FEDERAL RESERVE BANK *of* ATLANTA

MCMC Method for Markov Mixture Simultaneous-Equation Models: A Note

Christopher A. Sims and Tao Zha

Working Paper 2004-15
June 2004

WORKING PAPER SERIES

FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

MCMC Method for Markov Mixture Simultaneous-Equation Models: A Note

Christopher A. Sims and Tao Zha

Working Paper 2004-15
June 2004

Abstract: This paper extends the methods developed by Hamilton (1989) and Chib (1996) to identified multiple-equation models. It details how to obtain Bayesian estimation and inference for a class of models with different degrees of time variation and discusses both analytical and computational difficulties.

JEL classification: C3

Key words: simultaneity, identification, time variation, volatility, Bayesian method

The authors thank Dan Waggoner for helpful discussions. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to Christopher A. Sims, Department of Economics, 104 Fisher Hall, Princeton University, Princeton, New Jersey 08544-1021, 609-258-4033, sims@princeton.edu, or Tao Zha, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street N.E., Atlanta, Georgia 30309-4470, 404-498-8353, tao.zha@atl.frb.org.

MCMC METHOD FOR MARKOV MIXTURE SIMULTANEOUS-EQUATION MODELS: A NOTE

I. Introduction

We consider nonlinear stochastic dynamic simultaneous equations of the structural form:

$$y_t'A_0(s_t) = x_t'A_+(s_t) + \varepsilon_t', \quad t = 1, \ldots, T, \tag{1}$$

$$\Pr(s_t = i \mid s_{t-1} = k) = p_{ik}, \quad i, k = 1, \ldots, h, \tag{2}$$

where s is an unobserved state, y is an $n \times 1$ vector of endogenous variables, x is an $m \times 1$ vector of exogenous and lagged endogenous variables, A_0 is an $n \times n$ matrix of parameters, A_+ is an $m \times n$ matrix of parameters, T is a sample size, and h is the total number of states.

Denote the longest lag length in the system of equations (1) by ν. The vector of right-hand variables, x_t, is ordered from the n endogenous variables for the first lag down to the n variables for the last (ν^{th}) lag with the last element of x_t being the constant term.

For $t = 1, \ldots, T$, denote

$$Y_t = \{y_1, \ldots, y_t\}.$$

We treat as given the initial lagged values of endogenous variables $Y_0 = \{y_{1-\nu}, \ldots, y_0\}$. Structural disturbances are assumed to have the distribution:

$$\pi(\varepsilon_t \mid Y_{t-1}) = \mathcal{N}\left(\underset{n\times 1}{\mathbf{0}}, \mathbf{I}_n\right),$$

where $\mathcal{N}(a,b)$ refers to the normal pdf with mean a and covariance matrix b and $\mathbf{I}_n$ is an $n \times n$ identity matrix. Following Hamilton 1989 and Chib 1996, we impose no restrictions on the transition matrix $P = [p_{ik}]$.

The reduced-form system of equations implied by (1) is:

$$y_t' = x_t' B(s_t) + u_t'(s_t), \quad t = 1, \ldots, T; \tag{3}$$

where

$$B(s_t) = A_+(s_t) A_0^{-1}(s_t), \tag{4}$$

$$u_t(s_t) = A_0'^{-1}(s_t)\,\varepsilon_t, \tag{5}$$

$$E\left(u_t(s_t) u_t(s_t)'\right) = \left(A_0(s_t) A_0'(s_t)\right)^{-1}. \tag{6}$$

In the reduced form (4)-(6), $B(s_t)$ and $u_t(s_t)$ involve the structural parameters and shocks across equations, making it impossible to distinguish regime shifts from one structural equation to another. In contrast, the structural form (1) allows one to identify each structural equation, such as the policy rule, for regime switches.

II. Prior Restrictions

II.1. **Restrictions on time variation.** If we let all parameters vary across states, it is relatively straightforward to apply the existing methods of Chib 1996 and Sims and Zha 1998 to the model estimation because $A_0(s_t)$ and $A_+(s_t)$ in each given state can be estimated independent of the parameters in other states. But with such an unrestricted form for the time variation, if the system of equations is large or the lag length is long, the number of free parameters in the model becomes impractically large. For a typical monthly model with 13 lags and 6 endogenous variables, for example, the number of parameters in $A_+(s_t)$ is of order 468 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 468. 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 A_+ as

$$\underset{m\times n}{A_+(s_t)} = \underset{m\times n}{D(s_t)} + \underset{m\times n}{\overline{S}}\, \underset{n\times n}{A_0(s_t)}. \tag{7}$$

where

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

If we place a prior distribution on $D(s_t)$ that has mean zero, the specification of $\overline{S}$ is consistent with the reduced-form random walk feature implied by existing Bayesian VAR models (Sims and Zha 1998). As can be seen from (4) and (7), this form 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 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.

We consider the following three cases of restricted time variations for $A_0(s_t)$ and $D(s_t)$:

$$a_{0,j}(s_t), d_{ij,\ell}(s_t), c_j(s_t) = \begin{cases} \bar{a}_{0,j}, \bar{d}_{ij,\ell}, \bar{c}_j & \text{Case I} \\ \bar{a}_{0,j}\xi_j(s_t), \bar{d}_{ij,\ell}\xi_j(s_t), \bar{c}_j\xi_j(s_t) & \text{Case II} \\ a_{0,j}(s_t), \bar{d}_{ij,\ell}\lambda_{ij}(s_t), c_j(s_t) & \text{Case III} \end{cases}, \tag{8}$$

where $\xi_j(s_t)$ is a scale factor for the j^{th} structural equation, $a_{0,j}(s_t)$ is the j^{th} column of $A_0(s_t)$, $d_j(s_t)$ is the j^{th} column of $D(s_t)$, $d_{ij,\ell}(s_t)$ is the element of $d_j(s_t)$ for the i^{th} variable at the ℓ^{th} lag, the last element of $d_j(s_t)$, $c_j(s_t)$, is the constant term for equation j. The parameter $\lambda_{ij}(s_t)$ changes with variables but does not vary across lags. The variability across variables 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 bar symbol over $a_{0,j}$, $d_{ij,\ell}$, and c_j means that these parameters are state-independent (i.e., constant across time).

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. Case III represents a structural equation with time-varying coefficients. [1]

II.2. **Identifying restrictions.** It is well known that the model (1) would be unidentified without further identifying restrictions. We follow the identified VAR literature and apply linear restrictions on A_0 and D, which imply the following relationships (Waggoner and Zha 2003a)

$$\underset{nh\times 1}{a_j} = \underset{nh\times o_j}{U_j}\ \underset{o_j\times 1}{b_j}\ , j = 1, \ldots, n, \tag{9}$$

$$\underset{mh\times 1}{d_j} = \underset{mh\times r_j}{V_j}\ \underset{r_j\times 1}{g_j}\ , j = 1, \ldots, n, \tag{10}$$

$$a_j = \begin{bmatrix} a_{0,j}(1) \\ \vdots \\ a_{0,j}(h) \end{bmatrix}, \quad d_j = \begin{bmatrix} d_j(1) \\ \vdots \\ d_j(h) \end{bmatrix},$$

where b_j and g_j are the free parameters "squeezed" out of a_j and d_j by the linear restrictions, o_j and r_j are the numbers of the corresponding free parameters, columns of U_j are orthonormal vectors in the Euclidean space $\mathbb{R}^{nh}$, and columns of V_j are orthonormal vectors in $\mathbb{R}^{mh}$.

