{\Sigma}_{\delta_j(k)}\right), \tag{32}$$

where

$$\begin{aligned}
\hat{\Sigma}_{\delta_j(k)}^{-1} &= \Delta_j' \mathrm{diag}\left(g_{\psi_{j,k}}\right) V_j' \Sigma_{xx,k} V_j \mathrm{diag}\left(g_{\psi_{j,k}}\right) \Delta_j, \\
\tilde{\Sigma}_{\delta_j(k)}^{-1} &= \bar{\Sigma}_{\delta_j(k)}^{-1} + \hat{\Sigma}_{\delta_j(k)}^{-1}, \\
\hat{\mu}_{\delta_j(k)} &= \Delta_j' \mathrm{diag}\left(g_{\psi_{j,k}}\right) V_j' \left[\sum_{t\in\{t:\, s_{1t}=k\}} \xi_j^2(s_{2t}) x_t \left(y_t' + x_t' W_j\right)\right] U_j b_j(k), \\
\tilde{\mu}_{\delta_j(k)} &= \tilde{\Sigma}_{\delta_j(k)}\left(\hat{\mu}_{\delta_j(k)} - \hat{\Sigma}_{\delta_j(k)}^{-1}\bar{\delta}_j\right).
\end{aligned}$$

Similarly, from the conditional likelihood (18), the prior distribution (27), and the restriction (31), we obtain the posterior density kernel of ψ_j conditional on S_T, A, Ξ, Q, and δ_j as

$$\prod_{k=1}^{h_1} \exp\left(-\frac{1}{2}\psi_j'\bar{\Sigma}_{\psi_j}^{-1}\psi_j\right) \times \prod_{t\in\{t:\, s_{1t}=k\}} \exp\left(-\frac{\xi_j^2(s_{2t})}{2}\left(\left(y_t' + x_t' W_j\right) U_j b_j(k) - x_t' V_j \mathrm{diag}\left(g_{\delta_j(k)}\right)\Psi_{j,k}\psi_j\right)^2\right).$$

Rearranging the terms in the above equation gives

$$p\left(\psi_j \mid Y_T, Z_T, S_T, A, \Xi, Q, \delta_j\right) = \mathrm{normal}\left(\delta_j(k) \mid \tilde{\mu}_{\psi_j}, \tilde{\Sigma}_{\psi_j}\right), \tag{33}$$

where

$$\hat{\Sigma}_{\psi_j}^{-1} = \sum_{k=1}^{h_1} \Psi'_{j,k} \text{diag}\left(g_{\delta_j(k)}\right) V_j' \Sigma_{xx,k} V_j \text{diag}\left(g_{\delta_j(k)}\right) \Psi_{j,k},$$

$$\tilde{\Sigma}_{\psi_j}^{-1} = \bar{\Sigma}_{\psi_j}^{-1} + \hat{\Sigma}_{\psi_j}^{-1},$$

$$\hat{\mu}_{\psi_j} = \sum_{k=1}^{h_1} \Psi'_{j,k} \text{diag}\left(g_{\delta_j(k)}\right) V_j' \left[\sum_{t\in\{t:\, s_{1t}=k\}} \xi_j^2(s_{2t})\, x_t \left(y_t' + x_t' W_j\right)\right] U_j b_j(k),$$

$$\tilde{\mu}_{\psi_j} = \tilde{\Sigma}_{\psi_j} \hat{\mu}_{\psi_j}.$$

III.3.3. *Conditional posterior density of* $\xi_j^2(k)$. Let $T_{2,k}$ be the number of elements in $\{t : s_{2t} = k\}$ for $k = 1,\ldots,h_2$. It follows that

$$p\left(\xi_j^2(k) \mid Y_T, Z_T, S_T, A, G, Q\right) = \gamma\left(\xi_j^2(k) \mid \tilde{\alpha}_j(k), \tilde{\beta}_j(k)\right), \tag{34}$$

where

$$\tilde{\alpha}_j(k) = \bar{\alpha}_j + \frac{T_{2,k}}{2},$$

$$\tilde{\beta}_j(k) = \bar{\beta}_j + \frac{1}{2} \sum_{t\in\{t:\, s_{2t}=k\}} \left(y_t' a_j(s_{1t}) - x_t' f_j(s_{1t})\right)^2.$$

III.4. **Other types of Markov processes.** The previous analysis can be easily extended to other types of Markov processes. If we wish to synchronize the two state variables s_{1t} and s_{2t} into one state variable s_t, we simply need to replace these two independent state variables by this one common state variable s_t in the likelihood function. If we wish to have an independent Markov process for the coefficients in each equation, s_{1t} becomes a composite state variable consisting of $s_{j,1t}$ for $j = 1,\ldots,n$. In this case, we simply replace s_{1t} by $s_{j,1t}$ for the time-varying coefficients in equation j in the likelihood function.

