Finance and Economics Discussion Series
Divisions of Research & Statistics and Monetary Affairs
Federal Reserve Board, Washington, D.C.

Estimating (Markov-Switching) VAR Models without Gibbs Sampling: A Sequential Monte Carlo Approach

Mark Bognanni and Edward P. Herbst

2015-116

Please cite this paper as:
Bognanni, Mark, and Edward P. Herbst (2015). "Estimating (Markov-Switching) VAR Models without Gibbs Sampling: A Sequential Monte Carlo Approach," Finance and Economics Discussion Series 2015-116. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.116.

Estimating (Markov-Switching) VAR Models without Gibbs Sampling: A Sequential Monte Carlo Approach

Mark Bognanni*
Federal Reserve Bank of Cleveland

Edward Herbst†
Federal Reserve Board of Governors

This Draft: December 10, 2015

Abstract. Vector autoregressions with Markov-switching parameters (MS-VARs) fit the data better than do their constant-parameter predecessors. However, Bayesian inference for MS-VARs with existing algorithms remains challenging. For our first contribution, we show that Sequential Monte Carlo (SMC) estimators accurately estimate Bayesian MS-VAR posteriors. Relative to multi-step, model-specific MCMC routines, SMC has the advantages of generality, parallelizability, and freedom from reliance on particular analytical relationships between prior and likelihood. For our second contribution, we use SMC's flexibility to demonstrate that the choice of prior drives the key empirical finding of Sims, Waggoner, and Zha (2008) as much as does the data.

JEL: C11, C18, C32, C52, E3, E4, E5
Keywords: Vector Autoregressions, Sequential Monte Carlo, Regime-Switching Models, Bayesian Analysis

We thank Todd Clark, Ron Gallant, Eric Leeper, James Hamilton, Giorgio Primiceri, Juan F. Rubio-Ramírez and Frank Schorfheide for helpful comments and conversations. We also thank Dan Waggoner for a particularly helpful conference discussion and the other participants of various conferences for their feedback. The views expressed in this paper do not necessarily reflect those of Federal Reserve Bank of Cleveland, the Federal Reserve Board of Governors, or the Federal Reserve System.

*Correspondence: Mark Bognanni, Federal Reserve Bank of Cleveland, PO Box 6387, Cleveland, OH 44101-1387. Email: email.markbognanni@gmail.com, Web: http://markbognanni.com.

†Correspondence: Board of Governors of the Federal Reserve System, 20th Street and Constitution Avenue N.W., Washington D.C. 20551. Email: edward.p.herbst@frb.gov.

1. Introduction

The use of vector autoregressions (VARs) has grown steadily since Sims (1980) and VARs now serve as a vital element of the macroeconomist's toolkit. Bayesian methods have come to dominate the literature on VAR applications for two main reasons. Firstly, VARs' large number of parameters relative to data in typical macroeconomic applications has led researchers to favor the additional parameter discipline that Bayesian priors can provide. Secondly, researchers have developed methods that make Bayesian estimation of VARs straightforward. Posterior sampling is often the most challenging aspect of Bayesian inference, but for VARs a known family of priors yields (conditional) posteriors amenable to an efficient posterior sampling algorithm called the Gibbs sampler.

In recent years the interests of economists have moved beyond the basic VAR to extensions with time-varying parameters. One such extension, and the focus of this paper, is the VAR with Markov-switching parameters (MS-VAR) pioneered by Sims and Zha (2006). As a byproduct of their inquiry into the cause of the "Great Moderation," Sims and Zha (2006) document superior data fit of every MS-VAR they estimate relative to the constant-coefficient VAR (CC-VAR). Yet, despite the data's demonstrated preference for Markov-switching models, few researchers have used MS-VARs in economic applications.

We suspect that the sparse use of MS-VARs owes to the complicatedness of the estimation process; MS-VARs do not admit MCMC samplers that possesses the efficiency or simplicity of their constant-parameter predecessors. Sims, Waggoner, and Zha (2008) expound upon the methods used in Sims and Zha (2006) and describe the following four-step procedure for MS-VAR estimation and model comparison. First, search in the model's high dimensional parameter space for the posterior mode, from which one initializes the MCMC algorithm. Second, code and deploy a highly model-specific Gibbs sampler, which relies on so-called Metropolis-within-Gibbs steps. Third, impose both sign and state-labeling normalizations on the posterior draws at the post-processing stage, which is necessary for the stability of the estimator in step 4. Fourth and finally, code a nontrivial augmentation of the modified harmonic mean (MHM) algorithm for

estimating the marginal data density (MDD), which is necessary for Bayesian model comparison.

In a recent paper investigating the macroeconomic effects of financial crises, and which is a notable exception to the hesitance of economists to use MS-VARs, Hubrich and Tetlow (2015) use the algorithm of Sims et al. (2008) and summarize the length of the process as follows, "Computation of a specification's posterior mode and the marginal data density takes a minimum of 6 hours in clock time and can take as long as 8 days, depending on the specifics of the run. Adding lags, imposing restrictions on switching on variances and restricting switching in equation coefficients is costly in terms of computing times."[1] Of course, even at the end of this process uncertainty remains about whether or not one found the true posterior mode in the first step.

Motivated by these difficulties, we estimate MS-VARs using an alternative class of algorithms called Sequential Monte Carlo (SMC). Our SMC algorithm begins by propagating a set of "particles" from the prior distribution, where each particle contains a vector of values for the model's parameters. The algorithm then moves and reweights the particles to iteratively approximate a sequence of distributions, each of which combines the prior with partial information from the likelihood. Each distribution in the sequence uses more information from the likelihood than its predecessor and the algorithm concludes once the full likelihood has been incorporated. When the algorithm concludes, the researcher has a set of particles that serve as a discrete distribution approximating the model's true posterior.

Using SMC to estimate MS-VARs allows us to sidestep many of the aforementioned challenges. In particular, SMC has four key features that make it attractive for our purposes. First, the algorithm's initialization with many random draws from the prior negates both the need for a time-consuming mode search and any risk of residual dependence on a unique starting value. Second, the algorithm is generic and does not rely on any particular analytical convenience of the posterior. Rather, one needs only the ability to evaluate a posterior kernel pointwise, which

[1] The quotation comes from the Online Appendix of Hubrich and Tetlow (2015), in which the authors describe the details of the estimation process.

negates the need to code a multi-step, model-specific Gibbs Sampler or for the model to even allow one to derive a Gibbs Sampler.[2] Third, the algorithm generates an estimate of the model's MDD as a byproduct, which negates the need for time consuming post-processing and coding a unique MHM algorithm. Fourth, unlike MCMC algorithms, which must run serially, SMC's computations admit almost arbitrary parallelization, which makes SMC an increasingly practical approach as modern computer architectures continue to expand their parallel potential.

While the aforementioned properties of SMC are desirable, researchers have yet to demonstrate that SMC can effectively estimate the posteriors of high-dimensional time-series models when using a computationally feasible quantity of particles. Our first contribution is to show that SMC algorithms can indeed perform this task. We show this by demonstrating the algorithm's ability to estimate MDDs in two settings in which we know the true MDD in closed form, settings which thus provide a gold standard for assessing SMC's performance.

The first test setting is the familiar reduced-form CC-VAR with conjugate prior, which we consider the simplest possible test relevant to our interests. For the CC-VAR we show solid performance by SMC under a variety of choices for the algorithm's tuning parameters and highlight a few small changes to existing SMC implementations that yield particularly dramatic performance improvements for VARs. One can use our change to the algorithm generally, but its performance gains for VARs owes to its improved accounting for the correlation structure among parameters that is typically present in both VAR priors and posteriors. The second test setting is a mixture of reduced-form CC-VAR posteriors, which imitates the multi-modality of more complicated models. Remarkably, when confronted with the multi-modal posterior the SMC algorithm estimates MDDs as well or better than standard MDD estimators *even when we provide the standard*

[2]In principle one can use the same basic SMC algorithm to estimate reduced-form VARs, structural and exactly-identified VARs, structural and over-identified VARs, VARs with steady-state priors, and MS-VARs, each of which relies on a unique posterior sampler when using MCMC for estimation. We consider the algorithm's genericness to be an argument in favor of its use. As Geweke (2004) emphasizes, reliance on model-specific Gibbs samplers for posterior simulation typically involves a lengthy processes of tedious algebra and coding, both of which lend themselves well to making difficult-to-detect errors.

estimators with iid draws from the posterior.

Having established SMC's viability for high dimensional time series models, we exploit its genericness to make our second contribution: a demonstration of the importance of the prior for Bayesian model selection among MS-VAR specifications. We use SMC to estimate a suite of MS-VARs similar to those of Sims et al. (2008) under a variety of priors. The main empirical result of both Sims and Zha (2006) and Sims et al. (2008) is that the data clearly favors MS-VARs with regime switching only in shock variances, and not other economic dynamics. In this sense, the authors come down on the side of "good luck" in the debate over the cause of the "Great Moderation." We find that, when using the best fitting prior, the posterior probability that the best model includes *both* time-varying shock variances *and* time-varying economic dynamics shifts from 6% to 43%. This result suggests that prior choice deserves particularly careful attention when comparing competing MS-VAR models, a point that should be well taken generally by researchers using the Bayesian approach to comparing different specifications of densely parameterized models.

Lastly, our positive results on SMC's usefulness in practical applications constitute a contribution of general interest to economists. SMC provides a way forward when MCMC algorithms are either inefficient or reliant on the use of undesirable priors for efficiency. As economists estimate increasingly complicated models, it seems less likely that they will find priors that both yield a posterior amenable to Gibbs sampling and perfectly represent economists' *a priori* beliefs. Indeed, the existence of such a prior seems more likely to be the result of divine coincidence than the norm.

With regards to the estimation algorithm, our paper builds on the recent work by Durham and Geweke (2012) and Herbst and Schorfheide (2014), who also explore the use of SMC algorithms for estimating econometric models. Durham and Geweke (2012) emphasize the massive parallelization possibilities for SMC algorithms, particularly for use with GPUs. Herbst and Schorfheide (2014) apply SMC algorithms to the estimation of DSGE models and show that DSGE-model posteriors can possess multi-modality that random walk Metropolis-Hastings algorithms fail to uncover in reasonable amounts of time. We also make use of a

number of advances from the statistics literature, on which we elaborate further in the next section.

From here the rest of the paper proceeds as follows. In Section 2 we describe the estimation problem, our estimation algorithm, and its place within the larger SMC literature. In Section 3 we demonstrate the algorithm's effectiveness in settings in which we have closed-form expressions for the objects we estimate. In Section 4 we describe the MS-VAR models, the three priors we consider, and our estimation results. In Section 5 we conclude.