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

The restricted form (9) - (10) encompasses many existing models. For example, the restrictions imposed on the three-equation system of Taylor 1999 (one of the equations is the widely-used Taylor rule) fall into this form where $a_{0,j}(1) = \cdots = a_{0,j}(h)$ and $d_j(1) = \cdots = d_j(h)$ for all j's. Similarly, the Rudebusch and Svensson 2002's empirical model is of the form with different lag structures imposed on different equations, which is summarized by (10).

II.3. **Prior distributions.** In addition to these identifying restrictions, we use the reference prior on A_0 and D in the existing literature. The prior distributions take the Gaussian form:

$$\pi(a_{0,j}(k)) = \mathcal{N}(\mathbf{0}, H_{0j}),\ k = 1,\ldots,h,\ j = 1,\ldots,n; \tag{11}$$

$$\pi(d_j(k)) = \mathcal{N}(\mathbf{0}, H_{+j}),\ k = 1,\ldots,h,\ j = 1,\ldots,n. \tag{12}$$

We follow Sims and Zha 1998 to incorporate into the model the $n+1$ "dummy observations" formed from the initial observations (Y_0). 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 (1) and X_d be an $(n+1)\times m$ matrix of dummy observations on the right hand side. It follows from Sims and Zha 1998 that

$$(X_d'X_d + H_{+j}^{-1})^{-1}(X_d'Y_d + H_{+j}^{-1}\overline{S}) = \overline{S},$$

$$Y_d'Y_d + H_{0j}^{-1} + \overline{S}'H_{+j}^{-1}\overline{S} - \overline{\Sigma}_{0j}^{-1} = H_{0j}^{-1},$$

where

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

These results, combined with (9), (10), (11), and (12), lead to the prior distributions for the free parameters b_j and g_j:

$$\pi(b_j) = \mathcal{N}(\mathbf{0}, \overline{H}_{0j}), \tag{13}$$

$$\pi(g_j) = \mathcal{N}(\mathbf{0}, \overline{H}_{+j}), \tag{14}$$

where

$$\overline{H}_{0j} = \left(U_j'(I \otimes H_{0j}^{-1})U_j\right)^{-1},$$

$$\overline{H}_{+j} = \left(V_j'(I \otimes (X_d'X_d + H_{+j}^{-1}))V_j\right)^{-1}.$$

The prior distribution for $\xi_j(k)$ is taken as $\pi(\zeta_j(k)) = \Gamma(\alpha_\zeta, \beta_\zeta)$ for $k \in \{1,\ldots,h\}$, where $\zeta_j(k) \equiv \xi_j^2(k)$ and $\Gamma(\cdot)$ denotes the standard gamma pdf with β_ζ being a scale factor (not an inverse scale factor as in the notation of some textbooks). The prior pdf for $\lambda_{ij}(k)$ is $\mathcal{N}(0, \sigma_\lambda^2)$ for $k \in \{1,\ldots,h\}$.

The prior of the transition matrix P takes a Dirichlet form as suggested by Chib 1996. For the k^{th} column of P, p_k, the prior density is

$$\pi(p_k) = \pi(p_{1k}, \dots, p_{hk}) = \mathcal{D}(\alpha_{1k}, \dots, \alpha_{hk}) \propto p_{1k}^{\alpha_{1k}-1} \cdots p_{hk}^{\alpha_{hk}-1}, \tag{15}$$

where $\alpha_{ik} > 0$ for $i = 1, \dots, h$.

There are three steps in setting up a prior for p_k. First, the prior mode of p_{ik} is chosen to be υ_{ik} such that $\sum_{i=1}^{h} \upsilon_{ik} = 1$. Let

$$\upsilon_{zj} = \max\{\upsilon_{1j}, \dots, \upsilon_{hj}\}.$$

In general, we set $z = j$ or $\upsilon_{zj} = \upsilon_{jj}$, reflecting beliefs in the concentration on the diagonal of P.

In the second step, choose the prior variance of p_{zj} to be ε_{zj} such that is sufficiently small to have a real solution for α_{zj} in the following third order polynomial

$$\phi_3 \alpha_{zj}^3 + \phi_2 \alpha_{zj}^2 + \phi_1 \alpha_{zj} + \phi_0 = 0,$$

where

$$\begin{aligned} \psi_{zj} &= \frac{1 - \upsilon_{zj}}{\upsilon_{zj}}, \\ \phi_3 &= \varepsilon_{zj}(1 + \psi_{zj})^3, \\ \phi_2 &= \varepsilon_{zj}(1 + \psi_{zj})^2 \left[3(h - \psi_{zj}) - 2\right] - \psi_{zj}, \\ \phi_1 &= (h - \psi_{zj} - 1) \left\{ \varepsilon_{zj}(1 + \psi_{zj}) \left[3(h - \psi_{zj}) - 1\right] - 1 \right\}, \\ \phi_0 &= \varepsilon_{zj}(h - \psi_{zj})(h - \psi_{zj} - 1)^2. \end{aligned}$$

Given the prior modes $\{\upsilon_{1j}, \dots, \upsilon_{hj}\}$ and the solution α_{zj}, the third step involves solving for all the other elements of the vector α_j as

$$\underset{(h-1)\times 1}{\alpha_j^\star} = \underset{(h-1)\times(h-1)}{B_j^{\star -1}} \left[\underset{(h-1)\times 1}{\varsigma_j^\star} - \alpha_{zj} \underset{(h-1)\times 1}{\beta_j^\star} \right],$$

where $\alpha_j^\star$ is the $(h-1) \times 1$ subvector of α_j without the z^{th} element, $B_j^\star$ is the $(h-1) \times (h-1)$ submatrix of the following matrix without the z^{th} row and z^{th} column

$$\begin{bmatrix} 1 - \upsilon_{1j} & -\upsilon_{1j} & \dots & -\upsilon_{1j} \\ \upsilon_{2j} & 1 - \upsilon_{2j} & \dots & -\upsilon_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ -\upsilon_{hj} & -\upsilon_{hj} & \dots & 1 - \upsilon_{hj} \end{bmatrix},$$