III.5. **Normalization.** To obtain the accurate posterior distributions of functions of θ (such as long run responses, historical decompositions, and impulse responses), one must normalize signs of structural equations; otherwise, the posterior distributions will be symmetric with multiple modes, making statistical inferences of interest meaningless. Such normalization is also essential to achieving efficiency in evaluating the marginal data density for model comparison. We choose the Waggoner and Zha (2003b) normalization rule to determine the signs of columns of $A(k)$ and $F(k)$ for any given $k \in \{1,\ldots,h\}$. 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).

Scale normalization on $\delta_j(k_1)$ and $\xi_j(k_2)$ imposes 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 $k_2 \in \{1,\ldots,h_2\}$, where the notation $\mathbf{1}_{r_{\delta,j}\times 1}$ denotes the $r_{\delta,j} \times 1$ vector of 1's. One could use other normalization rules (e.g., restricting each set of time-varying parameters on the unit circle). The marginal data density, however, is invariant to scale normalization, as long as the Jacobian transformation is properly taken into account.

We do not perform any permutation of state-dependent parameters in our MCMC algorithm. For each posterior draw of the parameters, the $h!$ permutations of these parameters give the same posterior density; thus we follow Geweke (2006) and store the $h!$ copies in our MCMC runs conceptually but not literally . In principle, one could normalize the labelling of states as suggested by Hamilton, Waggoner, and Zha (2004) or by Sims and Zha (2006). For the same reason as outlined by Geweke (2006), this labelling does not affect the value of the marginal data density.

IV. Blockwise Optimization Algorithm

In spite of the complexity inherent in the multiple-equation models considered in this paper, it is essential to obtain the estimate of θ at the peak of the posterior distribution (7). The posterior estimate or the maximum likelihood estimate, serving as a starting point for our MCMC algorithm, ensures that an unreasonably long sequence of posterior draws do not get stuck in the low probability region. Used as a reference point in normalization, moreover, it helps avoid distorting the statistical inferences likely to be produced by inappropriate normalization. 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 expectation-maximizing (EM) algorithm for a simple Markov-switching model. For multivariate dynamic models, however, the expectation 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.

As shown in Sims and Zha (2006), these MC simulations can be very expensive computationally. 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 $p(\theta, Q \mid Y_T, Z_T)$ given by (7). Sims (2001) uses this approach for his single-equation model. But for a system of multivariate dynamic equations, the number of model parameters can be too large for a straight maximization routine to be reliable.

In this paper, we propose a different algorithm. We use the Gibbs-sampling idea to break the parameters θ, Q into two blocks of parameters θ and Q. In the multivariate dynamic models considered in this paper, we break the block of parameters θ further into three sub-blocks, one containing $b_j(k)$ for $k = 1, \ldots, h_1$, one containing $g_j(k)$ for $k = 1, \ldots, h_1$, and third sub-block containing $\xi_j^2(k)$ for $k = 1, \ldots, h_2$. Given an initial guess of the values of the parameters, one can use the standard hill-climbing optimization routine (e.g., the Quasi-Newton BFGS algorithm) to find the values of each block of parameters that maximizes the posterior density while holding other blocks of parameters fixed at the previous values. Iterate this algorithm across blocks until it converges. For each iteration, we also employ a *constrained* optimization method to check whether there are boundary solutions associated with Q or any other model parameters.

V. New Implementation of the MHM Method

For many empirical models, the modified harmonic mean (MHM) method of Gelfand and Dey (1994) is a widely used method to compute 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. 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 be $p(\theta)$, both of which must have proper probability densities instead of their kernels. Given these two objects, the marginal data density is defined as

$$p(Y_T) = \int p(Y_T \mid \theta) p(\theta)\, d\theta. \tag{35}$$

The MHM method used to approximate (35) numerically is based on a theorem that states

$$p(Y_T)^{-1} = \int_\Theta \frac{h(\theta)}{p(Y_T \mid \theta)p(\theta)} p(\theta \mid Y_T) d\theta, \tag{36}$$

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 (36) can be accomplished in principle through the Monte Carlo (MC) integration

$$\hat{p}(Y_T)^{-1} = \frac{1}{N} \sum_{i=1}^{N} m(\theta^{(i)}), \tag{37}$$

where $\theta^{(i)}$ is the i^{th} 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 Gaussian density 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 may be quite small at the sample mean, especially when the posterior density has multiple peaks. Second, a truncated Gaussian density function may be a poor local approximation to the posterior density. Third, as one can see from (8), the likelihood tends to be zero in the interior points of the domain Θ.