2. Sequential Monte Carlo Methods

Let θ be the parameters of a model and Y be the data relevant for the model's likelihood function. The Bayesian researcher is interested in the posterior density $p(\theta|Y)$, which is given by

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)}, \quad \text{where } p(Y) = \int p(Y|\theta)p(\theta)d\theta \,, \tag{1}$$

$p(\theta)$ denotes the prior density, and $p(Y|\theta)$ denotes the likelihood. The term $p(Y)$ is known as the "marginal data density" (MDD) or "marginal likelihood", an important measure of model fit in Bayesian statistics. For ease of exposition, in this section we abbreviate these objects by $\pi(\theta) = p(\theta|Y)$, $f(\theta) = p(Y|\theta)p(\theta)$, and $Z = p(Y)$, which gives an equivalent expression to (1) as

$$\pi(\theta) = \frac{f(\theta)}{Z} \,. \tag{2}$$

An unfortunate feature of Bayesian inference is that in most applications of practical interest ones does not know the moments of $\pi(\theta)$ in closed-form. Hence, posterior inference often relies on devising a method to sample from $\pi(\theta)$. To put these facts in the context of our applications, note that for VARs the literature has previously concentrated on families of priors that induce a posterior such that either $\pi(\theta)$ can be sampled directly or there exists a partitioning of the parameters $\theta = [\theta^1, \ldots, \theta^n]$ such that each conditional posterior can be sampled directly,

yielding draws from the posterior through a Gibbs sampler.[3] While for MS-VARs no known priors induce posteriors from which we can sample with a pure Gibbs Sampler or for which we know Z in closed-form.

2.1 Overview of the Sequential Monte Carlo Method

In our applications we use SMC algorithms to approximate $\pi(\theta)$ and Z.[4] Since importance sampling (IS) serves as the keystone of SMC, we begin our description of SMC methods with a brief description of IS.[5] IS approximates the target density $f(\cdot)$ by a different, easy-to-sample density $g(\cdot)$, which is sometimes known as the "source density." IS is based on the identity

$$E_\pi[h(\theta)] = \int h(\theta)\pi(\theta)d\theta = \frac{1}{Z}\int_\Theta h(\theta)w(\theta)g(\theta)d\theta, \quad \text{where} \quad w(\theta) = \frac{f(\theta)}{g(\theta)}, \tag{3}$$

If $\theta^i \overset{iid}{\sim} g(\theta)$, $i = 1, \ldots, N$, then, under suitable regularity conditions—see Geweke (1989)—the Monte Carlo estimate

$$\bar{h} = \sum_{i=1}^{N} h(\theta^i)\tilde{W}^i, \quad \text{where} \quad \tilde{W}^i = \frac{w(\theta^i)}{\frac{1}{N}\sum_{j=1}^{N} w(\theta^j)}, \tag{4}$$

converges almost surely (a.s.) to $E_\pi[h(\theta)]$ as $N \longrightarrow \infty$. The set of pairs $\{(\theta^i, \tilde{W}^i)\}_{i=1}^N$ provides a discrete distribution that approximates $\pi(\theta)$. The $\tilde{W}^i$'s are known as the (normalized) importance weights assigned to each particle value θ^i. The distance between $g(\cdot)$ and $f(\cdot)$ determines the accuracy of the approximation (per particle) and the uniformity (or lack thereof) of the distribution of weights reflects the size of this distance. If the distribution of weights is very uneven,

[3] Researchers usually estimate VARs under a conjugate prior of the Normal-Inverse Wishart form. One can efficiently estimate structural VARs that have linear over-identifying restrictions by using the algorithm described in Waggoner and Zha (2003a).

[4] We describe only our particular algorithm here. Chopin (2002), Del Moral, Doucet, and Jasra (2006), Creal (2012), and Herbst and Schorfheide (2014) offer additional details on SMC implementation.

[5] Indeed, the SMC method we use in this paper is sometimes known as Iterated Batch Importance Sampling.

the Monte Carlo approximation $\bar{h}$ is inaccurate, because only a few particles contribute meaningfully to the estimate. On the other hand, uniform weights arise if $g(\cdot) \propto f(\cdot)$, which means that we are sampling directly from $\pi(\theta)$.

Unfortunately, constructing "good" importance distributions, $g(\cdot)$, is difficult when the econometrician knows little about the shape of f.[6] The SMC algorithm we use attacks this problem by recursively building particle approximations to a *sequence* of distributions, starting from a known distribution, the prior, and then slowly adding information from the likelihood until we have obtained a particle approximation to the posterior. Specifically, we use n to index a sequence of distributions of the form

$$\pi_n(\theta) = \frac{f_n(\theta)}{Z_n} = \frac{[p(Y|\theta)]^{\phi_n} p(\theta)}{\int [p(Y|\theta)]^{\phi_n} p(\theta) d\theta}, \quad n = 1, \ldots, N_\phi. \tag{5}$$

and choose an increasing sequence of values for the scaling parameter, ϕ_n, such that $\phi_1 = 0$ and $\phi_{N_\phi} = 1$. The choice of $\phi_1 = 0$ means that the initial target distribution, $\pi_1(\theta)$, is simply the prior, $p(\theta)$. Hence, one initializes the algorithm by propagating the particles as random draws from the prior. The choice of $\phi_{N_\phi} = 1$ means that the final target distribution, $\pi_{N_\phi}(\theta)$, is the posterior. Thus the final particles approximate the distribution of interest to the researcher.[7]

2.2 The Sequential Monte Carlo Algorithm

Algorithm 1 describes the three steps to construct a particle approximation to π_n from a particle approximation to π_{n-1}, in the terminology of Chopin (2002). The general form of Algorithm 1 is the same as the one used in Herbst and Schorfheide (2014), but we describe its key features here for completeness. The

[6] There is a history of using importance sampling techniques for VAR estimation when direct samplers and Gibbs samplers are unavailable, namely Leeper, Sims, Zha, Hall, and Bernanke (1996), Uhlig (1997), and Kadiyala and Karlsson (1997) used importance samplers for parts of VAR posteriors. However since the late 1990s researchers estimating VARs have largely abandoned this approach because of the difficulty of finding a good "g," which resulted in inefficient algorithms.

[7] The "likelihood tempering" formulation is not the only avenue one could have pursued. For example Durham and Geweke (2012) propose a GPU-based SMC algorithm as a blackbox for many time series economic models with $f_n(\theta) = p(Y_{1:n}|\theta)p(\theta)$. Durham and Geweke (2012)'s "data tempering" approach is attractive for obtaining on-line parameter estimates.

Algorithm 1: Simulated Tempering SMC

Initialization. ($\phi_1 = 0$). Draw the initial particles from the prior:

$$\theta_1^i \overset{iid}{\sim} p(\theta), \quad W_1^i = 1, \quad i = 1, \dots, N.$$

for $n = 2, \dots, N_\phi$ **do**

1. Correction. Reweight the particles from stage $n-1$ by defining the incremental and normalized weights

$$\tilde{w}_n^i = [p(Y|\theta_{n-1}^i)]^{\phi_n - \phi_{n-1}}, \quad \tilde{W}_n^i = \frac{\tilde{w}_n^i W_{n-1}^i}{\frac{1}{N}\sum_{i=1}^N \tilde{w}_n^i W_{n-1}^i}, \quad i = 1, \dots, N.$$

2. Selection. Compute the effective sample size

$$ESS_n = N \Big/ \left(\frac{1}{N} \sum_{i=1}^N (\tilde{W}_n^i)^2 \right)$$

if $ESS_n < N_{part}/2$ **then**

Resample the particles via multinomial resampling and reinitialize the weights to uniform, i.e.

$$W_n^i = 1, \quad \hat{\theta}_n^i \sim \{\theta_{n-1}^j, \tilde{W}_n^j\}_{j=1,\dots,N}, \quad i = 1, \dots, N$$

else

$$W_n^i = \tilde{W}_n^i, \quad \hat{\theta}_n^i = \theta_{n-1}^i$$

end

3. Mutation. Propagate each particle $\{\hat{\theta}_n^i, W_n^i\}$ via M steps of an MCMC algorithm with transition density $\theta_n^i \sim K_n(\theta_n|\hat{\theta}_n^i; \zeta_n)$ and stationary distribution $\pi_n(\theta)$. (See Algorithm 2 for details and the definition of ζ_n).

end

Compute posterior moments. An approximation of $\mathbb{E}_{\pi_n}[h(\theta)]$ is given by

$$\bar{h}_{n,N} = \frac{1}{N} \sum_{i=1}^N h(\theta_n^i) W_n^i. \tag{6}$$

This approximation is valid using the particle approximations, $\{\theta_{n-1}^i, \tilde{W}_n^i\}_{i=1}^{N_{part}}$, $\{\hat{\theta}_n^i, W_n^i\}_{i=1}^{N_{part}}$ and $\{\theta_n^i, W_n^i\}_{i=1}^{N_{part}}$ after the correction, selection, and mutation step, respectively.

algorithm initializes with particles sampled from $p(\theta)$ and assigned uniform weights. We then enter the recursions. We enter any stage n of the recursion with a particle approximation $\{\theta_{n-1}, \tilde{W}^i_{n-1}\}_{i=1}^{N_{part}}$ of π_{n-1}. In the first step of stage n, the *correction* step, the particles are reweighted according to π_n. This is an importance sample of π_n using π_{n-1} as the proposal distribution. In the second step, *selection*, if the sample is unbalanced in the sense that only a few particles have meaningful weight, the particles are rejuvenated using multinomial resampling. This process ensures that the sampler avoids the well-known issue of particle impoverishment. On the other hand, the resampling itself induces noise into the simulation, and so we avoid doing it unless necessary. In the third and final step, *mutation*, particles are moved around the parameter space, using M iterations of a Metropolis-Hastings algorithm on each individual particle.

The last step, mutation, is crucial. Mutation allows particles to move towards areas of higher density of π_n and ensures diversity across replicated particles when resampling occurs during the selection step. Were the algorithm to run without mutation, repeated resampling of the corrected particles would leave only a few unique values surviving until the final stage, resulting in a poor approximation to the posterior.

From a computational perspective, a point to stress about the mutation step is that each particle operates independently of one another, in a sense forming N_{part} independent Markov chains. This stands in contrast to MCMC algorithms, which rely on a single chain. The independence of particles during mutation allows us to exploit parallel computations during the mutation step, which provides the benefit of greatly speeding up the algorithm, as highlighted by both Durham and Geweke (2012) and Herbst and Schorfheide (2014).

We follow Herbst and Schorfheide (2014) in our specification for the tempering schedule, $\{\phi_n\}_{n=1}^{N_\phi}$, and choose a schedule which follows

$$\phi_n = \left(\frac{n-1}{N_\phi - 1}\right)^\lambda. \tag{7}$$

The hyperparameter $\lambda(> 0)$ controls the rate at which "information" from the likelihood is added to the sampler. If $\lambda = 1$, then the schedule is linear, and, very

roughly speaking, each stage has the same contribution. We use $\lambda > 1$ which means that we add only small increments of the likelihood to the prior in the early stages of the sampler and add larger increments in the later stages. We discuss the role of λ in more detail in Section 3, in which we test the algorithm under various choices for the tuning parameters.

2.3 MCMC Transition Kernel

Algorithm 1 presents the generic SMC algorithm for estimating Bayesian models, but does not specify the exact nature of the MCMC transition kernel used for particle mutation. As we show in Section 3, the form of the MCMC kernel can crucially affect the performance of the sampler. Our base mutation kernel is a block random walk Metropolis-Hasting (RWMH) sampler, detailed in Algorithm 2. Block MH algorithms have been useful in the estimation of DSGE models (see, for example, Chib and Ramamurthy (2010) and Herbst (2012)). Breaking the parameter vector into blocks reduces the dimensionality of the target density for each MCMC step, making it easier to well approximate it by the proposal density.