$\beta_j^\star$ is the $(h-1)\times 1$ subvector of the z^{th} column of the above matrix without the z^{th} element, and $\varsigma_j^\star$ is the $(h-1)\times 1$ subvector of the following vector without the z^{th} element

$$\begin{bmatrix} 1-h\upsilon_{1j} \\ \vdots \\ 1-h\upsilon_{hj} \end{bmatrix}.$$

It can be verified that the values of all elements of α_j constructed above imply

$$\varepsilon_{zj} \geq \text{Var}(p_{kj}),\ \forall k \in \{1,\ldots,h\}.$$

III. Posterior Estimate

As shown in Section II.3, the restricted parameters (through identification and the degree of time variation) are functions of the free parameters. We gather different groups of free parameters as follows, with the understanding that we sometimes interchange the use of free parameters and original (but restricted) parameters.

$$p = \{p_k, k=1,\ldots,h\};$$
$$\gamma = \begin{cases} \zeta = \left\{\zeta_j(k),\ j=1,\ldots,n, k=1,\ldots,h\right\}, & \text{for Case II;} \\ \lambda = \left\{\lambda_{ij}(k),\ i,j=1,\ldots,n, k=1,\ldots,h\right\}, & \text{for Case III;} \end{cases}$$
$$g = \left\{g_j,\ j=1,\ldots,n\right\};$$
$$b = \left\{b_j,\ j=1,\ldots,n\right\};$$
$$\theta = \{p,\gamma,g,b\}.$$

The overall likelihood function $\pi(Y_T \mid \theta)$ can be obtained by integrating over unobserved states the conditional likelihood at each time t and by recursively multiplying these conditional likelihood functions forward (Kim and Nelson 1999):

$$\pi(Y_T \mid \theta) = \prod_{t=1}^{T} \left\{ \sum_{s_t=1}^{h} \left[\pi(y_t \mid Y_{t-1}, s_t, \theta)\,\text{Pr}(s_t \mid Y_{t-1},\theta)\right] \right\}, \tag{16}$$

where

$$\begin{aligned} \pi(y_t \mid Y_{t-1}, s_t, \theta) = (2\pi)^{-\frac{n}{2}}\, |A_0(s_t)| \exp\Bigg\{ -\frac{1}{2}\sum_{j=1}^{n} \Big[& a'_{0,j}(s_t)\, y_t y'_t\, a_{0,j}(s_t) \\ & -2a'_{+,j}(s_t)\, x_t y'_t\, a_{0,j} + a'_{+,j}(s_t)\, x_t x'_t\, a_{+,j}(s_t)\Big] \Bigg\}, \end{aligned} \tag{17}$$

$$\text{Pr}(s_t \mid Y_{t-1},\theta) = \sum_{s_{t-1}=1}^{h} \left[\text{Pr}(s_t \mid s_{t-1})\text{Pr}(s_{t-1} \mid Y_{t-1},\theta)\right]. \tag{18}$$

The probability $\Pr(s_{t-1} \mid Y_{t-1}, \theta)$ can be updated recursively. We begin by setting $\Pr(s_0 \mid Y_0, \theta) = \Pr(s_0) = 1/h$. For $t = 1, \ldots, T$, the updating procedure involves the following computation:

$$\Pr(s_t \mid Y_t, \theta) = \frac{\pi(y_t \mid Y_{t-1}, s_t, \theta) \Pr(s_t \mid Y_{t-1}, \theta)}{\sum_{s_t=1}^{h} \left[\pi(y_t \mid Y_{t-1}, s_t, \theta) \Pr(s_t \mid Y_{t-1}, \theta)\right]}. \tag{19}$$

From the Bayes rule, the posterior distribution of θ conditional on the data is

$$\pi(\theta \mid Y_T) \propto \pi(\theta)\pi(Y_T \mid \theta), \tag{20}$$

where the prior $\pi(\theta)$ is specified in Section II.3.

In order to avoid very long startup periods for the MCMC sampler, it is important to begin with at least an approximate estimate of the peak of the posterior density (20). Moreover, such an estimate is used as a reference point in normalization to obtain likelihood-based statistical inferences. Because the number of parameters is quite large for our models (over 500), we used an eclectic approach, combining the stochastic expectation-maximizing algorithm with various optimization routines.

IV. Inference

Our objective is to obtain the posterior distribution of functions of θ such as impulse responses, forecasts, historical decompositions, and long-run responses of policy. It involves integrating over large dimensions many highly nonlinear functions. Because most of these dimensions are related to unobserved states, there is no analytical solution for $\pi(\theta|Y_T)$, nor is it possible to simulate from its distribution. One can, however, use a Gibbs sampler to obtain the joint distribution $\pi(\theta, S_T \mid Y_T)$ where

$$S_t = \{s_0, s_1, \ldots, s_t\}, \quad \forall t \in \{1, \ldots, T\}.$$

The Gibbs sampler we propose here involves sampling alternatively from the following conditional posterior distributions:

$$\begin{gathered}\Pr(S_T \mid Y_T, p, \gamma, g, b),\\ \pi(p \mid Y_T, S_T, \gamma, g, b),\\ \pi(\gamma \mid Y_T, S_T, p, g, b),\\ \pi(g \mid Y_T, S_T, p, \gamma, b),\\ \pi(b \mid Y_T, S_T, p, \gamma, g).\end{gathered}$$

It has been shown in the literature that such a Gibbs sampling procedure produces the unique limiting distribution that is the posterior distribution of S_T and θ (e.g., Geweke 1999).

IV.1. **Conditional posterior distribution of S_T.** Denote

$$S^t = \{s_t, \ldots, s_T\},\, Y^t = \{y_t, \ldots, y_T\}, \quad \forall t \in \{1, \ldots, T\}.$$

Paths of S_T are simulated recursively backward from the last conditional posterior distribution whose pdf is $\Pr(S_T \mid Y_T, p, \gamma, g, b)$ or $\Pr(S_T \mid Y_T, \theta)$. To see how this recursion is accomplished, observe that

$$\Pr(S_T \mid Y_T, \theta) = \Pr(s_T \mid Y_T, \theta) \cdots \Pr(s_t \mid Y_T, S^{t+1}, \theta) \cdots \Pr(s_0 \mid Y_T, S^1, \theta); \tag{21}$$

and

$$\begin{aligned}\Pr(s_t \mid Y_T, S^{t+1}, \theta) &\propto \Pr(s_t \mid Y_t, \theta) \Pr(Y^{t+1}, S^{t+1} \mid Y_t, s_t, \theta)\\ &\propto \Pr(s_t \mid Y_t, \theta) \Pr(s_{t+1} \mid s_t, \theta) \Pr(Y^{t+1}, S^{t+2} \mid Y_t, s_t, s_{t+1}, \theta)\\ &\propto \Pr(s_t \mid Y_t, \theta) \Pr(s_{t+1} \mid s_t, \theta),\end{aligned} \tag{22}$$

because $\Pr(Y^{t+1}, S^{t+2} \mid Y_t, s_t, s_{t+1}, \theta)$ is independent of s_t when s_{t+1} is given. Relationship (22) implies that

$$\Pr(s_t \mid Y_T, S^{t+1}, \theta) = \frac{\Pr(s_t \mid Y_t, \theta) \Pr(s_{t+1} \mid s_t, \theta)}{\sum_{s_t=1}^{h} [\Pr(s_t \mid Y_t, \theta) \Pr(s_{t+1} \mid s_t, \theta)]}. \tag{23}$$

The backward recursion begins by drawing the last state s_T from $\Pr(s_T | Y_T, \theta)$ according to (19) and drawing s_t recursively given the path S^{t+1} according to (23). It can be seen from (21) that draws of S_T this way come from $\Pr(S_T | Y_T, \theta)$.