To deal with these potential 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 $\hat{\theta}$ and 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 i^{th} draw from the posterior simulator and N is the sample size. Computing the posterior mode is typically more expensive than computing the sample mean (see Section IV), but it greatly improves efficiency of the MHM method. Instead of the Gaussian family of distributions, we use 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 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. Simply draw x from a 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, 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 should be chosen so that its cumulative density closely matches the estimated one. There are many ways to accomplish this task. For instance, f could be chosen to be a step function such that the cumulative density is piecewise-linear approximation of the estimated cumulative density. We chose a somewhat simpler technique. The density f has support on $[a,b]$ and is defined by

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

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 is 0.1 and the probability of $r < c_{90}$ from f_0 is 0.9. These choices translate into

$$v = \frac{\log(1/9)}{\log(c_{10}/c_{90})},\ b = \frac{c_{90}}{0.9^{1/v}}. \tag{38}$$

For the reasons elaborated below, we set the value of a to c_1. With the nonzero value of a and the values of v and b specified in (38), one should note that the probability of $r < c_p$ from f will not be exactly p, where $p = 0.1$ or $p = 0.9$.

We now turn to the method we use to truncate the elliptical distribution g. Let U be a positive number and Θ_U be the region defined by

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

The weighting function h 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 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. Since we can take i.i.d. draws from an elliptical distribution, the estimate of q_U has a binomial distribution and its accuracy can be readily obtained. The lower the truncation value of U is, the larger the effective sample size of a sequence of MCMC draws is, but the less acceptable the value of $\hat{q}_U$ becomes. Therefore, there is a balance between having a low 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 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, then h is given by

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

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

The new MHM method developed here is computationally more demanding than the standard MHM implementation, but it avoids the potential problems associated with ill-behaved patters of posterior draws of $m(\theta)$ when a Gaussian approximation to the posterior distribution is poor. 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 random 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. Note that $\hat{q}_L$ has a binomial distribution and depends on the number of MCMC draws and the sample simulated from $h()$. If $\hat{q}_L < 1.0e-06$, this estimate is unreliable because three or four standard deviations will include the value zero. As a rule of thumb, we keep $\hat{q}_L \geq 1.0e-04$.
(3) For each value of L, estimate the marginal data density according to (37).

This procedure can also be implemented by selecting a good value of the upper bound U. 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.

We have thus far discussed our new MHM procedure based on the constant-parameter case. For the 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 II.4. The transition matrix parameters w_j's are treated separately from θ and we use a Dirichlet density instead of a truncated power density as the weighting function for w_j.

VI. APPLICATION

In this section we apply our method developed in the previous sections to a regime-switching three-variable VAR model with five lags. The three variables are those commonly used by recent DSGE models: log GDP (x_t), GDP-deflator inflation (π_t), and the federal funds rate (R_t). The data are quarterly from 1959:I to 2005:IV. Recent debate on changes in monetary policy has focused on whether the coefficients in the policy equation have changed or the variance sizes for structural shocks have changed. Using the notation in Section III.1, we let

$$y_t = \begin{bmatrix} x_t & \pi_t & R_t \end{bmatrix}'.$$

Following the identification of Christiano, Eichenbaum, and Evans (2005), we consider the upper triangular matrix $A(s_t)$ where the last equation is the interest rate equation. We study a large number of models and compare their fits to the data. The types of models are described as follows

#v: Each equation is of Case II with # states under one common Markov process. We call this type of model "variance-only."

#vm: Each equation is of Case III with # states under one common Markov process. We call this type of model "all-change" (i.e., both variances and means changing).

#vRm: The interest rate (R) equation (i.e., the third equation in our application) is of Case III and the other two equations are of Case II with # states under one common Markov process. We call this type of model "policy-change" (i.e., both variances and coefficients in the policy equation changing).

$\#_1$v$\#_2$m: Each equation is of Case III, with $\#_1$ states under one Markov process for $a_j(s_{1t})$ and $f_j(s_{1t})$ and with $\#_2$ states under the other independent Markov process for $\xi_j(s_{2t})$, where $j = 1,\ldots,n$. We call this type of model "variance-with-all-change."

$\#_1$v$\#_2$Rm: The interest rate equation is of Case III and the other equations are of Case II, with $\#_1$ states under one Markov process for $a_j(s_{1t})$ and $f_j(s_{1t})$ and with $\#_2$ states under the other independent Markov process for $\xi_j(s_{2t})$, where $j = 1,\ldots,n$. We call this type of model "variance-with-policy-change."