A key consideration affecting the efficiency of any RW-MH algorithm is the construction of the proposal variance. Our choice of proposal covariance departs from Herbst and Schorfheide (2014) in a simple but important way; we use the multivariate normal approximation to the *conditional* variance for block b, while Herbst and Schorfheide (2014) use the estimate of the *marginal* variance for the block b parameters. To be more precise, we use

$$\hat{\Sigma}_{b,n} = [\hat{\Sigma}_n]_{b,b} - [\hat{\Sigma}_n]_{b,-b}[\hat{\Sigma}_n]^{-1}_{-b,-b}[\hat{\Sigma}_n]_{-b,b} \,, \tag{8}$$

rather than

$$\hat{\Sigma}_{b,n} = [\hat{\Sigma}_n]_{b,b} \,. \tag{9}$$

The marginal variance ignores the relationship between the parameters in block b and the other "conditioning" parameters, which makes it a poor choice for (MS)VARs because of the nontrivial correlation structures in standard priors and

Algorithm 2: Mutation Step

Let $\{B_n\}_{n=2}^{N_\phi}$ be a sequence of random partitions of the parameter vector. For a given partition B_n, let b denote the block of the parameter vector so that $\theta^i_{b,n}$ refers to the b elements of the ith particle. Further let $\theta^i_{<b,n}$ the denote the subpartition of B_n refering to elements of θ^i_n partitioned before the bth set and so on.

At each stage, n, obtain a particle estimate of the covariance of the parameters after selection but before mutation,

$$\hat{\Sigma}_n = \sum_{i=1}^{N_{part}} W^i_n(\hat{\theta}^i_n - \hat{\mu}_n)(\hat{\theta}^i_n - \hat{\mu}_n)' \text{ with } \hat{\mu}_n = \sum_{i=1}^{N_{part}} W^i_n \hat{\theta}^i_n.$$

Denote a covariance matrix for the b-th block, at stage n, which is some function $\zeta(.)$ of $\hat{\Sigma}_n$ as,

$$\hat{\Sigma}_{b,n} = \zeta(\hat{\Sigma}_n).$$

We consider two different functions $\zeta(.)$, which we describe, and compare the performance of, in the text.

Let M be an integer (≥ 1) defining the number of Metropolis-Hastings steps in the mutation stage. Introduce an additional subscript m so that $\theta^i_{m,b,n}$ refers to the bth block of the nth stage, ith particle after m Metropolis-Hastings steps. Set $\theta^i_{0,b,n} = \hat{\theta}^i_{b,n}$.

for $m = 1, \ldots, M$ **do**

  for $b \in B_n$ **do**

    1. Draw a proposal $\theta^*_b \sim N\left(\theta^i_{m-1,b,n}, \hat{\Sigma}_{b,n}\right)$.

       Denote $\theta^* = \left[\theta^i_{m,<b,n}, \theta^*_b, \theta^i_{m-1,>b,n}\right]$ and $\theta^i_{m,n} = \left[\theta^i_{m,<b,n}, \theta^i_{m-1,\geq b,n}\right]$.

    2. With probability,

$$\alpha = \min\left\{\frac{[p(Y|\theta^*)]^{\phi_n} p(\theta^*)}{[p(Y|\theta^i_{m,n})]^{\phi_n} p(\theta^i_{m,n})}, 1\right\}$$

       Set $\theta^i_{m,b,n} = \theta^*_b$. Otherwise set $\theta^i_{m,b,n} = \theta^i_{m-1,b,n}$.

  end

end

Retain the last step of the Metropolis-Hastings sampler. Set $\theta^i_{b,n} = \theta^i_{M,b,n}$ for all $b \in B_n$.

posteriors. In Section 3 we show that this simple change greatly improves the algorithm's efficiency.

2.4 Theoretical Considerations

We make no theoretical contributions in this paper, so we will not go into detail about the formal arguments proving the strong law of large numbers (SLLN) and central limit theorem (CLT) for the particle approximation in (6) at the conclusion of Algorithm 1. Readers interested in the details of the SLLN and CLT should refer to Chopin (2002), which provides a recursive characterization of the SLLN and CLT that apply after each of the correction, selection, and mutation steps. Herbst and Schorfheide (2014) characterize the high level assumptions sufficient for the SLLN and CLT to apply when the mutation stage is adaptive; that is, when features of the MCMC algorithm depend on previous particle approximations. While difficult to verify in practice, the extension of the SLLN and CLT provides at least a basis for the use of such a transition kernel. Finally, the variances associated with the CLTs have the formulation given in Chopin (2002), but the recursive form is, unfortunately, not useful in practice. Hence, to characterize the uncertainty around our estimates in subsequent sections we use numerical standard errors computed across multiple independent runs of the algorithm, as in Durham and Geweke (2012).

3. Sequential Monte Carlo in Two Controlled Experiments

Since we know of no other research evaluating SMC's effectiveness in applications comparable to ours, we find it worthwhile to verify that SMC can reliably estimate the posteriors of VAR-sized models. We verify our SMC algorithm's effectiveness by demonstrating the accuracy of its MDD estimates for two models for which we know the true MDD in closed-form: 1) a VAR with conjugate prior and, as a more challenging test, 2) a mixture of VAR posteriors.

3.1 The Constant-Parameter VAR

The MS-VARs we estimate in subsequent sections build off of a parameterization of the VAR model known as the "structural" form. The structural VAR

has the form

$$y_t'A = \sum_{l=1}^{p} y_{t-l}'F_l + F_0 + \varepsilon_t', \quad \varepsilon_t \sim \mathcal{N}(0, I_n), \quad \text{for } 1 \le t \le T \tag{10}$$

where y_t is an $n \times 1$ vector of observables at time t, y_{t-l} is the time $t-l$ realization of the same observables, p is the number of lags of the observables, and ε_t is an $n \times 1$ vector of structural shocks. Letting $x_t = [y_{t-1}', \ldots, y_{t-p}', 1]'$ and $F = [F_1', \ldots, F_p', F_0']'$, we can write the VAR more compactly as

$$y_t'A = x_t'F + \varepsilon_t', \quad \varepsilon_t \sim \mathcal{N}(0, I), \tag{11}$$

and refer to its parameters as $\theta_S = (A, F)$.

Standard priors for θ_S described in Sims and Zha (1998) and Waggoner and Zha (2003a) do not admit a closed form expression for the model's MDD, which would make it difficult to assess our algorithm. However, in the absence of overidentifying restrictions on the matrices A and F, one can derive a prior for θ_S as a change of variables from a prior over an alternative ("reduced-form") parameterization, with parameters θ_{RF}. The reduced-form parameters are defined as

$$\theta_{RF} = (\Sigma, \Phi) = g(\theta_S) = ((AA')^{-1}, FA^{-1}), \tag{12}$$

in which case the VAR is written as

$$y_t' = x_t'\Phi + u_t', \quad u_t \sim \mathcal{N}(0, \Sigma), \tag{13}$$

Standard priors $p_{RF}(\cdot)$ admit a closed-form expression for the VAR's MDD. Using this distribution and change of variables, the resulting prior density for θ_S is simply

$$p_{RFB}(\theta_S) = p_{RF}(g(\theta_S))|J(g(\theta_S), \theta_S)|, \tag{14}$$

where $J(g(\theta_S), \theta_S)$ is the Jacobian of g. We use RFB to indicate that prior for

the structural parameters θ_S is based on this change of variables from a standard reduced-form prior. The details of p_{RFB} are described in subsequent sections, but for our present purposes it suffices to know that 1) we can easily sample from it, 2) we have closed form expressions for p_{RF} and J that admit pointwise evaluation, and 3) it gives us an exact expression for the model's MDD. In other words, we can estimate the model with SMC and and compare SMC's MDD estimate to the true MDD. To sample from $p(\theta_S)$ we first sample from $p(\theta_{RF})$ and then transform θ_{RF} into θ_S via the function g^{-1}, which we define as

$$g^{-1}(\theta_{RF}) = ((chol(\Sigma)')^{-1}, \Phi(chol(\Sigma)')^{-1}) = (A, F) = \theta_S \,, \tag{15}$$

where $chol(\cdot)$ refers to the lower triangular Cholesky matrix. [8]

3.2 Experiment 1: SMC Accuracy for VAR Posteriors

We first test SMC's performance on a VAR with $n = 3$ variables and $p = 3$ lags. The data for our test consists of observations on the output gap, inflation (GDP deflater), and the Federal Funds Rate from 1959:Q1 to 2005:Q4. We use the exact dataset from the empirical example of Sims et al. (2008), which we also use when estimating Markov-switching models in Section 4.

Recall that the SMC sampler described in Section 2 features a number of tuning parameters that must be set by the user. For our baseline experiment, we set $N_{part} = 2000$, $N_\phi = 500$, $M = 1$, $N_{blocks} = 3$ (random), and $\lambda = 4$.[9] We run 20 Monte Carlo replications of the sampler and examine the distribution of ln(MDD) estimates. The first row of Table I shows the results under the baseline

[8] It is well known in the literature that our choice of g^{-1} is not the unique definition for which $g(g^{-1}(\theta_{RF})) = \theta_{RF}$: multiplying both A and F from the right by an orthogonal matrix would yield an alternative $\tilde{\theta}_S$ for which $g(\tilde{\theta}_S)$ would equal the same θ_{RF}. In the present context our definition of g^{-1} is just a normalization and its lack of uniqueness is irrelevant. Both the VAR likelihood and our choice of prior are invariant to orthogonal rotations, so the model's MDD is invariant to alternative choices of g^{-1}. Rubio-Ramírez, Waggoner, and Zha (2010) document that the density of any prior for θ_S derived from the θ_{RF} parameterization will be invariant to orthogonal rotation.

[9] Appendix A provides a thorough examination of SMC's effectiveness under a variety of values for the tuning parameters. In particular, Table A-1 represents a matrix of approximate "partial derivatives" of MDD estimation accuracy with respect to various tuning parameters.

TABLE I
ACCURACY OF SMC ESTIMATES OF ln(MDD): EFFECT OF RW-MH PROPOSAL COVARIANCE MATRIX

Σ_{prop}	RMSE
Conditional	0.21
Unconditional	1.90

Notes: The other SMC tuning parameters for this exercise are $N_{part} = 2000$, $N_{blocks} = 3$, Random Blocking, $N_\phi = 500$, $M = 1$, and $\lambda = 4$. The VAR model has $p = 3$ lags. RMSE is the root mean squared error of the estimates of of ln(MDD). The true value of the ln(MDD) is 1791.9.

tuning parameters in terms of root mean squared error (RMSE) of ln(MDD) estimates. We can see that the sampler is quite accurate.

The second row of Table I demonstrates the ramifications of using (9) as the RWMH proposal variance rather the conditional approximation, given by (8). The sampler using (9) most closely resembles the one used for DSGE models by Herbst and Schorfheide (2014). Using the unconditional variance estimate in the block RWMH leads to substantial deterioration in the sampler's performance, as the RMSE of the ln(MDD) estimates increases by nearly an order of magnitude. To contextualize the efficiency gains from our modification of the Herbst and Schorfheide (2014) proposal variance, we find that the gains in accuracy from using the conditional approximation are significantly greater than the gains from doubling the number of particles (or even moving from 1000 to 5000 particles). One reason for this is that the VAR prior exhibits substantial correlation among key parameters. When this correlation structure is not accounted for, the sampler performs very poorly in the early stages when the prior dominates the likelihood contribution.