IV.2. **Conditional posterior distribution of p.** Given the path S_T, the posterior distribution of p is reduced to the following distribution:

$$\pi(p \mid S_T) \propto \Pr(S_T \mid p) \pi(p).$$

The likelihood for S_T given p has the multinomial form:

$$\Pr(S_T \mid p) = \binom{T}{n_{11}, \ldots, n_{h1}, \ldots, n_{1h}, \ldots, n_{hh}} p_{s_1 s_0} p_{s_2 s_1} \cdots p_{s_T s_{T-1}} \Pr(s_0), \tag{24}$$

where n_{ij} is the total number of one-step transitions of s from state j to state i over the entire sample and

$$T = \sum_{i=1}^{h} \sum_{j=1}^{h} n_{ij}.$$

The resulting conditional posterior pdf is also of Dirichlet:

$$\pi(p_k \mid S_T) = \mathcal{D}(\alpha_{1k} + n_{1k}, \ldots, \alpha_{hk} + n_{hk}),\ \forall k \in \{1, \ldots, h\}. \tag{25}$$

IV.3. **Conditional posterior distribution of γ, g, or b.** The likelihood function conditional on S_T and θ is

$$\pi(Y_T \mid S_T, \theta) = \prod_{t=1}^{T} \pi(y_t | Y_{t-1}, s_t, \theta), \tag{26}$$

where $\pi(y_t|Y_{t-1}, s_t, \theta)$ is given by (17). The joint posterior pdf of γ, g, and b conditional on Y_T, S_T, and the other parameters is proportional to the conditional likelihood (26) multiplied by the Gamma prior distribution of ζ, the Gaussian prior distribution of λ, and the Gaussian prior distributions of g and b, specified by (13) and (14). This joint distribution involves a large number of parameters and has an unrecognized analytical form that is impossible to simulate accurately. In what follows, we derive the conditional posterior distribution of each individual group of parameters from which random values of the parameters can be accurately simulated. We begin with Case III and work backward to Case II.

IV.3.1. *Additional Notation.* The following notation, which will be repeatedly used for Cases II and III, is now introduced.

$$\underset{n^2\times n}{\Upsilon} = \begin{bmatrix} \mathbf{e}'_{n,1} \\ \mathbf{0}_n \\ \mathbf{e}'_{n,2} \\ \mathbf{0}_n \\ \vdots \\ \mathbf{e}'_{n,n-1} \\ \mathbf{0}_n \\ \mathbf{e}'_{n,n} \end{bmatrix}, \; \underset{nv\times 1}{d_{j,1}(k)} = \begin{bmatrix} d_{1j,1}(k) \\ \vdots \\ d_{nj,1}(k) \\ \vdots \\ d_{ij,\ell}(k) \\ \vdots \\ d_{1j,v}(k) \\ \vdots \\ d_{nj,v}(k) \end{bmatrix}, \; \underset{nv\times 1}{\bar{d}_{j,1}} = \begin{bmatrix} \bar{d}_{1j,1} \\ \vdots \\ \bar{d}_{nj,1} \\ \vdots \\ \bar{d}_{ij,\ell} \\ \vdots \\ \bar{d}_{1j,v} \\ \vdots \\ \bar{d}_{nj,v} \end{bmatrix}, \; \underset{n\times v}{\overline{D}_{j,1}} = \begin{bmatrix} \bar{d}_{1j,1} & \cdots & \bar{d}_{1j,v} \\ \vdots & \ddots & \vdots \\ \bar{d}_{nj,1} & \cdots & \bar{d}_{nj,v} \end{bmatrix},$$

where $\mathbf{e}_{n,i}$ is a standard unit vector of size $n \times 1$ with the i^{th} element being 1 and $\mathbf{0}_n$ is an $n \times n$ matrix of zeros. Let T_k be the total number of observations such that $s_t = k$. The notation for a block diagonal matrix is defined as

$$\text{diag}\left[\{B_k\}_{k=1}^{h}\right] = \begin{bmatrix} B_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & B_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & B_h \end{bmatrix}.$$

Additional notation for λ is introduced as

$$\underset{nh\times n}{\lambda_j} = \begin{bmatrix} \lambda_j(1) \\ \vdots \\ \lambda_j(k) \\ \vdots \\ \lambda_j(h) \end{bmatrix}, \quad \underset{n\times 1}{\lambda_j(k)} = \begin{bmatrix} \lambda_{1j}(k) \\ \vdots \\ \lambda_{ij}(k) \\ \vdots \\ \lambda_{nj}(k) \end{bmatrix},$$

$$\underset{n\times n}{\Lambda_j(k)} = \begin{bmatrix} \lambda_{1j}(k) & 0 & \cdots & 0 \\ 0 & \lambda_{2j}(k) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{nj}(k) \end{bmatrix}, \quad \underset{m\times m}{\Delta_j(k)} = \begin{bmatrix} \mathbf{I}_\nu \otimes \Lambda_j(k) & \underset{n\nu\times 1}{\mathbf{0}} \\ \underset{1\times n\nu}{\mathbf{0}} & 1 \end{bmatrix}.$$

IV.3.2. *Case III.* Note the relationship

$$d_{j,1}(k) = \left[\mathbf{I}_n \otimes \Lambda_j(k)\right] \bar{d}_{j,1} = (\overline{D}_{j,1} \otimes \mathbf{I}_n)\Upsilon\lambda_j(k).$$