For all these quarterly models, the tightness values for the BVAR reference prior are, in the notation of Sims and Zha (1998), $\lambda_0 = 1.0, \lambda_1 = 1.0, \lambda_2 = 1.0, \lambda_3 = 1.2, \lambda_4 = 0.1, \mu_5 = 1.0$, and $\mu_6 = 1.0$. These hyperparameter values determine the prior covariance matrices $\bar{\Sigma}_{b_j}$ and $\bar{\Sigma}_{\psi_j}$. For other prior settings, we follow Sims and Zha (2006) and set $\sigma_\delta = 50$, $\bar{\alpha}_j = 1.0$, and $\bar{\beta}_j = 1.0$. For the prior distribution of the transition probability q_j as discussed in Section II.4, we first begin with the case where q_j is unrestricted, as this case is commonly considered in the literature. We set $\alpha_{i,j} = 1$ for $i \neq j$ and

$$\alpha_{j,j} = \frac{p_{j,\text{dur}}(h-1)}{1 - p_{j,\text{dur}}}, \tag{39}$$

where $p_{j,\text{dur}} = Eq_{j,j}$ is the expected value of the probability of staying in the same state (here state j). This prior setting, differing from that of Sims and Zha (2006), allows the possibility that the posterior estimate of $q_{j,j}$ may be one (i.e., allowing the j^{th} state to be irreversible). For our quarterly data, we set $p_{j,\text{dur}} = 0.85$, implying a prior belief that the average duration of staying in the same state is between 6 and 7 quarters. For the four-state case, it follows from (39) that

$$\alpha_{j,j} = 17,\ \alpha_{i,j} = 1 \text{ for } i \neq j. \tag{40}$$

In our application, we restrict the transition matrix in the pattern of (2) when the number of states for a given Markov process is greater than two. Thus, in the case of four states, the transition matrix is restricted as

$$Q = \begin{bmatrix} \pi_1 & (1-\pi_2)/2 & 0 & 0 \\ 1-\pi_1 & \pi_2 & (1-\pi_3)/2 & 0 \\ 0 & (1-\pi_2)/2 & \pi_3 & 1-\pi_4 \\ 0 & 0 & (1-\pi_3)/2 & \pi_4 \end{bmatrix}. \tag{41}$$

Take as an example the first two columns of Q in the case of (41). Expressing the restrictions on q_1 and q_2 in the form of (1), we have

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

If we take as given the values of $\alpha_{i,j}$ specified in (39) (as supplied by a user who is used to working on an unrestricted transition matrix) and transform them to $\beta_{i,j}$ as

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

we have

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

According to (4), we have

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

With the values specified in (40), we have $Eq_{j,j} = Ew_{j,j} = 0.94$, implying a prior belief that the average duration of staying in the same state is about 17 quarters, much longer than the prior belief when Q is unrestricted. Although this is a prior we use for our application, we recommend that in future research one should specify the prior on $w_{i,j}$ directly to maintain the same prior belief on the average duration whether one works on an unrestricted or restricted transition matrix.

Using the blockwise optimization algorithm described in Section IV, we obtain the posterior estimates of the model parameters.[13] With this estimate or any random value near the neighborhood of this estimate as a starting point, we simulate a sequence of 20 million MCMC draws to compute the marginal data density using the new method described in Section V.[14] For the case of 3 states, the restricted transition matrix takes the form of (2). For the case of 4 states,

Table 1 reports log values of marginal data densities for nine different models. The MDDs are not sensitive to the cutoff value L for our new MHM method. In the table, we report only one value and the corresponding $\hat{q}_L$. For each sequence of MCMC draws, we use the software *R* to compute an effective sample size (ESS) (i.e., the sample size adjusted

[13]For our C program, this algorithm takes less than 1 minute while the EM algorithm takes about 9 hours on a Pentium-4 personal desktop computer.

[14]It takes about 20 minutes to simulate one million MCMC draws.

for serial correlation of MCMC draws) according to Plummer, Best, Cowles, and Vines (2005). For all the models studied in Table 1, the computed ESSs are near one million.[15] Based on the ESS, thus, the numerical standard error on the estimated MDD is trivially small. On the similar magnitude, we obtain very small numerical standard errors based on the procedure of Newey and West (1987).

It is known, however, that these measures tend to deliver much smaller numerical standard errors than the actual ones. We propose a different measure by breaking a sequence of 20 million MCMC draws into 10 successive blocks with each block having 2 million draws. For each block, we compute log value of the inverse of the estimated mean of $m(\theta)$ (by a proper scaling to avoid an overflow in computation). The standard error of log MDD is then computed according to the differences of log MDD across blocks and reported in Table 1. As we can see, the standard error is much smaller for the 2-state variance-only model than that for the 4-state variance-only model. In general, the standard error increases with the degree of time variation. Figures 1 and 2 plot the values of log MDD across blocks for the 2-state and 3-state variance-only models. As can be seen, the estimated log MDD is quite stable across blocks for the 2-state case where a Gaussian approximation is likely to be good. For the 3-state variance-only model, however, we begin to see noticeable differences across blocks.