3.3 Experiment 2: SMC Accuracy for a Mixture of VAR Posteriors

Sims et al. (2008) stress that the posterior of MS-VARs "tends to be non-Gaussian" and may well contain "multiple peaks." Indeed, when estimating MS-VARs in Section 4 we find evidence of fat-tailed and multipeaked posterior densities in our posterior draws, even after normalizing them. To determine

whether or not SMC can stand up to such irregularities, we conduct a Monte Carlo simulation on a bimodal target density for which: 1) we know the integrating constant in closed-form, which provides an absolute measure of success, 2) we can sample the target distribution directly and then apply existing MDD estimation techniques, which provides a relative measure of success, and 3) the distribution is similar to the SMC-estimated posterior of the MS-VARs we consider in Section 4, which provides our controlled experiment with empirical relevance.

We construct the bimodal target distribution as the mixture of two posteriors for a parameter vector θ that share a common prior, but are informed by different observations. Letting $p(\theta)$ be a prior, $p(Y|\theta)$ the model's likelihood function for observations Y, and

$$p(Y) = \int_{\Theta} p(\theta)p(Y|\theta)d\theta \,, \tag{16}$$

our target density is given by

$$\tilde{p}(\theta|Y_1, Y_2) = \alpha \left(\frac{p(\theta)p(Y_1|\theta)}{p(Y_1)} \right) + (1-\alpha) \left(\frac{p(\theta)p(Y_2|\theta)}{p(Y_2)} \right) \,, \quad \alpha \in [0,1] \,.$$

We take α as given and known, so we implicitly condition on this value. Defining

$$\tilde{L}(\theta|Y_1, Y_2) = \alpha p(Y_2)p(Y_1|\theta) + (1-\alpha)p(Y_1)p(Y_2|\theta) \,, \tag{17}$$

which we call a pseudo-likelihood, and

$$\tilde{p}(Y_1, Y_2) = p(Y_1)p(Y_2) \,, \tag{18}$$

and performing some simple algebra, we can write the target distribution as

$$\tilde{p}(\theta|Y_1, Y_2) = \frac{p(\theta)\tilde{L}(\theta|Y_1, Y_2)}{\tilde{p}(Y_1, Y_2)} \,. \tag{19}$$

For the mixture components in our experiment we use posteriors of the VAR model described in Section 3.1 and hence we know $p(Y_1)$ and $p(Y_2)$ (and thus $\tilde{p}(Y_1, Y_2)$) in closed-form. We execute 50 replications of SMC estimation of

$\tilde{p}(\theta|Y_1, Y_2)$, which include estimates of the MDD, $\tilde{p}(Y_1, Y_2)$.

To provide a benchmark for SMC, we also sample directly from $\tilde{p}(\theta|Y_1, Y_2)$ and estimate the MDD with standard techniques.[10] In particular we estimate the MDD from the direct sample with two versions of the modified harmonic mean method: the version originally described in Geweke (1989), which we refer to as "MHM," and the version adapted for better performance with non-Gaussian distributions in Sims et al. (2008), which we refer to as "MHM-SWZ."[11]

Although we call this exercise a "benchmark" for SMC, it is in fact a very high bar. The task of posterior sampling is often extremely challenging, to say nothing of MDD estimation from the resulting sample. Researchers typically simulate posterior draws using MCMC algorithms for which iid draws represent a practical upper bound on the quality of the posterior approximation.[12] Thus the benchmark exercise actually sidesteps one of the major challenges of the conventional approach to estimating MDDs. One might then say that our benchmark for SMC is, in fact, an upper bound on the performance of the conventional approach. Meanwhile, we charge SMC with the doubly difficult task of simultaneously sampling the posterior and MDD estimation.

Table II shows the results of our simulation for a VAR($n = 3, p = 5$), from which we arrive at four main conclusions. Firstly, and most central to our interests, SMC estimates the MDD as well as, or better than, either MHM estimator *when we give the MHM estimators 10,000 i.i.d. draws*. From this we conclude that the SMC algorithm shows superior performance in the presence of substantive multimodality. And furthermore these simulation results give us a compelling reason to trust our numerical estimates in Section 4.

Secondly, the MHM-SWZ estimator performs well (compared to traditional MHM) in the presence of bimodality. Even though the MHM-SWZ estimator con-

[10] With known α, we can easily sample $\tilde{p}(\theta|Y_1, Y_2)$ directly. See Appendix D.1 for the direct sampling algorithm.

[11] Frühwirth-Schnatter (2004) documents the poor performance of Chib's estimator for even small mixture models, so we do not consider it here.

[12] In principle, it is possible to use other Monte Carlo methods to obtain estimates of model moments that are even more precise than those achieved with iid draws (i.e., antithetic variates). In practice, these methods are impractical in the environments we study and they are not widely used by econometricians. See Geweke (2005) for details on this class of methods.

TABLE II

ACCURACY OF ln(MDD) ESTIMATES FOR MULTIMODAL TARGET DENSITY WITH DIFFERENT METHODS OF POSTERIOR SAMPLING AND MDD ESTIMATION.

Posterior Sampler	MDD Estimator	RMSE
SMC: $N_{part} = 2000$	SMC	0.421
SMC: $N_{part} = 5000$	SMC	0.337
Direct: 10,000 draws	MHM - SWZ	0.311
	MHM	1.068
Single Mode	MHM - SWZ	0.812
	MHM	0.829

Notes: VAR($n = 3, p = 5$), true $\ln p(Y) = 1725.289$. Values are based on 50 replications. "MHM" refers to the original implementation of the modified harmonic mean estimator from Geweke (1989). "MHM - SWZ" refers to the adaptation of MHM proposed and implemented in Sims et al. (2008). VAR algorithm settings: SMC sampler uses $\lambda = 4$, $n_\phi = 500$, $N_{blocks} = 8$; and MHM estimate uses $p = 0.9$ for truncation.

structs its approximating density around only one of the distribution's modes, the approximating density has fat enough tails to effectively incorporate information throughout the parameter space.

Thirdly, bimodality renders MDD estimation via the MHM method of Geweke (1989) hopeless, as it fails even under large numbers of draws from the target distribution. Since the average bias is in terms of units of $\ln p(Y)$, we can interpret these values as approximately percentage errors of $p(Y)$. Hence, for the VAR simulation, the MHM estimator tends to overstate $p(Y)$ by more than 50% of its true value.

Lastly, the extent to which multimodal target densities pose problems for MCMC methods remains a subject of debate, a debate whose waters we do not care to wade into any more deeply than necessary, but we do wish to document the stakes of proper posterior sampling. The basic concern when using MCMC is that the sampler may not mix properly in a reasonable amount of time. In the worst-case scenario, the MCMC sampler never leaves a neighborhood around the mode nearest to the point from which the algorithm initialized.[13] The rows in

[13]Celeux, Hurn, and Robert (2000) document degeneracies of this nature when posterior sampling with simple MCMC for mixture models. However, Geweke (2007) shows that there

Table II labeled "Single Mode" show MDD estimates computed from a caricature of a failed MCMC algorithm, i.e. the draws are simulated from only one of the two modes. In such a situation the results are disastrous.

4. Sequential Monte Carlo in Practice: MS-VAR Estimation

While conceptually straightforward, just a cursory glance at Sims et al. (2008) reveals that inference for MS-VARs is messy in practice. In this section we revisit the empirical application of Sims et al. (2008) using SMC estimation and two alternative prior specifications. We show that the use of an off-the-shelf prior commonly used in reduced-form VARs significantly improves data fit for MS-VARs and meaningfully alters the posterior probability assigned to models that allow changes to macroeconomic dynamics.

4.1 Structural MS-VAR Model

We estimate MS-VAR models of the form

$$y_t' A(s_t) = x_t' F(s_t) + \varepsilon_t' \Xi(s_t)^{-1} , \quad \varepsilon_t \sim iid\mathcal{N}(0_n, I_n) \tag{20}$$

$$\Xi(s_t) = diag([\xi_1(s_t), \ldots, \xi_n(s_t)]) \tag{21}$$

$$p(s_t | S_{t-1}, Y_{t-1}, \theta, q) = q_{s_t, s_{t-1}} \tag{22}$$

$$q_{s_t=i, s_{t-1}=j} = q_{i,j} , \quad \text{for } t > 0 \tag{23}$$

where $\Xi(s_t)$ is an $n \times n$ diagonal matrix, s_t is the time t realization of a discrete latent process that we call a "state." S_{t-1} is the history of states up to and including $t-1$, and Y_{t-1} is the history of observations up to and including $t-1$.

---

exist MCMC methods, such as the method of Frühwirth-Schnatter (2001), able to handle the *a priori*-known-and-symmetric multimodality of mixture models that results from the arbitrariness of state labeling. Unlike the examples in Celeux et al. (2000) and Geweke (2007), our experience estimating MS-VARs in the subsequent section indicates the presence of asymmetric posterior multimodality, also known as "genuine multimodality," in addition to the typical symmetric multimodality which can, in theory, be normalized away. The target density in the present section possesses only genuine multimodality by construction.

Let H be the total number of states in the latent process and let

$$A = \{A(h)\}_{h\in\{1,\ldots,H\}}, \quad F = \{F(h)\}_{h\in\{1,\ldots,H\}}, \quad \Xi = \{\Xi(h)\}_{h\in\{1,\ldots,H\}} . \tag{24}$$

We then let $\theta = \{A, F, \Xi\}$. Note that we use the set notation in (24) to collect only the unique parameters in each set of matrices; nothing about our framework so far assumes that all parameters of $A(1)$ and $A(2)$, or any other two states, are unique.[14] The state of the latent process at time t may be determined by the joint realization of K independent latent processes, which each govern a different subset of θ. We will refer to the set of parameters corresponding to only process k as θ_k. The notation s_t refers to the joint state of all latent processes, while the notation s_t^k refers to the state of only process k.

The MS-VAR has the likelihood

$$p(Y_T|\theta, q) = \prod_{t=1}^{T} p(y_t|\theta, q, Y_{t-1}) \tag{25}$$

where

$$p(y_t|\theta, q, Y_{t-1}) = \sum_{h=1}^{H} p(y_t|\theta, q, s_t, Y_{t-1})\, p(s_t|\theta, q, Y_{t-1}) . \tag{26}$$

To evaluate (26) note that

$$\begin{aligned} p(y_t|\theta, q, s_t, Y_{t-1}) &= (2\pi)^{n/2} |\det(A(s_t)^{-1'} \Xi(s_t)^{-1} A(s_t)^{-1})|^{-1/2} \\ &\quad \times \exp\left\{ -\frac{1}{2} (y_t' A(s_t) - x_t' F(s_t))\, \Xi(s_t)^2\, (y_t' A(s_t) - x_t' F(s_t)) \right\} . \end{aligned} \tag{27}$$

and one can evaluate $p(s_t|\theta, q, Y_{t-1})$, using the filtering algorithms in Sims et al. (2008).