With this and some algebraic work, one can show that the conditional posterior distributions of λ, g, and b have the following forms:

$$\pi(\lambda_j(k) \mid Y_T, S_T, \bar{d}_{j,1}, c_j(k), a_{0,j}(k)) = \mathcal{N}(\widetilde{\lambda}_j(k), \widetilde{\Psi}_{4j}(k)), \tag{27}$$

$$\pi(g_j \mid Y_T, S_T, \lambda_j, b_j) = \mathcal{N}(\widetilde{g}_j, \widetilde{\Psi}_{6j}), \tag{28}$$

$$\pi(b \mid Y_T, S_Y, \gamma, g) \propto \prod_{k=1}^{h} |A_0(k)|^{T_k} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}\left(b_j' \Psi_{8j}^{-1} b_j - 2\Psi_{7j} b_j\right)\right\}, \tag{29}$$

where

$$\widetilde{\lambda}_j(k) = \widetilde{\Psi}_{4j}(k)\Psi_{3j}(k),$$

$$\widetilde{\Psi}_{4j}^{-1}(k) = \Upsilon'\left(\overline{D}_{1,j} \otimes \mathbf{I}_n\right)\Psi_{2j,11}(k)\left(\overline{D}_{1,j}' \otimes \mathbf{I}_n\right)\Upsilon + \frac{1}{\sigma_\lambda^2}\mathbf{I}_n,$$

$$\Psi_{2j}(k) = \sum_{t\in\{t:\, s_t=k\}} \left[x_t x_t'\right] = \begin{bmatrix} \underset{n\nu\times n\nu}{\Psi_{2j,11}(k)} & \underset{n\nu\times 1}{\Psi_{2j,12}(k)} \\ \underset{1\times n\nu}{\Psi_{2j,21}(k)} & \underset{1\times 1}{\Psi_{2j,22}(k)} \end{bmatrix},$$

$$\Psi_{3j}(k) = \Upsilon'\left(\overline{D}_{1,j} \otimes \mathbf{I}_n\right)\left[\Psi_{1j,1\boldsymbol{\cdot}}(k)\, a_{0,j}(k) - \Psi_{2j,12}\, c_j(k)\right],$$

$$\Psi_{1j}(k) = \sum_{t\in\{t:\, s_t=k\}} \left[x_t y_t' - x_t x_t' \overline{S}\right] = \begin{bmatrix} \underset{n\nu\times n}{\Psi_{1j,1\boldsymbol{\cdot}}(k)} \\ \underset{1\times n}{\Psi_{1j,2\boldsymbol{\cdot}}(k)} \end{bmatrix},$$

$$\widetilde{g}_j = \widetilde{\Psi}_{6j} V_j' \Psi_{5j} b_j,$$

$$\widetilde{\Psi}_{6j}^{-1} = V_j' \operatorname{diag}\left[\left\{\Delta_j(k)' \Psi_{2j}(k) \Delta_j(k)\right\}_{k=1}^{h}\right] V_j + \overline{H}_{+j}^{-1},$$

$$\Psi_{5\,j} = \operatorname{diag}\left[\left\{\Delta_j(k)'\Psi_{1\,j}(k)\right\}_{k=1}^{h}\right]U_j,$$
$$\Psi_{8\,j}^{-1} = U_j'\operatorname{diag}\left[\left\{\Psi_{0\,j}(k)\right\}_{k=1}^{h}\right]U_j + \overline{H}_{0j}^{-1},$$
$$\Psi_{0\,j}(k) = \sum_{t\in\{t:\,s_t=k\}}\left[y_ty_t' - 2\overline{S}'x_ty_t' + \overline{S}'x_tx_t'\overline{S}\right],$$
$$\Psi_{7\,j} = \operatorname{diag}\left[\left\{d_j(k)'\Psi_{1\,j}(k)\right\}_{k=1}^{h}\right]U_j.$$

Except (29), one can generate random values directly from the other two conditional posterior distributions (27) - (28). As for the conditional posterior density of b, we use the Gibbs sampling idea of Waggoner and Zha 2003a to sample b_j one at a time conditional on b_i for $i \neq j$ and the other parameters. Unlike constant-parameter simultaneous-equation models, however, the posterior density of b_j conditional on all the other parameters in our case has no recognized form. We thus use a Metropolis algorithm with the following proposal density for the transition from b_j to $b_j^\star$

$$J\left(b_j^\star \mid b_j, Y_T, S_T, b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_n, \gamma, g\right) = \mathcal{N}\left(\underset{o_j\times 1}{\mathbf{0}},\ \kappa_{j,\mathrm{III}}\Psi_{8\,j}\right), \tag{30}$$

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

IV.3.3. *Case II.* If the j^{th} structural equation is Case II,[2] $\lambda_{ij}(k)$ is equal to 1 for all $i \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,h\}$. Thus, $\Lambda_j(k) = \mathbf{I}_n$ and $\Delta_j(k) = \mathbf{I}_m$ $\forall k \in \{1,\ldots,h\}$. The conditional posterior distributions of g and b have the same densities as (28) and (29) except the terms $\Psi_{2\,j}(k)$, $\Psi_{1\,j}(k)$, and $\Psi_{0\,j}(k)$ are now replaced by

$$\Psi_{2\,j}(k) = \sum_{t\in\{t:\,s_t=k\}}\left[\zeta_j(s_t)\,x_tx_t'\right],$$
$$\Psi_{1\,j}(k) = \sum_{t\in\{t:\,s_t=k\}}\left[\zeta_j(s_t)\,\left(x_ty_t' - x_tx_t'\overline{S}\right)\right],$$
$$\Psi_{0\,j}(k) = \sum_{t\in\{t:\,s_t=k\}}\left[\zeta_j(s_t)\,\left(y_ty_t' - 2\overline{S}'x_ty_t' + \overline{S}'x_tx_t'\overline{S}\right)\right].$$

The conditional posterior distribution of $\zeta_j(k)$ has the following pdf

$$\begin{aligned}\pi(\zeta_j(k) \mid Y_T, S_T, g_j, b_j) &= \Gamma\left(\frac{T_k}{2} + \alpha_\zeta,\ \frac{1}{\widetilde{\zeta}_j(k)/2 + 1/\beta_\zeta}\right) \\ &\propto \zeta_j(k)^{T_k/2+\alpha_\zeta-1}\,\mathrm{e}^{-\left(\widetilde{\zeta}_j(k)/2+1/\beta_\zeta\right)\zeta_j(k)},\end{aligned} \tag{31}$$

[2] Other equations may be Case III or II.

where

$$
\begin{aligned}
\widetilde{\zeta}_j(k) = \bar{a}'_{0,j} \sum_{t\in\{t:\, s_t=k\}} & \left[y_t y_t' - 2\bar{S}' x_t y_t' + \bar{S}' x_t x_t' \bar{S}\right] \bar{a}_{0,j} \\
& - 2\bar{d}'_j \sum_{t\in\{t:\, s_t=k\}} \left[x_t y_t' - x_t x_t' \bar{S}\right] \bar{a}_{0,j} \\
& \qquad + \bar{d}'_j \sum_{t\in\{t:\, s_t=k\}} \left[x_t x_t'\right] \bar{d}_j,
\end{aligned}
$$

and $\bar{d}_j = [\bar{d}'_{j,1} \quad \bar{c}_j]'$.