The best-fit model is 3-state or 4-state variance-only model, which seems to dominate all other models by taking into account the standard error of the estimated log MDD. Among the models with changing coefficients, the 3v2Rm variance-with-policy-change model is the best, which does not improve upon the 3-state variance-only model. The conclusion that the variance-only model dominates remains if the Schwarz criterion is applied to the posterior kernel.

To examine whether there exists any bias from our procedure in favor of variance-only models or models with independent Markov processes, we simulate a series of 2000 data points from the 2vRm model where the coefficients in the third equation switch between 2 states and the Markov process is the same for both coefficients and shock variances. We apply our procedure to this artificial data set. The 2vRm model has the best fit with the

[15] Because of some memory management problems associated with the program *R*, the ESSs are estimated on the smaller sample thinned by every twenty MCMC draws.

estimated log MDD being 19763.6. The second-best models are 3v2Rm with log MDD being 19754.49 and 2v3Rm with log MDD being 19753.54. The other models have even lower values of log MDD.

Our exercises point to the fact that accurate calculation of the MDD is an extremely challenging task and give reasons why our method is useful when the posterior distribution is non-Gaussian.

VII. Conclusion

We have developed methods of inference for a class of multiple-equation Markov-switching models with restricted transition matrices. The methods apply to both structural and unrestricted switching VARs. We have described a blockwise optimization algorithm that proves to be much more efficient in these models than the EM algorithm that has been widely applied to similar models. Our variant on the MHM method deals explicitly with the problem of zero likelihood in the interior points of the parameter space. This problem makes many of the usual estimates of the accuracy of results from MCMC simulations unreliable, and we suspect that the problem may be present and unrecognized in some of the recent macroeconomic literature that reports posterior odds ratios on models.

We have proposed a new weighting function used by the MHM method, which is key to obtaining reasonable estimates of marginal data densities in our exercises. This weighting function is likely to be of general use, as in model comparison one often needs a reasonable approximation of the posterior density whose distribution may be very non-Gaussian.

We hope the various ideas we have presented make possible wider use of this class of models, since it represents one convenient approach to accounting for a salient fact about economic time series — their volatilities, and occasionally their dynamic responses, change over time.

TABLE 1. Marginal data densities by new MHM method

	2v	2vm	2vRm	2v2m	2v2Rm
log(MDD)	1821.70	1831.72	1833.41	1857.80	1837.61
s.d. of log(MDD)	0.051	0.43	0.042	0.045	0.034
log(L)	1689.68	1647.12	1699.42	1640.25	1703.42
$\hat{q}_L$	0.381	1.6e-5	2.0e-3	1.72e-4	3.06e-4
	2v3Rm	3v	3v2Rm	4v	
log(MDD)	1839.47	1865.70	1863.04	1867.96	
s.d. of log(MDD)	0.35	0.446	0.11	0.13	
log(L)	1664.24	1719.14	1691.35	1717.53	
$\hat{q}_L$	8.0e-6	5.77e-5	2.2e-5	5.1e-5	