The probability model for the data described in (20)-(23) belongs to the class of models considered in Sims et al. (2008) and matches the general form of their

[14] For example, one could restrict the regime-switching so that $A(1)$ and $A(2)$ differ by only their last column, which is one specification considered in both Sims and Zha (2006) and Sims et al. (2008).

empirical application. Each MS-VAR we estimate uses the same lag length (5 quarters) and exactly the same data set as Sims et al. (2008), which we also used in Section 3. We also follow Sims et al. (2008) in assuming that $\theta_1 = \{A, F\}$ and $\theta_2 = \{\Xi\}$ follow independent regime-switching processes. Our only point of departure from Sims et al. (2008) is that we do not restrict the parameters multiplying variable i in equation j at each lag l to change only proportionally across regimes. We find that not imposing those restrictions allows the model to achieve superior data fit.

Since $\{A, F\}$ determine the conditional mean of y_t and $\{\Xi\}$ determines the volatility of the structural shocks, we refer to the state of $\{A, F\}$ at time t as s_t^m and the state of Ξ at time t as s_t^v. Denote the number of regimes for $\{A, F\}$ as H_m and the number of regimes for Ξ as H_v. If a model has $H_m = 2$ and $H_v = 3$, then we refer to it using the shorthand 2m3v. Since we assume that the two processes evolve independently, a 2m3v model has 6 joint states.

We estimate MS-VAR models under a variety of choices for H_m and H_v. For each choice of H, we estimate the model under three different priors for (A, F), to the description of which we now turn.

4.2 Priors for MS-VAR Coefficients

For each MS-VAR the priors on $\{(A(h_m), F(h_m))\}_{h_m=1}^{H_m}$ and $\{\Xi(h_v)\}_{h_v=1}^{H_v}$ are independent and identical across h_m and h_v respectively.

1. SZ Prior. This is the prior originally described in Sims and Zha (1998) and used in Sims et al. (2008). For each state h_m, the prior takes the form

$$a(h_m) \sim \mathcal{N}(0, I_n \otimes H_0) \tag{28}$$

$$f(h_m)|a(h_m) \sim \mathcal{N}(vec(\bar{S}A(h_m)), I_n \otimes H_+) \,, \tag{29}$$

where

$$a(h_m) = vec(A(h_m)) \,, \quad f(h_m) = vec(F(h_m)) \,, \quad \bar{S} = \begin{bmatrix} I_n \\ 0_{(n(p-1)+1)\times n} \end{bmatrix} \tag{30}$$

and H_0, H_+ are prior parameters.[15] In practice, the prior is implemented with dummy observations as described in Sims and Zha (1998). The dummy observations depend on a few moments constructed from the data, $\bar{y}$ and $\bar{s}$, and vector of hyperparameters that control the influence of different subsets of the dummy observations. The standard implementation sets $\bar{y}$ as the mean of the observations used to initialize the lags of the VAR and $\bar{s}$ as the standard deviations of the residuals from univariate autoregressions for each data series, both of which we follow here.[16] For this prior we set Λ identically to Sims et al. (2008) at $\lambda_0 = 1.0$, $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.2$, $\lambda_4 = 0.1$, $\mu_5 = 1.0$, and $\mu_6 = 1.0$ and we refer to this set of values as Λ_{SWZ}.

2. Reduced-Form-Based (RFB) Prior. We take up the suggestion of Sims and Zha (1998) and derive a prior for (A, F) by placing a prior distribution over the reduced-form dynamics, summarized by Φ and Σ, then mapping to $(A(h_m), F(h_m))$ via (15) in Section 3.[17] Appendix B.2 gives the exact expressions for the density of the RFB prior.

Discussion. We pause here to describe the key features of, and relationship between, the SZ and RFB priors. A key aspect of the SZ prior is its centering of A at 0. Sims and Zha (1998) note that the prior for A in (28) is equivalent to what one would derive from inverse-Wishart beliefs about $\Sigma = (AA')^{-1}$ with $n+1$ degrees of freedom and while also ignoring the Jacobian term for the transformation from $A \to \Sigma$. With appropriate choices of hyperparameters, the RFB prior for A differs from the SWZ prior only in that it includes the Jacobian, which serves to recenter beliefs about A away from 0. One can see these differences in the prior A densities clearly in Figure 1.[18] Since the VAR's forecast errors have covariances $(AA')^{-1}$, centering beliefs about A at 0 amounts to centering beliefs about the VAR's

[15] As pointed out in Rubio-Ramírez et al. (2010), the SZ prior has the desirable properties of invariance of the density with respect to orthogonal rotations of (A, F).

[16] As has been common since Litterman (1986), we use six lags in the univariate autoregressions from which we estimate $\bar{s}$.

[17] Sims and Zha (1998) state that, "A better procedure, which, however, would not have been very different in practice, would have been to derive our prior on A_0 from a natural prior on Σ^{-1}, the Wishart."

[18] The values in the figure have been "sign normalized" to positive values.

forecast errors at ∞. In Appendix B.2.2 we derive that $p_{RFB}(F|A) = p_{SZ}(F|A)$ and thus the entirety of the difference between the SZ and RFB priors derives from their differences for $p(A)$. Thus the RFB prior differs from the SZ prior with respect to $p(A)$, while maintaining the choices for Λ in Sims et al. (2008) mentioned above.

Though an undesirable property, the SZ prior's mode of $A = 0$ has little effect on the posterior of constant-parameter VARs. Sims and Zha (1998) note that the sample size in typical macro applications is large enough that this will typically be the case: the Jacobian term consists of

$$|J(g(A), A)| = 2^n \prod_{j=1}^{n} a_{jj}^j \,, \tag{31}$$

while the likelihood contains $\prod_{i=1}^{n} a_{ii}^T$, hence ignoring the Jacobian will have little effect on posterior estimates as long as T is "considerably larger than" n. However, effective sample sizes informing regime-specific parameters of MS-VARs may well be small enough that the omitted Jacobian term has a substantial effect on inference. The basic logic is simple (and hardly new): prior beliefs have a larger effect on posteriors when sample sizes are smaller. Thus undesirable features of a prior typically employed for models informed by large sample sizes can distort inference when reemployed for models with "smaller" sample sizes.

3. Reduced-Form-Based Hierarchical (RFB-Hier.) Prior. It is known in the literature that VAR posteriors can be sensitive to the choice of hyperparameters, Λ. However, it is not obvious to us what type of, or how much, shrinkage one should impose in MS-VARs. Should we increase shrinkage to restrict the "size" of the large parameter space inherent in MS-VARs? Or should we decrease the standard types of shrinkage to let the parameters of different, possibly highly transitory, regimes take on values that one might consider unreasonable in constant parameter VARs?

For these reasons we follow the approach of Giannone, Lenza, and Primiceri (2015) and form a hierarchical model in which we put priors over some elements of Λ, as well as $\bar{y}$ and $\bar{s}$, treating them as an additional vector of parameters

to estimate.[19] Following Giannone et al. (2015), we estimate λ_0, μ_5, μ_6, and $\bar{s}$. We also estimate λ_4 and $\bar{y}$ since it seems reasonable to us that the MS-VAR might favor additional flexibility for the constant term (controlled by λ_4 and the average level of variables (controlled by $\bar{y}$).[20] When estimating λ_0, λ_4, μ_5, and μ_6, we give each parameter a prior from the Gamma distribution with a mode at the value used in the SZ and RFB priors and a standard deviation of one. For $\{\bar{y}_j\}_{j=1}^n$, we set identical and independent Normal distributions centered at 0 with a standard deviation of 0.1. For $\{\bar{s}_j\}_{j=1}^n$, we use relatively diffuse independent Inverse Gamma distributions as in Giannone et al. (2015).

The hierarchical approach gives us 10 additional parameters to estimate. However, when using the SMC algorithm, estimating $(\Lambda, \bar{y}, \bar{s})$ does not introduce any additional complications. To be sure, the efficacy of the sampler will diminish slightly because of the increased dimensionality of the parameter space, but estimation proceeds without any modification of the algorithm. Were we to use MCMC methods to estimate the model with the RFB-Hier. prior, we would have to include an additional Metropolis-Hastings step in our sampler, as there are not natural conditionally conjugate relationships for all of hyperparameters. Moreover, estimating these parameters would deteriorate the performance of the algorithm, given the obvious relationships between Λ and (A, F) and the fact that posteriors of some of the hyperparameters–in particular $\bar{y}$–are nonstandard.[21]

Priors on Other Parameters. The priors for all MS-VARs we consider share common specifications for the volatilities and transition regimes. For the volatilities, $p(\xi_j(h_v))$ are independent and identically distributed across j and h_v such that

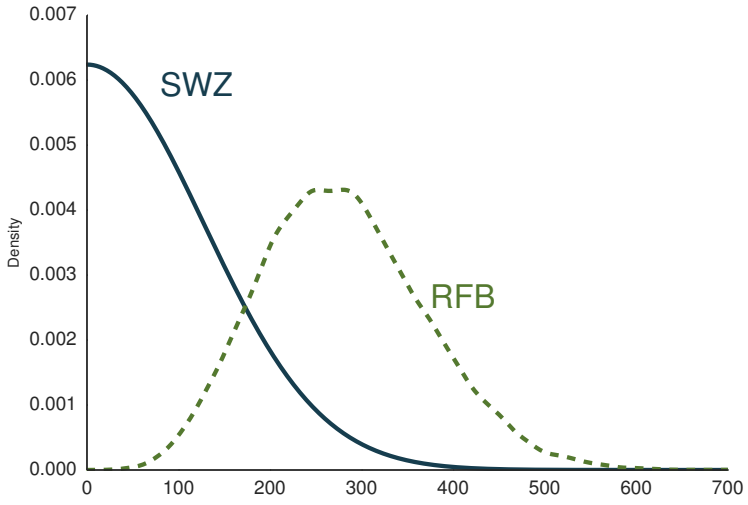
$$\xi_j^2(h_v) \sim \mathcal{G}(\bar{\alpha}_j, \bar{\beta}_j) \tag{32}$$

[19] Sims and Zha (1998) also suggest the potential to form a hierarchical prior by putting prior beliefs over Λ, stating that, "In principle, these hyperparameters can be estimated or integrated out in a hierarchical framework."

[20] We follow Giannone et al. (2015) and fix the other parameters.

[21] See the discussion in Herbst and Schorfheide (2015) on parameter blocking in Metropolis-Hastings algorithms.

FIGURE 1.—Prior Densities for A_{22}



Notes: The figure shows the prior distribution for the parameter A_{22} under the SWZ and RFB priors.

and we set $\bar{\alpha}_j = 1$ and $\bar{\beta}_j = 1$ for all j and k, as in Sims et al. (2008). Additionally, we normalize the first state of volatilities so that $\xi_j(1) = 1$ for all j.

Priors over the transition probabilities q_{ij} for both the mean and shock regimes are of the unrestricted Dirichlet form from Sims et al. (2008). For an n state process i, this distribution is parameterized by n hyperparameters, $\{\alpha_{ij}\}_{j=1}^{n}$ which SWZ suggest eliciting by introspection about the persistence of each regime. For every specification (regardless of the number of regimes), we set

$$\alpha_{i,j} = 5.667, i = j \text{ and } \alpha_{i,j} = 1, i \neq j.$$

For a two state process, this implies an average duration of a given regime of about 6.5 quarters. As the number of states increases, this expected length decreases.