IV.3.4. *Additional Case.* There may be situations where one is willing to consider identifying restrictions for different lag structures across equations to reduce the number of parameters while imposing no restrictions on the degree of time variation in $a_{0,j}(s_t)$ and $d_j(s_t)$. With little modification, the analytical results derived in the previous sections apply to this case as well. Because such identifying restrictions are still in the form of (9) and (10), it can be seen that the conditional posterior distributions of g_j and b are the same as (28) and (29), except $\Delta_j(k) = I_m \; \forall k \in \{1,\ldots,h\}$ in the expressions for $\widetilde{\Psi}^{-1}_{6j}$ and Ψ_{5j} in Section IV.3.2.

V. Normalization

To obtain accurate posterior distributions of functions of θ (such as long run responses and historical decompositions), we must normalize both the signs of structural equations and the labels of states; otherwise, the posterior distributions will be symmetric with multiple modes, making statistical inferences of interest meaningless. Such normalization is also necessary to achieve efficiency in evaluating the marginal likelihood for model comparison.[3] For both purposes, we normalize the signs of structural equations the same way. Specifically, we use the Waggoner and Zha (2003b) normalization rule to determine the column signs of $A_0(k)$ and $A_+(k)$ for any given $k \in \{1,\ldots,h\}$.

Two other normalizations, scale normalization on $\zeta_j(k)$ and $\lambda_j(k)$ and label normalization on the states, require additional treatment. The rule applied to inference is different from that used for evaluating the marginal likelihood.

V.1. **Error bands.** We first obtain the posterior estimate of θ under the restrictions $\zeta_j(k) = 1$ and $\lambda_j(k) = \mathbf{1}_{h\times 1}$ for all $j \in \{1,\ldots,n\}$ and $k \in \{1,\ldots,h\}$, where the notation $\mathbf{1}_{h\times 1}$ denotes the $h \times 1$ vector of 1's. Let $\hat{\zeta}_j(k)$ and $\hat{\lambda}_j(k)$ be the posterior estimates of $\zeta_j(k)$ and $\lambda_j(k)$.

[3] Note that the marginal data density is invariant to the way parameters are normalized, as long as the Jacobian transformations of the parameters are taken into account explicitly.

We then normalize these posterior estimates around the unit circle as

$$\mathring{\zeta}_j(k) = \frac{\hat{\zeta}_j(k)}{\sum_{i=1}^h \hat{\zeta}_j(i)},\ \forall j \in \{1,\dots,n\}, k \in \{1,\dots,h\}, \tag{32}$$

$$\mathring{\lambda}_{ij}(k) = \frac{\hat{\lambda}_{ij}(k)}{\sqrt{\sum_{i=1}^h \hat{\lambda}_{ij}^2(i)}},\ \forall i,j \in \{1,\dots,n\}, i \in \{1,\dots,h\}, \tag{33}$$

and adjust the other affected parameters accordingly. This normalization simply rescales the posterior estimates and has no material consequences. But for statistical inference, the normalization this way on each posterior draw of $\zeta_j(k)$ and $\lambda_j(k)$ creates a better-behaved normalized posterior distribution.

To continue this scale normalization, we simulate Markov chain Monte Carlo (MCMC) draws of θ as detailed in Section IV with $\zeta_j(k) = 1$ and $\lambda_j(k) = \mathbf{1}_{h\times 1}$ for all $j \in \{1,\dots,n\}$ and $k \in \{1,\dots,h\}$. For each posterior draw of $\zeta_{jj}(k_2)$ and $\lambda_j(k_1)$, we normalize these random values in the manner of (32) and (33) and adjust accordingly the values of the other affected parameters such as A_0 and D.

To perform label normalization, compute $\sum_{t=1}^T P(s_t = k|\hat{\theta})$ for $k = 1,\dots,h$, where $\hat{\theta}$ is the posterior estimate of θ. Reorder the states so that state i corresponds to the state that has the i^{th} largest sum. We then label the simulated states from the posterior distribution as follows. [4]

(1) Let $i = 1$.

(2) For each MCMC draw of states $\tilde{s}_t$ for all $t = 1,\dots,T$, permute the states such that

$$\sum_{t\in\{t:\tilde{s}_t=i\}} P(s_t = i|\hat{\theta}) \geq \sum_{t\in\{t:\tilde{s}_t=k\}} P(s_t = i|\hat{\theta}), \quad k = 2,\dots,h.$$

(3) For $i = 2,\dots,h-1$, repeat the previous step successively.

The normalized draws of θ after rescaling and relabeling are used in constructing the error bands of functions of θ for statistical inference. None of the normalized draws, however, is used in the Gibbs sampling procedure for two reasons. First, it can be seen from (32) and (33) that the scaling normalization leads to the change of the unnormalized prior to the normalized prior that has an unrecognized pdf form (because of the complicated Jacobian term). Second, if the labeling normalization involves a swap of, say, $\lambda_j(1)$ and $\lambda_j(k)$ for $k \neq 1$, one of the swapped priors will be the inverse of a normal prior distribution. In

[4] This label normalization requires the number of permutations only on the order of h^2 and thus is a computationally efficient way to approximate Wald normalization discussed by Hamilton, Waggoner, and Zha 2003 or the normalization that maximizes the correlation of the time path of probabilities of drawn states with the time path of posterior estimates of state probabilities. The latter two normalizations require $h!$ permutations. The normalization described here works well if there are a few enduring states. Otherwise, one could obtain reduced-form residuals of some important variable such as the interest rate and normalize the states from the smallest residual to the largest residual.

both situations, it is very inefficient (if not impossible) to sample from the corresponding normalized conditional posterior distribution at each Gibbs sampling step. Such problems associated with *explicit* changes of prior pdfs do not exist for normalized draws because the Jacobian terms are *implicitly* taken into account after the normalization.

V.2. **Marginal data density.** The marginal data density (or marginal likelihood) is invariant to normalization. To see this point, let ρ be the normalized θ and note

$$\begin{aligned}\pi(Y_T) &= \int \pi(Y_T|\theta)\pi_\theta(\theta)\,d\theta \\ &= \int \pi(Y_T|\rho)\,\pi_\theta(\rho)|\partial\theta/\partial\rho|\,d\rho \\ &= \int \pi(Y_T|\rho)\,\pi_\rho(\rho)\,d\rho.\end{aligned} \tag{34}$$

Since ρ is simply a normalized version of θ, the likelihoods $\pi(Y_T|\theta)$ and $\pi(Y_T|\rho)$ are the same. One can compute the marginal data density through either θ or its normalized parameterization. Without normalization, it would be very inefficient to compute the marginal data density because of the multi-mode nature of the unnormalized posterior distribution. If one works with the normalized parameterization, on the other hand, the implicit Jacobian term $|\partial\theta/\partial\rho|$ may be difficult or even impossible to derive analytically. We recommend some, not all, normalization in computing the marginal data density. Specifically, we simulate MCMC posterior draws of θ with $\zeta_i(k)=1$ and $\lambda_j(k)=\mathbf{1}_{h\times 1}$ for all $j\in\{1,\ldots,n\}$ and $k\in\{1,\ldots,h\}$. For each posterior draw, we apply the label normalization for $k>1$ only. We do not perform the scaling normalization at all. Unlike Section V.1, the normalized draw this way is used for the next Gibbs step. In this normalized Gibbs process, all priors are the identical and no Jacobian terms are involved.