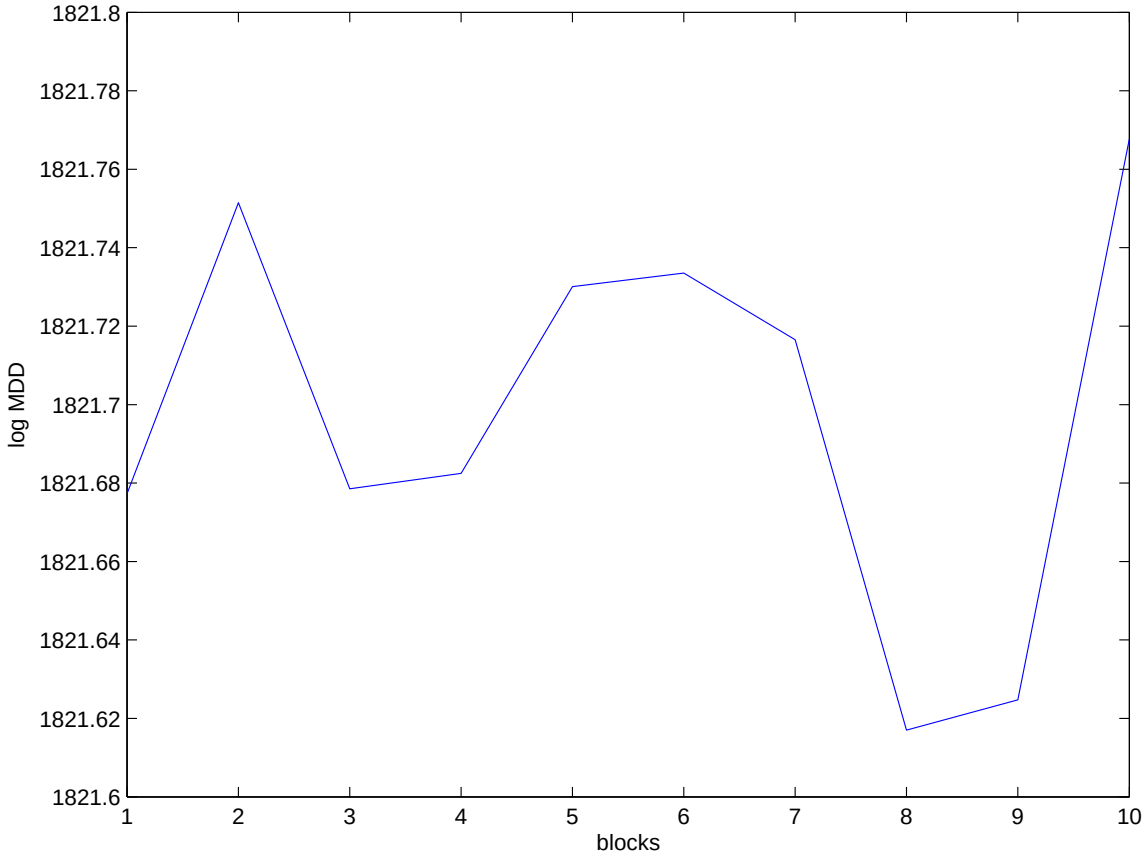

FIGURE 1. The 2-state variance-only (2v) model.

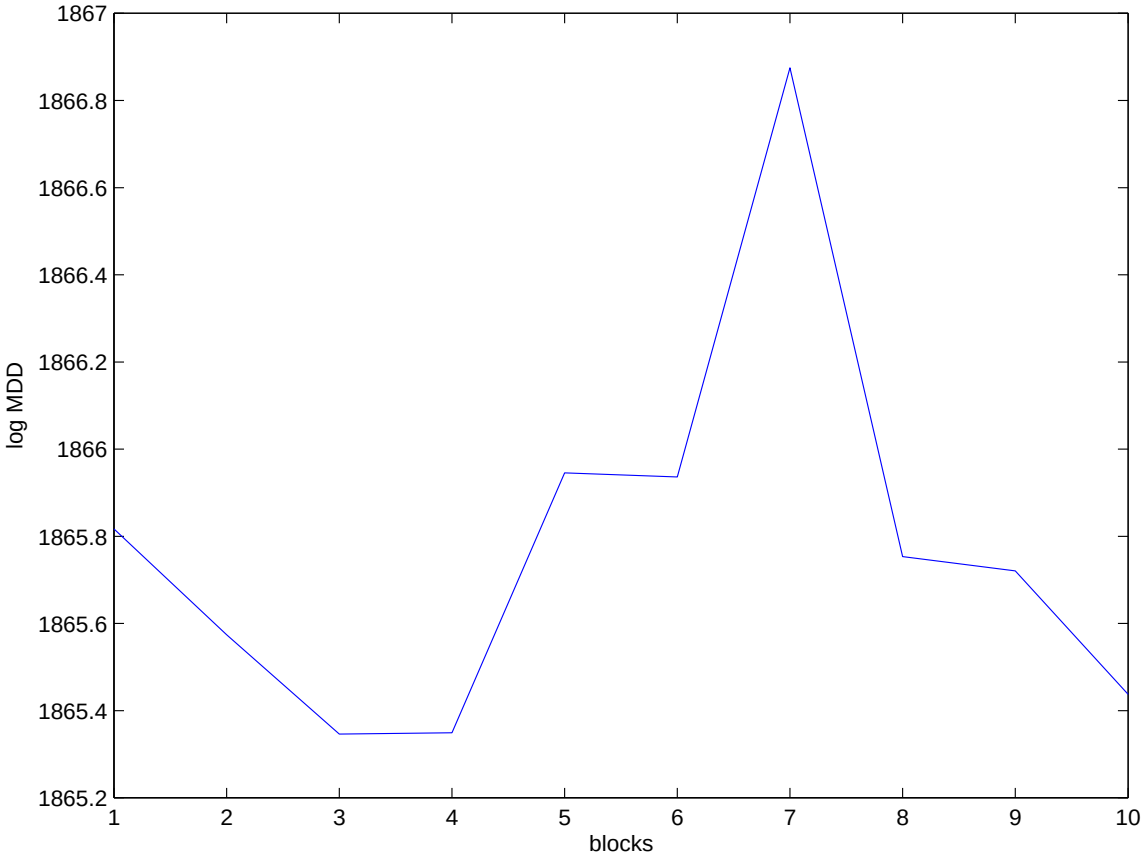

FIGURE 2. The 3 variance-only (3v) model.