4.3 Estimation Details

Under each prior, we estimate MS-VARs for $H_m = 1, 2$ and $H_v = 1, \ldots, 5$ using the SMC algorithm described in Section 2. We set $N_{part} = 4000$, $N_{blocks} = 12$ (random) and $M = 1$, using the conditional variance given by the normal

approximation for the mutation step. For the tempering schedule we set $\lambda = 4$ and $N_\phi = 2000$.

We have run our SMC sampler in both `Fortran` and `MATLAB`. In `Fortran`, estimation of a given model takes between one and ten minutes, with likelihood evaluations parallelized across the 12 cores of a desktop with an Intel Xeon x5670 CPU. The `Matlab` version executing on the same machine roughly takes between twenty minutes to six hours, depending on the number of states.[22]

For each specification we estimate the model with 50 independent runs of the algorithm and report both the point estimate and standard error of the model's log(MDD). Owing to the difficulty of MCMC estimation of MS-VARs, previous researchers have been able to report standard errors only from different subsets of draws along a single MCMC chain.[23] Since we initialize each run of SMC from an independent draw of initial particles, there is no risk of our standard error estimates being spuriously small because of influential initial conditions. Hence, we interpret the precision of our estimates to, in fact, reflect accuracy.

4.4 Estimation Results: MS-VAR Model Selection

Table III shows the point estimates and associated standard errors of log(MDD) values of MS-VARs, including the constant-parameter VAR (the special case of 1m1v), for each of the three priors. Figure 2 shows the results graphically. From our estimation results we deduce four main findings. Firstly, and consistent with the key findings of Sims and Zha (2006) and Sims et al. (2008), the best fitting model for each prior is a 1m3v or 1m4v model and, furthermore, regime-switching in shock variances is critical to fitting the data. Indeed, the worst fitting

[22] In principle, we could also simulate from the posteriors using the sampler proposed by Sims et al. (2008) for the SWZ prior and modify the Metropolis-within-Gibbs steps of the sampler to accommodate the RFB prior. However, like the other researchers we quoted in Section 1, we have found the MCMC estimation process cumbersome and lengthy. Experimentation across models indicated difficulties with reliably finding the posterior mode, making the batch estimation exercise tedious. On a subset of models with the SWZ prior, which we successfully repeatedly sampled using MCMC, the SMC and MCMC posteriors more-or-less coincided. The SMC posteriors were slightly wider than the MCMC ones, which generally indicates a more thorough posterior exploration.

[23] For MCMC estimation, Gelman and Rubin (1992) emphasize the importance of using multiple independent chains, with each chain initialized from a different starting value.

TABLE III
SMC ESTIMATES OF ln(MDD) FOR MS-VAR MODELS.

Model		Prior					
		SWZ		RFB		RFB-Hier.	
		ln(MDD)	(S.E.)	ln(MDD)	(S.E.)	ln(MDD)	(S.E.)
1m	1v	1759.10	(0.07)	1754.77	(0.08)	1778.15	(0.78)
1m	2v	1869.51	(0.09)	1873.24	(0.13)	1877.93	(0.71)
1m	3v	1872.64	(0.11)	1877.83	(0.18)	1880.92	(0.81)
1m	4v	1872.57	(0.12)	1879.17	(0.14)	1880.07	(1.07)
1m	5v	1871.27	(0.16)	1878.03	(0.15)	1878.82	(1.29)
2m	1v	1845.23	(1.45)	1836.78	(2.84)	1857.68	(2.73)
2m	2v	1867.48	(0.33)	1873.70	(0.55)	1879.94	(0.74)
2m	3v	1869.98	(0.45)	1877.34	(0.55)	1880.32	(0.93)
2m	4v	1869.55	(0.27)	1878.22	(0.47)	1879.58	(1.17)
2m	5v	1868.26	(0.45)	1876.83	(0.43)	1877.49	(1.55)

Notes: ln(MDD) estimates are means from 50 independent runs of the algorithm for each model. We give standard errors of the log(MDD) estimates, computed across the 50 runs, in parentheses. The SMC algorithm hyperparameters are $N_{part} = 4000$, $\lambda = 4$, $N_{blocks} = 12$, $N_{\phi} = 2000$, and $M = 1$.

regime-switching specification is always the 2m1v model by a large margin.

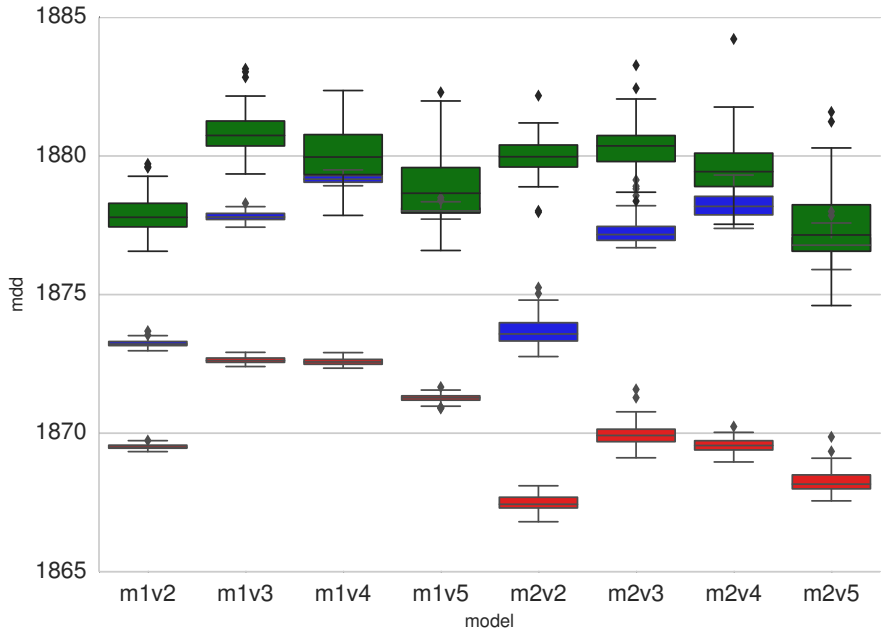
Secondly, changing the prior from SWZ to RFB to RFB-Hier increases the MDD of all regime-switching models (other than 2m1v). We take this to mean that the two variants of the RFB prior are first and foremost favored by the data rather than any particular specification. The MDD improvements are large for any particular specification, in many cases exceeding 10 log points.[24]

Thirdly, in addition to improving data fit for all models, changing the prior from SWZ to RFB to RFB-Hier dramatically increase the posterior probability of a 2m specification being the correct model.[25] Table IV shows the posterior probability on 2m specifications conditional on each prior, along with the unconditional posterior probability on all specifications for each prior. Under the SWZ prior,

[24]Note that, with equal prior odds on two models, a 10 point difference in their log(MDD)s puts the posterior odds in favor of the better fitting model above 20,000 to 1.

[25]In calculations of posterior probabilities, we assume all models are *a priori* equiprobable.

FIGURE 2.—log(MDD) estimates for each MS-VAR specification and prior.



Notes: The figure shows box plots for log MDD of each specification and prior, computed from 50 independent runs of the SMC algorithm for each specification-prior combination. We omit the 1m1v and 2m1v because they are the worst fitting models by wide margins.

there is negligible probability (0.06) on changes in the mean parameters; that is, the 2m models. Under the RFB prior, in which we have only taken into account the Jacobian term and the IW degrees of freedom, the probability increases to 0.29. Under the RFB-Hier. the probability increases further to 0.43, nearly a coin-flip with the only-variances-change explanation. This finding contrasts with the results in Sims and Zha (2006) and Sims et al. (2008) who find a landslide victory (10 log points in Sims et al. (2008)) for the only-variances-change specification. Recall that a key difference between our model and the models in Sims and Zha (2006) and Sims et al. (2008) is that we do not impose the additional restriction of only proportional switching across the coefficients multiplying variable i in equation j. Sims et al. (2008) express the concern that allowing all parameters to

TABLE IV
POSTERIOR PROBABILITY OF 2M MODEL CONDITIONAL ON PRIOR.

Prior	$P(2m\|Prior,Y)$	$P(Prior\|Y)$
SWZ	0.06	0.00
RFB	0.29	0.12
RFB - Hierarchical	0.43	0.88

Notes: The second column gives the posterior probability of the 2m models, conditional on the prior. The third column gives the posterior probability on all models estimated with a particular prior.

change would over-parameterize the model and such models would be heavily penalized for their complexity in the MDD calculation. Our results show that these fears are unwarranted.[26]

Fourth, across all priors and model specifications, Markov-switching parameters offer large gains in model fit compared to constant-parameter specifications, as was also found in Sims and Zha (2006). For all models with at least 2 volatility regimes, the MDD gains exceed a staggering 100 log points.

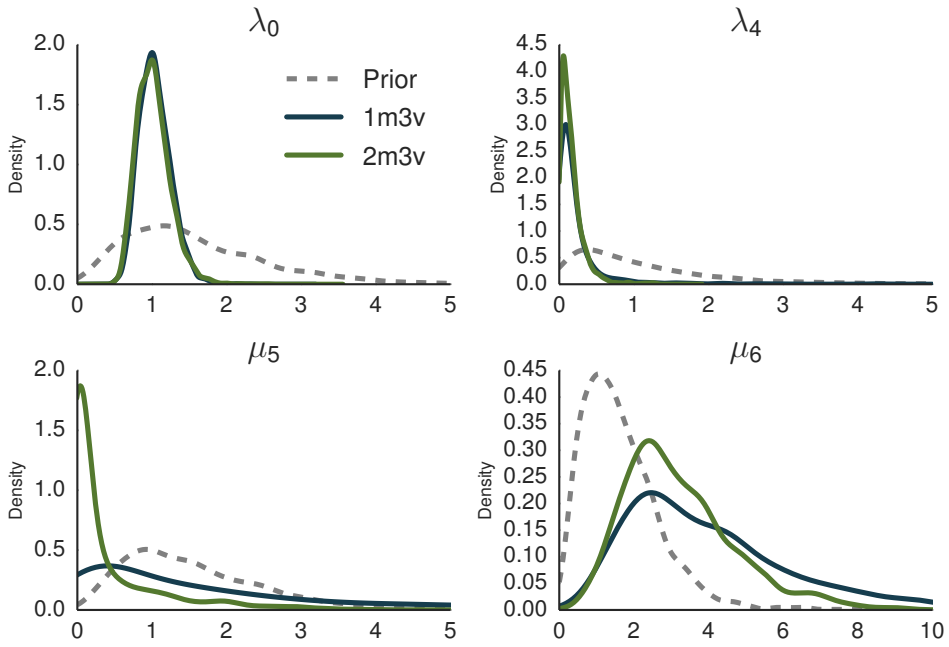
4.5 Estimation Results: Examining the 2m3v Model

4.5.1 Hyperparameter Posteriors

Figure 3 shows the priors and posteriors for the estimated elements of Λ under both 1m3v and 2m3v specifications and Appendix E contains tables summarizing the posteriors of all estimated hyperparameters, for all specifications. The most important feature of the estimated hyperparameter posteriors is the difference of the μ_5 posteriors (lower left panel of Figure 3) under the 1m3v and 2m3v specifications. The hyperparameter μ_5 controls shrinkage towards the "sums-of-coefficients" dummy observations. In particular, the 2m3v model wants virtually no influence for these observations. This represents a dimension in which the model strongly favors weaker prior restrictions to make the best use of the

[26] Some researchers also find the proportionality restrictions undesirable on theoretical grounds. As is well-known, and was pointed out particularly starkly in Benati and Surico (2009), one would expect all coefficients of the VAR representation of a DSGE model to change if one changes the DSGE model's policy rule parameters.