VI. Model Fit

To select a model that fits best to the data, we need to estimate the marginal data density $\pi(Y_T)$ for each model and then compare the marginal data densities among different models. We apply both the modified harmonic mean method (MHM) of Gelfand and Dey 1994 and the method of Chib and Jeliazkov 2001. The MHM method is quite efficient for most models considered in this paper, but it may give unreliable estimates for some models whose posterior distributions have multiple modes. In such a situation, we also use the Chib and Jeliazkov to check the robustness of the estimate.

The method of Chib and Jeliazkov 2001 utilizes the following identity

$$\begin{aligned}\pi(Y_T) &= \int \pi(Y_T \mid \theta)\pi(\theta)\,d\theta \\ &= \frac{\pi\left(Y_T \mid \theta^*\right)\pi\left(\theta^*\right)}{\pi\left(\theta^* \mid Y_T\right)},\end{aligned}$$

where θ^* can be any point in the support of the θ parameter space. The prior ordinate $\pi(\theta^*)$ is readily available by direct calculation. It is also straightforward to compute the likelihood ordinate $\pi(Y_T \mid \theta^*)$ using (16). The evaluation of the posterior ordinate $\pi(\theta^* \mid Y_T)$ demands intensive computation because it involves reduced MCMC runs for the following $n+3$ blocks of conditional probability densities

$$\begin{aligned}\pi(\theta^* \mid Y_T) = \prod_{j=1}^{n} \pi\left(b_j^* \mid Y_T, b_1^*, \ldots, b_{j-1}^*\right) \\ \pi(g^* \mid Y_T, b^*)\, \pi(\gamma^* \mid Y_T, g^*, b^*)\, \pi(w^* \mid Y_T, \gamma^*, g^*, b^*)\,. \end{aligned} \tag{35}$$

The estimate of the posterior ordinate

$$\pi\left(b_i^* \mid Y_T, b_1^*, \ldots, b_{i-1}^*\right),$$

requires simulating random values of $\{b_i, b_{i+1}, \ldots, b_n, S_T, \delta\}$ with the reduced conditional probability densities:

$$\begin{gathered}\pi\left(b_i \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_{i+1}, \ldots, b_n, S_T, \delta\right), \\ \pi\left(b_{i+1} \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_i, b_{i+2}, \ldots, b_n, S_T, \delta\right), \\ \vdots \\ \pi\left(b_n \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_i, \ldots, b_{n-1}, S_T, \delta\right), \\ \pi\left(S_T, \delta \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_i, \ldots, b_n\right),\end{gathered}$$

and simulating random values of $\{b_{i+1}, \ldots, b_n, S_T, \delta\}$ with the reduced conditional probability densities:

$$\begin{gathered}\pi(b_{i+1} \mid Y_T, b_1^*, \ldots, b_i^*, b_{i+2}, \ldots, b_n, S_T, \delta), \\ \pi(b_{i+2} \mid Y_T, b_1^*, \ldots, b_i^*, b_{i+1}, b_{i+3}, \ldots, b_n, S_T, \delta), \\ \vdots \\ \pi(b_n \mid Y_T, b_1^*, \ldots, b_i^*, b_{i+1}, \ldots, b_{n-1}, S_T, \delta), \\ \pi(S_T, \delta \mid Y_T, b_1^*, \ldots, b_i^*, b_{i+1}, \ldots, b_n), \\ J\left(b_i \mid b_i^*, Y_T, b_1^*, \ldots, b_{i-1}^*, b_{i+1}, \ldots, b_n, S_T, \delta\right).\end{gathered}$$

A choice of b_i^* proves less straightforward. Initially, we set $b_i^* = \hat{b}_i$. Because the joint posterior distribution of $b_1, \ldots, b_n$ is highly non-Gaussian in simultaneous equation models like this, the posterior estimate $\hat{b}_i$ may turn out to be in the very low probability region of the *conditional* or *marginal* distribution of b_i, rendering $\hat{b}_i$ to be a poor choice. To choose an efficient point, we perform an additional reduced Gibbs run to set b_i^* at the value of b_i'

that maximizes the ratio in the jumping probability of the Metropolis algorithm

$$\frac{\pi(b_i' \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_{i+1}, \ldots, b_n, S_T, \delta)}{\pi(b_i \mid Y_T, b_1^*, \ldots, b_{i-1}^*, b_{i+1}, \ldots, b_n, S_T, \delta)}.$$

For other posterior ordinates in (35), we use the method proposed by Chib 1995. Estimation of $\pi\left(g^* \mid Y_T, b^*\right)$ involves simulating random values of $\{S_T, g, \gamma, p\}$ from $\pi\left(S_T, g, \gamma, p \mid Y_T, b^*\right)$. The value of g^* is set to $(1/Q)\sum_{q=1}^{Q} \widetilde{g}_j^{(q)}$. Estimation of $\pi\left(\gamma^* \mid Y_T, b^*, g^*\right)$ involves simulating random values of $\{S_T, \gamma, p\}$ from $\pi\left(S_T, \gamma, p \mid b^*, g^*\right)$. Because $\pi\left(\gamma \mid Y_T, b^*, g^*, S_T^{(q)}, p^{(q)}\right)$ is non-Gaussian for Case II, we generate several random draws of γ and set γ^* to be the draw that gives that highest value of $\pi\left(\gamma \mid Y_T, b^*, g^*, S_T^{(q)}, p^{(q)}\right)$ in our reduced Gibbs run. Finally, $\pi\left(p^* \mid Y_T, b^*, g^*, \gamma^*\right)$ is estimated with random values of $\{S_T, p\}$ simulated from $\pi\left(p, S_T \mid Y_T, b^*, g^*, \gamma^*\right)$ and p^* is set to the average value of the posterior means:

$$p_{ik}^* = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\alpha_{ik} + n_{ik}^{(q)}}{\sum_{i=1}^{h}(\alpha_{ik} + n_{ik}^{(q)})}.$$

The value of θ^* so chosen changes randomly as the number of MC draws increases. This feature safeguards our procedure from producing an erroneous evaluation of $\pi(Y_T)$ likely to occur with any fixed value of θ that happens to be in the tail of one of the n conditional posterior distributions in (35). The computational cost can be quite high for large models with the non-Gaussian shape of the posterior distribution.