APPENDIX A. COMPUTING U_j, V_j, AND W_j

We assume that a_j and f_j satisfy linear restrictions of the form

$$Q_j \left[\begin{array}{c} a_j \\ f_j \end{array} \right] = 0$$

where Q_j is a $(n+k)\times(n+k)$ with $k = np+m$. The matrix Q_j will not be of full rank. We show that there exist a $n \times q_j$ matrix U_j with orthogonal columns, a $(pn+m)\times r_j$ matrix V_j with orthogonal columns, and a such that $(pn+m)\times n$ matrix W_j with $W_j'V_j = 0$ such that

$$\begin{aligned} a_j(k) &= U_j b_j(k) \\ f_j(k) &= V_j g_j(k) - W_j U_j b_j(k) \end{aligned}$$

To prove this we rely on the following result:

Proposition 5. Given any $r \times s$ matrix X with $r \geq s$, there exist an invertiable $r \times r$ matrix Y and a $s \times s$ orthogonal matrix $\left[\begin{array}{cc} \hat{Z} & Z \end{array} \right]$ where Z is a $s \times q$ matrix and $\hat{Z}$ is a $s \times (s-q)$ such that

$$Y^{-1}X = \left[\begin{array}{c} \hat{Z}' \\ 0 \end{array} \right]$$

Proof. This follows directly from the singular value decomposition of X. Let $X = UDV'$ where U is an $r \times r$ orthogonal matrix, V is a $s \times s$ orthogonal matrix, and D is a $r \times s$ diagonal matrix where the first $s-q$ diagonal elements are non-zero and the last q diagonal elements are zero. The first $s-q$ columns of V will be $\hat{Z}$, the last q columns of V will be Z, and $Y = UE$ where E is the $r \times r$ diagonal matrix whose first $s-q$ diagonal elements are the first $s-q$ diagonal elements of D, and the last $r-(s-q)$ diagonal elements are one. □

Applying the above propostion to the last k columns of Q_j, there exists a $(n+k)\times(n+k)$ invertiable matrix Y_1 and a $k \times k$ orthogonal matrix $\left[\begin{array}{cc} \hat{V}_j & V_j \end{array} \right]$ where $\hat{V}_j$ is $k \times \left(k - r_j\right)$ and V_j is $k \times r_j$ such that

$$Y_1^{-1} Q_j = \left[\begin{array}{cc} \hat{W}_j & \hat{V}_j' \\ \tilde{U}_j & 0 \end{array} \right]$$

Now applying the above proposition to the $(n+r_j)\times n$ matrix $\tilde{U}_j$, there exists a $(n+r_j)\times(n+r_j)$ invertiable matrix Y_2 and a $n\times n$ orthogonal matrix $\begin{bmatrix} \hat{U}_j & U_j \end{bmatrix}$ where $\hat{U}_j$ is $n\times(n-q_j)$ and U_j is $n\times q_j$ such that

$$\begin{bmatrix} I_{k-r_j} & 0 \\ 0 & Y_2^{-1} \end{bmatrix} Y_1^{-1} Q_j = \begin{bmatrix} \hat{W}_j & \hat{V}_j' \\ \hat{U}_j' & 0 \\ 0 & 0 \end{bmatrix}$$

Thus a_j and f_j satisfy the restrictions if and only if

$$\begin{bmatrix} \hat{W}_j & \hat{V}_j' \\ \hat{U}_j' & 0 \end{bmatrix} \begin{bmatrix} a_j \\ f_j \end{bmatrix} = 0.$$

Since both $\hat{V}_j'\hat{V}_j$ and $\hat{U}_j'\hat{U}_j$ are equal to a identity matrix, writing $a_j = U_j b_j + \hat{U}_j c_j$ and $f_j = V_j g_j + \hat{V}_j h_j$ gives

$$\begin{bmatrix} \hat{W}_j & \hat{V}_j' \\ \hat{U}_j' & 0 \end{bmatrix} \begin{bmatrix} a_j \\ f_j \end{bmatrix} = \begin{bmatrix} \hat{W}_j U_j b_j + \hat{W}_j \hat{U}_j c_j + h_j \\ c_j \end{bmatrix}.$$

This is zero if and only if $c_j = 0$ and $h_j = -\hat{W}_j U_j b_j$. If we define $W_j = \hat{V}_j \hat{W}_j$, then the result follows.

REFERENCES

BEYER, A., AND R. E. FARMER (2004): "What We Don't Know About the Monetary Transmission Mechanism and Why We Don't Know It," Centre for Economic Policy Research Discussion Paper No. 4811.