FIGURE 3.—Posterior of Λ



infrequently occurring additional conditional mean regime.

4.5.2 Conditional Mean Regimes

Rather than describe the time-series of regime probabilities conditional on an estimate of the model's posterior mode, Figure 4 shows the time-series of regime-probabilities for each of the 4000 particles from a single run of the SMC algorithm for the 2m3v model. In particular, the figure shows the time-series data used in estimation (scale on left axis) together with the posterior probabilities of estimated regimes (conditional means in top panel, shock variances in bottom panel). For example, if we look at the year 1990 in the top panel and see that the figure's background is uniformly white from top to bottom, then that means that virtually all 4000 particles are in agreement about the regime probability at that date.

From the top panel of Figure 4 one can see a substantial amount of disagreement across particles about the timing of regime occurrences. In particular, the posterior contains two modes which support alternative interpretations of the

$h_m = 1$ regime. To make the bimodality more visually apparent, we sorted the particles in Figure 4 in ascending order from the top according to the average probability of $h_m = 2$ over the 2 years of observations from 1965:Q1 to 1966:Q4. Near the top of the figure, one can see a set of particles favoring an $h_m = 1$ occurrence in 1965-1966, some of which favor a recurrence in the late 1990s. These same particles put less probability on $h_m = 1$ in the early 1980s than do the particles near the bottom of the panel. Documenting this relationship formally, there is a negative correlation of -0.34 between the average probability of $h_m = 1$ over 1965-1966 and the average probability of $h_m = 1$ over 1980-1981, thus revealing a substantial amount of multimodality.[27]

The key macroeconomic feature of the first mode is periods of rapid economic growth with little inflation and thus little movement in the nominal interest rate. One might interpret the parameter value corresponding to this mode as representing periods of a particularly flat Phillips Curve. Other time-series investigations have uncovered nonlinearities and/or time-variation in the Phillips curve that mesh well with this type of time-variation in economic dynamics. Stock and Watson (2010) and many references therein document a nonlinear relationship between the traditional gap measures and inflation, wherein the strongest Phillips curve relationship occurs in recessions. Their finding is roughly consistent with the interpretation of the parameter values generating this mode: that the relationship between inflation and economic slack deteriorates during (some) periods of quickly diminishing slack.

One can also examine how a fixed coefficient structural general equilibrium model fits the economic dynamics during the $h_m = 1$ period. That these periods might represent a structural change economic dynamics is, in a sense, visible

[27] When we describe the features of a particular regime's posterior, there is an issue about which of a given particle's parameter values represent which regime. In the statistics literature on mixture models, this is a well-known annoyance referred to as the "label switching problem." We refer readers interested in the issue to Jasra, Holmes, and Stephens (2005)'s excellent description and survey of solutions. Since one can always relabel regimes arbitrarily, a model with 2 regimes necessarily has 2 symmetric modes. The statistics literature on mixture models uses the term "genuine multimodality" to refer to multimodality in the posterior that exists even after normalizing draws around one of the inherently-symmetric modes. Appendix C contains the details on our handling of normalization and relabeling for the MS-VAR.

from the historical decompositions implied by the NK-DSGE model of Smets and Wouters (2007). The Smets and Wouters (2007) model interprets the second half of the 1990s as a period in which the joint dynamics of output growth and inflation are caused by a sequence of similarly sized negative "mark-up" shocks occurring for more than 5 years in a row. The "mark up" shocks in the Smets and Wouters (2007) model function largely as a time-varying slope to the Phillips Curve. The persistence of necessary mark up shocks suggests a dimension of model misspecification.

The key macroeconomic feature of the second mode (the particles whose time-series of probabilities are nearer the bottom of Figure 4's top panel) is an increased responsiveness of the nominal interest rate to changes in inflation. To document this formally we calculate the impulse response of the nominal interest rate to a one standard deviation sized inflation shock (the second shock in the structural system) on impact under each regime, conditional on the posterior draw belonging to the region around the second mode. Figure 5 shows density estimates of these two responses under each regime. While the IRFs under the second regime are not as sharply identified as those of the first, the greater probability of a larger response under regime 1 is still apparent in the figure.

4.5.3 Shock Variance Regimes

Turning to the shock volatility regimes, shown in the second panel of Figure 4, one can see that a single regime prevailed from the mid-1980s to the end of the sample and that same regime occurs in the late 1960s. For most posterior draws this regime has the lowest variance for all three structural shocks.

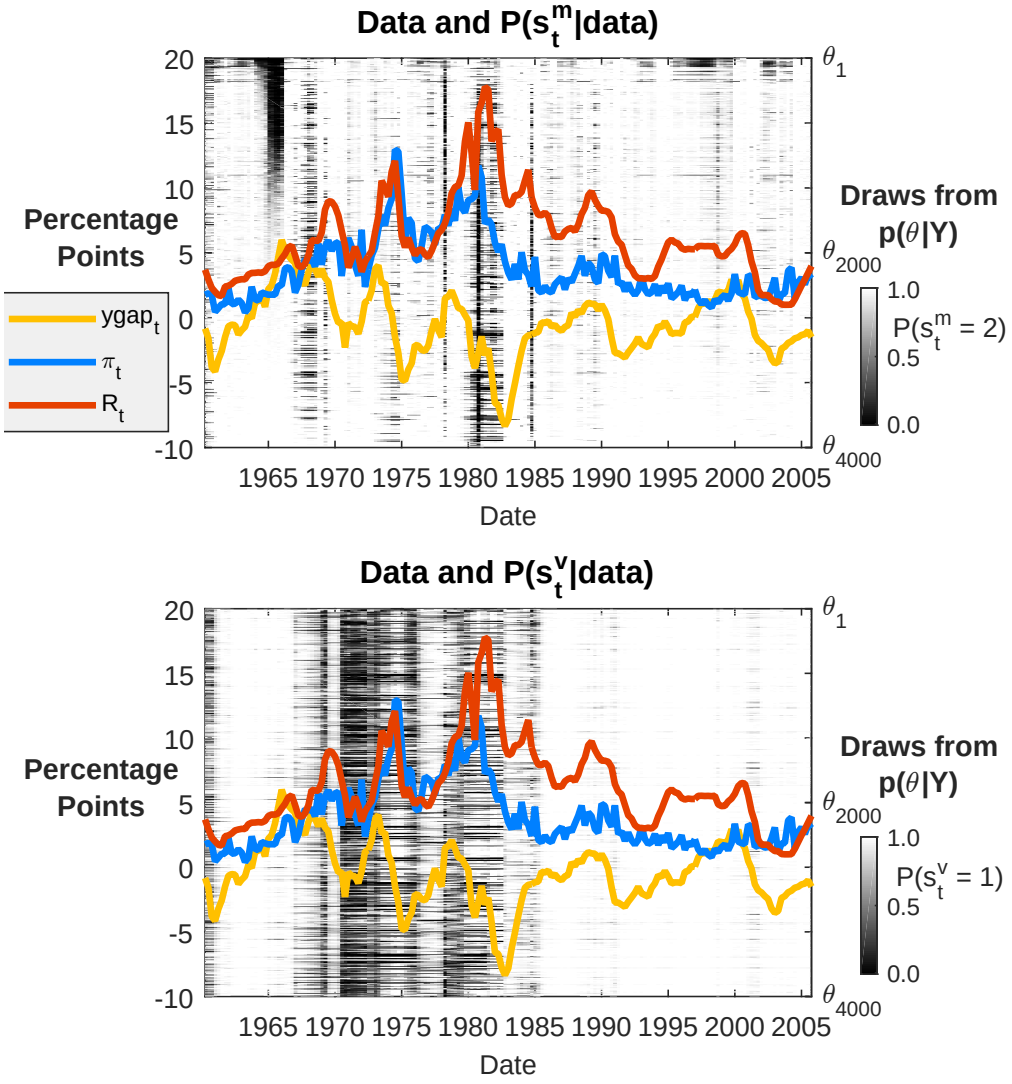
Not surprisingly, the regime with the largest shock standard deviations occurs during the mid-1970s and early 1980s, similar to the $1m$ models. Echoing the main result of Sims and Zha (2006) and Sims et al. (2008), our model interprets the Great Moderation as a once-and-for-all decrease in shock volatilities, in line with a "good luck" explanation. The same "good luck" regime also prevailed during the late 1960s.

5. Conclusion

Led by Sims and Zha (2006) and Sims et al. (2008), MS-VARs have played a prominent role the debate over whether or not any structural change to US macroeconomic dynamics has occurred in the last 50 years. In this paper we have shown that some small tweaks to recently-developed SMC algorithms allows us to apply them to MS-VAR estimation. SMC delivers fast, reliable characterization of posteriors and dramatically broadens the space of tractable priors. We use the ease of SMC implementation under alternative priors to show that, relative to the conclusions of Sims et al. (2008), the use of an off-the-shelf prior typically applied to reduced-form VARs improves data fit and substantially alters posterior beliefs about changes to economic dynamics. When using the hierarchical reduced-form-based prior we find a 43% chance that the true model features changing macroeconomic dynamics either in the form of a periodically flattening Phillips Curve or increased responsiveness of the monetary authority to inflation shocks.

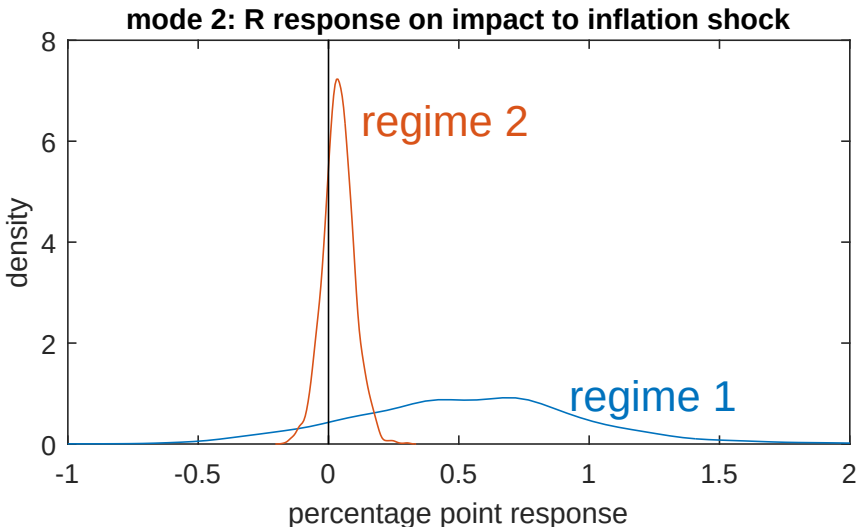
The results in our paper suggest that the choice of priors deserves careful attention when working with densely-parameterized models, such as MS-VARs. It may well be the case that appropriate priors for such models require us to depart from previous methods that were chosen for either analytical or computational tractability. Whether or not such departures are necessary is an empirical question, but this paper shows that it is a question whose answer will most likely be found by using SMC methods.