The modified harmonic mean (MHM) method of Gelfand and Dey 1994 seems to be a more efficient procedure, at least for the time-varying identified models studied here. Denote the support of $\pi(\theta|Y_T)$ by Θ_π. Let $\mathfrak{p}(\theta)$ be a *weighting* function that must be a pdf (*not* kernel) whose support is contained in Θ_π, Gelfand and Dey 1994 observe that

$$\pi(Y_T)^{-1} = \int_{\Theta_\pi} \frac{\mathfrak{p}(\theta)}{\pi(Y_T \mid \theta)\pi(\theta)}\, \pi(\theta \mid Y_T) d\theta. \tag{36}$$

A numerical evaluation of the integral on the right hand side of (36) can be done through the Monte Carlo integration

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

where $\theta^{(i)}$ is the i^{th} draw of θ from the posterior distribution $\pi(\theta \mid Y_T)$. A popular choice of $\mathfrak{p}(\theta)$ is a truncated normal pdf (Geweke 1999), which turns out to be a poor choice for our problems because each element of the parameter vector w in the transition matrix is bounded between 0 and 1. To deal with this problem, we regroup the parameter vector θ into two blocks. The first block, denoted by θ_1, contains all the parameters except w and the second block consists of w only. Denote the supports for these two blocks of parameters by

$\Theta_{1,\pi}$ and $\Theta_{2,\pi}$. Let $\bar{\theta}_{1,N_1}$ and $\overline{\Omega}_{1,N_1}$ be the sample posterior mean and the sample posterior covariance matrix of θ_1 with N_1 MCMC draws. For $\mathfrak{r} \in (0,1)$, define

$$\Theta_{N_1,\mathfrak{r}} = \left\{ \theta_1 : \, (\theta_1 - \bar{\theta}_{1,N_1})' \overline{\Omega}^{-1}_{1,N_1} (\theta_1 - \bar{\theta}_{1,N_1}) < \chi^2_{\mathfrak{r}}(\mathfrak{n}_1) \right\},$$

where $\mathfrak{n}_1$ is the dimension of θ_1 and $\chi^2_{\mathfrak{r}}(\mathfrak{n}_1)$ is the inverse of the chi-squared cdf with $\mathfrak{n}_1$ degrees of freedom for the probability $\mathfrak{r}$. Define the truncated normal pdf $\mathfrak{p}^{\star}_1(\theta_1)$ as

$$\mathfrak{p}^{\star}_1(\theta_1) = \frac{\chi(\theta_1 \in \Theta_{N_1,\mathfrak{r}})}{\mathfrak{r}} \frac{1}{(2\pi)^{\mathfrak{n}_1/2}} \left| \overline{\Omega}_{1,N_1} \right|^{-1/2} \exp\left\{ -\frac{1}{2} (\theta_1 - \bar{\theta}_{1,N_1})' \overline{\Omega}^{-1}_{1,N_1} (\theta_1 - \bar{\theta}_{1,N_1}) \right\}. \tag{38}$$

Recall that the indicator function $\chi(\,)$ returns 1 when the statement in parentheses is true and 0 otherwise. For Case II, since the parameter $\zeta_{jj}(k_2)$ is always greater than 0, the support $\Theta_{N_1,\mathfrak{r}}$ is not contained in $\Theta_{1,\pi}$. Following Geweke 1999, we redefine the support to be $\Theta_{N_1,\mathfrak{r}} \cap \Theta_{1,\pi}$ and calculating a rescaling constant, denoted by $q_{N_1,\mathfrak{r}}$, for $\mathfrak{p}^{\star}_1(\theta_1)$. The constant $q_{N_1,\mathfrak{r}}$ can be approximately by the proportion of i.i.d. draws of θ_1 from (38) that fall in $\Theta_{1,\pi}$. This approximation is very accurate so long as $q_{N_1,\mathfrak{r}}$ is not close to 0. The weighting function for θ_1 is

$$\mathfrak{p}_1(\theta_1) = \frac{\chi(\theta_1 \in \Theta_{1,\pi})}{q_{N_1,\mathfrak{r}}} \mathfrak{p}^{\star}_1(\theta_1).$$

Every parameter in the second block p is bounded between 0 and 1. If the weighting function $\mathfrak{p}_2(w)$ were truncated-normal, the rescaling constant would be very close to 0. Thus, we use a Dirichlet distribution as the weighting function of p_j $\forall j \in \{1,\ldots,h\}$ and $\forall x \in \{1,2\}$. Let $\bar{p}_{ij}$ and $\overline{V}(p_{ij})$ be the sample posterior mean and the sample posterior variance of p_{ij} with N_2 MCMC draws. The weighting function $\mathfrak{p}_2(p_j)$ is defined as

$$\mathfrak{p}_2(p_j) = \mathcal{D}(\bar{\kappa}_{1j}, \ldots, \bar{\kappa}_{hj}),$$

where

$$\bar{\kappa}_{ij} = \begin{cases} \bar{\bar{p}}_{ij} \bar{p}_{ij} & \text{if } \bar{\bar{p}}_{ij} > 0 \\ 1 & \text{if } \bar{\bar{p}}_{ij} < 0 \end{cases},$$

$$\bar{\bar{p}}_{ij} = \frac{\bar{p}_{ij}(1 - \bar{p}_{ij})}{\overline{V}(p_{ij})} - 1.$$

It can be shown that the mean and variance for $\mathfrak{p}_2(p_j)$ are exactly equal to $\bar{p}_{ij}$ and $\overline{V}(p_{ij})$.

The weighting function for evaluating (37) is

$$\mathfrak{p}(\theta) = \mathfrak{p}_1(\theta_1) \prod_{j=1}^{h} \mathfrak{p}_2(p_j).$$

VII. CONCLUSION

This paper extends the existing MCMC simulation methods to a system of simultaneous equations with hidden Markov chains. It overcomes analytical and computational difficulties that arise when one restricts the degree of time variation on the system. We derive the probability density functions of conditional posterior distributions used for the MCMC simulations and develope software that enables one to obtain the solution on a standard PC desktop. Sims and Zha 2004 have applied this method to addressing various questions regarding monetary policy. Despite intensive computation needed to get reliable results, we hope that further innovations in numerical methods and computer technology will make our method easier for applied researchers to use.

REFERENCES

CHIB, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

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

CHIB, S., AND I. JELIAZKOV (2001): "Marginal Likelihood From the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96(453), 270–281.

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.

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

HAMILTON, J. D., D. F. WAGGONER, AND T. ZHA (2003): "Normalization in Econometrics," Manuscript, University of California (San Diego) and Federal Reserve Bank of Atlanta.

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

RUDEBUSCH, G. D., AND L. E. SVENSSON (2002): "Eurosystem monetary targeting: lessons from U.S. data," *European Economic Review*, 46(3), 417–442.

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

——— (2004): "Were There Regime Switches in US Monetary Policy?," Manuscript, Princeton University and Federal Reserve Bank of Atlanta.

TAYLOR, J. B. (1999): "The robustness and efficiency of monetary policy rules as guidelines for interest rate setting by the European Central Bank," *Journal of Monetary Economics*, 43, 655–679.

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.

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY,FEDERAL RESERVE BANK OF ATLANTA
E-mail address: sims@princeton.edu, tzha@mindspring.com