CANOVA, F., AND L. GAMBETTI (2004): "Structural Changes in the US Economy: Bad Luck or Bad Policy," Manuscript, Universitat Pompeu Fabra.

CHIB, S. (1996): "Calculating Posterior Distributions and Model Estimates in Markov Mixture Models," *Journal of Econometrics*, 75, 79–97.

CHRISTIANO, L., M. EICHENBAUM, AND C. EVANS (2005): "Nominal Rigidities and the Dynamics Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

COGLEY, T., AND T. J. SARGENT (2002): "Evolving US Post-Wolrd War II Inflation Dynamics," *NBER Macroeconomics Annual*, 16, 331–373.

——— (2005): "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S.," *Review of Economic Dynamics*, 8, 262–302.

COOLEY, T. F., S. F. LEROY, AND N. RAYMON (1984): "Econometric policy evaluation: Note," *The American Economic Review*, 74, 467–470.

FARMER, R. E., D. F. WAGGONER, AND T. ZHA (2006): "Minimal State Variable Solutions to Markov-Switching Rational Expectations Models," Unpublished Manuscript.

GELFAND, A. E., AND D. K. DEY (1994): "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society (Series B)*, 56, 501–514.

GEWEKE, J. (1999): "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication," *Econometric Reviews*, 18(1), 1–73.

——— (2006): "Interpretation and Inference in Mixture Models: Simple MCMC Works," Manuscript, University of Iowa.

HAMILTON, J. D. (1989): "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57(2), 357–384.

——— (1994): *Times Series Analsis*. Princeton University Press, Princeton, NJ.

HAMILTON, J. D., D. F. WAGGONER, AND T. ZHA (2004): "Normalization in Econometrics," Federal Reserve Bank of Atlanta Working Paper 2004-13.

KIM, C.-J., AND C. R. NELSON (1999): *State-Space Models with Regime Switching*. MIT Press, London, England and Cambridge, Massachusetts.

LEEPER, E. M., AND T. ZHA (2003): "Modest Policy Interventions," *Journal of Monetary Economics*, 50(8), 1673–1700.

NEWEY, W. K., AND K. K. WEST (1987): "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

PLUMMER, M., N. BEST, K. COWLES, AND K. VINES (2005): "The coda Package," Version 0.10-2, November, plummer@iarc.fr.

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

ROBERTSON, J. C., AND E. W. TALLMAN (1999): "Vector Autoregressions: Forecasting and Reality," *Federal Reserve Bank of Atlanta Economic Review*, First Quarter, 4–18.

——— (2001): "Improving Federal-Funds Rate Forecasts in VAR Models Used for Policy Analysis," *Journal of Business and Economic Statistics*, 19(3), 324–330.

RUBIO-RAMÍREZ, J. F., D. F. WAGGONER, AND T. ZHA (2006): "Structural Vector Autoregressions: Theory and Application," Manuscript, Duke University and Federal Reserve Bank of Atlanta.

SIMS, C. A. (1987): "A rational expectations framework for short-run policy analysis," in *New approaches to monetary economics*, ed. by W. A. Barnett, and K. J. Singleton, pp. 293–308. Cambridge University Press, Cambridge, England.

——— (1993): "A 9 Variable Probabilistic Macroeconomic Forecasting Model," in *Business Cycles, Indicators, and Forecasting*, ed. by J. H. Stock, and M. W. Watson, vol. 28 of *NBER Studies in Business Cycles*, pp. 179–214. University of Chicago Press.

——— (1999): "Drift and Breaks in Monetary Policy," Manuscript, Princeton University.

——— (2001): "Stability and Instability in US Monetary Policy Behavior," Manuscript, Princeton University.

SIMS, C. A., AND T. ZHA (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39(4), 949–968.

——— (2006): "Were There Regime Switches in US Monetary Policy?," *The American Economic Review*, 96, 54–81.

STOCK, J. H., AND M. W. WATSON (2003): "Has the Business Cycles Changed? Evidence and Explanations," Monetary Policy and Uncertainty: Adapting to a Changing Economy, Federal Reserve Bank of Kansas City Symposium, Jackson Hole, Wyoming, August 28-30.

WAGGONER, D. F., AND T. ZHA (2003a): "A Gibbs Sampler for Structural Vector Autoregressions," *Journal of Economic Dynamics and Control*, 28(2), 349–366.

——— (2003b): "Likelihood Preserving Normalization in Multiple Equation Models," *Journal of Econometrics*, 114(2), 329–347.

ZHA, T. (In press): "Vector Autoregressions," in *The New Palgrave Dictionary of Economics*, ed. by L. E. Blume, and S. Durlauf. Palgrave Macmillan.

PRINCETON UNIVERSITY, FEDERAL RESERVE BANK OF ATLANTA, FEDERAL RESERVE BANK OF ATLANTA