FIGURE 4.—Observables and Regime Probabilities



Notes: The figure shows the time-series data used in estimation (left axis) together with the posterior probabilities for the conditional mean regimes (top panel) and volatility regimes (bottom panel), for each of 4000 particles from a single run of SMC for the 2m3v model. Top panel particles are sorted according to probability $h_m = 1$ averaged over 1965-1966.

FIGURE 5.—Density Estimates for IRF of R_t to ϵ_t^{π}.



Notes: Figure shows the density estimates for the impact response of the interest rate to a one standard deviation inflation shock for each regime, conditional on being on the second mode (described in text).

A. Additional Computational Results

A.1 Assessing the Importance of Tuning Parameters

The researcher applying SMC faces a few key questions about SMC's use in practice. How should one choose n_ϕ (or λ)? How many particles should one use? How many parameter blocks? While theoretical results on the optimal choice of the algorithm's tuning parameters are beyond the scope of this paper, in this section we exploit the relative transparency of VARs to move beyond the suggestions of Herbst and Schorfheide (2014) and find well-performing choices for tuning parameters.

To assess the importance of each of the algorithm parameters, we vary each component while holding the rest of the hyperparameters at the baseline case. This gives a rough "partial derivative" of each parameter's contribution to the effectiveness of the algorithm. In particular, we 1) consider the use of the proposal distribution for the mutation steps as described in Herbst and Schorfheide (2014), 2) vary the number of particles to 1000 and 5000, 3) vary the number of blocks and the mechanism for selecting them, 4) assess the trade-off between the number of bridge distributions and intermediate Metropolis-Hastings steps while keeping the number of likelihood evaluations fixed by setting $(N_\phi, M) = (50, 10)$, and 5) vary the ϕ schedule by testing $\lambda = 1, 7$. We run 20 Monte Carlo replications of the sampler for each configuration of hyperparameters and examine the distribution of the estimates of $\ln p(Y)$. Table A-1 shows the results of our Monte Carlo exercise. Each row after the first describes a deviation from the baseline tuning parameters and shows the estimation performance of the algorithm under that parameterization.

Under the baseline setting, the sampler using the structural parameterization is slightly more accurate. The primary reason for this is that the RWMH is restricted to draws which satisfy a positive definiteness condition for Σ. When a draw does not have this property, it is rejected, reducing the efficiency of the MH algorithm and hence the size of movements in the parameter space. The structural parameterization operates on the Cholesky decomposition of Σ thus negating the problem of drawing inadmissable parameterizations and allowing

TABLE A-1
SMC ESTIMATES OF $\ln p(Y)$ FOR VAR: EFFECTS OF ALGORITHM TUNING PARAMETERS

SMC Tuning Parameters							VAR Parameterization	
							Reduced Form	Structural
Σ_{prop}	N_{part}	N_{blocks}	Blocking	N_{ϕ}	M	λ	RMSE	RMSE
Cond	2000	3	Random	500	1	4	0.29	0.21
Un	-	-	-	-	-	-	1.37	1.90
-	1000	-	-	-	-	-	0.39	0.47
-	5000	-	-	-	-	-	0.19	0.11
-	-	1	-	-	-	-	0.61	0.50
-	-	2	-	-	-	-	0.39	0.38
-	-	2	(Φ, Σ)	-	-	-	0.44	3.95
-	-	3	Row	-	-	-	0.26	0.75
-	-	4	(Row, Σ)	-	-	-	0.18	1.51
-	-	-	-	50	10	-	0.43	1.33
-	-	-	-	-	-	1	1.87	4.02
-	-	-	-	-	-	7	0.41	0.37

Notes: The symbol " - " indicates inheritance of the parameter value from the baseline parameterization given in the first line of the table. RMSE is the root mean squared error of the estimates of $\ln p(Y)$. VAR has 3 variables, 3 lags, and a constant term. The true value is of the ln(MDD) is 1791.9.

for more effective moves.[28]

The first set of deviations we consider, line two in Table A-1, shows what happens when (9) is used as the RWMH proposal variance rather the conditional approximation, given by (8). This variation of the sampler most closely resembles the one used for DSGE models by Herbst and Schorfheide (2014). Using the unconditional variance estimate in the block RWMH leads to substantial deterioration in performance of the sampler. While the average log marginal data density still reliably estimates the true value, the standard deviation of the log

[28]Since our identification scheme is the Cholesky decomposition, negative elements along the diagonal should technically have zero density. However our prior density does not actually rule out these values and thus treats the sign of a column of A and F as simply a normalization.

MDD estimate across the twenty simulations has increased markedly: relative to the baseline algorithm the RMSE is about five times larger for the reduced-form parameterization and almost ten times larger for the structural. One reason for this is that the VAR prior exhibits substantial correlation among key parameters. When this is not accounted for, the sampler performs very poorly in the key early stages when the prior dominates the likelihood contribution. To contextualize the efficiency gains from our modification of the Herbst and Schorfheide (2014) proposal variance, we find that the gains in accuracy from using the conditional approximation are significantly greater than the gains from doubling the number of particles (or even moving from 1000 to 5000 particles).

The second set of deviations we consider, rows 3 and 4 of Table A-1, shows the effects of changing the quantity of particles. As one would expect, RMSEs fall as the number of particles increases, roughly in line with the central limit theorems in the previously mentioned literature.

The third set of deviations we consider, rows 5 through 9 of Table A-1, examines the role of the blocking configurations of the parameter vector during the mutation phase. First, we consider using a single block for all parameters and we can see that failing to break the parameters into smaller blocks yields RMSEs twice as large as our baseline configuration. Second, we allow for two blocks instead of the baseline number, three. These two blocks are chosen either randomly or by dividing the parameter vector in a "natural way," with one block for Φ and another for Σ. We also allow for a three block fixed scheme where the parameters are grouped by the row in which they enter this VAR. For the samplers using the reduced form parameterization, the effects of blocking is generally smaller. Reducing the number of blocks to 2, but maintaining the random assignment of parameters into blocks each stage, results in an increase in the RMSE to 0.39, relative to the baseline of 0.29, which has three blocks. Removing the randomization every stage and partitioning the parameter in the "natural way": $[\Phi, \Sigma]$, results in a modest increase in the RMSE. For the sampler using the structural parameterization, the quality of the marginal data density estimate deteriorates much more when using a fixed block scheme. Under the natural partitioning of θ into Φ and Σ, the RMSE of the log marginal data density

is 3.95, more then ten times the size when randomizing the blocks.

The fourth type of deviation we consider concerns the number of ϕ stages and mutation steps. Row 10 of Table A-1 shows the results when the number of stages N_ϕ is reduced to 50 but the number of intermediate MH steps is increased to 10, thus keeping the total number of likelihood evaluations the same as under the baseline configuration. We see that performance, measured in terms of RMSE is, deteriorates under this setting relative to the baseline. In the case of structural parameterization, the increase in RMSE is substantial. One reason for this is that the drop in the number of intermediate stages causes the "difference" between two subsequent distributions to increase substantially, in a way that the increased MH steps cannot compensate for. Another reason is that even though the blocks are randomized at each stage, the blocks are fixed *within* the sequence of mutation MH steps at a given stage, so that even a few "bad" configurations of blocks can deteriorate performance despite a large number of MH steps.

Finally, the fifth set of deviations we consider, the bottom two rows of Table A-1, shed light on the role of the ϕ schedule. When $\lambda = 1$, the schedule is linear, resulting in information being added too quickly. Only a few particles have meaningful weight as we move from the prior to the early stages of the schedule. This means that many particles at the end of the algorithm share a common ancestor, and this dependence manifests itself in poor estimates. Indeed, this configuration is the only one exhibiting meaningful bias. Moreover, the RMSE of the log marginal data density estimate under the structual parameterization is 4.02 more than twice that of the reduced form estimate, suggesting that the discrepancy between the prior and posterior is worse under the structural parameterization. Adding information "too" slowly does not incur the same penalty, as the results when $\lambda = 7$, show. While the RMSEs of 0.41 and 0.37 are slightly higher than under the baseline case, because of the relatively large differences in the distributions later in the sampler, the mean error is still quite small. One reason for this is that the shape of the posterior is largely determined when ϕ is quite small, so even large differences between ϕ later in the schedule don't result in radically different distributions.

Overall, the SMC algorithm works well across a wide range of values for the

hyperparameters under both the reduced form and structural parameterizations of the VAR.

B. VAR Priors

B.1 Conjugate Reduced-Form VAR Prior and MDD Expression

The standard conjugate prior for the parameters (Σ, Φ) of a reduced-form VAR specifies Inverse-Wishart beliefs about Σ and Gaussian beliefs about $vec(\Phi)|\Sigma$.

$$\Sigma \sim \mathcal{IW}(\Psi, \nu) \tag{33}$$

$$vec(\Phi)|\Sigma = \mathcal{N}(vec(\Phi^*), \Sigma \otimes \Omega^{-1}) \tag{34}$$

where Ψ, ν, Φ^*, and Ω prior hyperparameters specified by the econometrician. In practice, researchers typically implement VAR priors by supplementing the data matrices Y and X with dummy observations Y^* and X^*. The resulting posterior for Σ and Φ is identical under either approach as long as

$$\Omega = X^{*'}X^* \tag{35}$$

$$\Phi^* = \left(X^{*'}X^*\right)^{-1} X^{*'}Y^* \tag{36}$$

$$\Psi = \left(Y^{*'}Y^*\right) - \left(\Phi^{*'}\Omega\Phi^*\right) \tag{37}$$

$$d = T^* - m\,, \tag{38}$$

with T^* and m the number of rows and columns of X^* respectively.

Given the data and choices of prior hyperparameters and defining

$$\tilde{S} = (Y'Y + \Phi^{*'}\Omega\Phi^*) - (X'Y + \Omega\Phi^*)'(X'X + \Omega)^{-1}(X'Y + \Omega\Phi^*)\,, \tag{39}$$

the MDD of the VAR is given in closed form by the expression

$$p(Y) = (2\pi)^{-Tn/2}\left(\frac{|(X'X+\Omega)|^{-n/2}}{|\Omega|^{-n/2}}\right)\left(\frac{|\tilde{S}+\Psi|^{-(T+\nu)/2}}{|\Psi|^{-\nu/2}}\right) \times\left(\frac{2^{(T+\nu)n/2}}{2^{\nu n/2}}\right)\left(\frac{\Gamma_n((T+\nu)/2)}{\Gamma_n(\nu/2)}\right). \tag{40}$$

B.2 Reduced-Form-Based Prior

B.2.1 Prior Density for A

Our reduced-form-based prior for A is derived from the relationship $(AA')^{-1} = \Sigma$. As is standard in the analysis of reduced-form VARs, we give Σ a density of the inverse-Wishart family $\mathcal{IW}(\Psi, \nu)$, i.e.

$$p(\Sigma|\Psi,\nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu n/2}\Gamma_n(\nu/2)}|\Sigma|^{-(\nu+n+1)/2}\exp\left\{-\frac{1}{2}tr[\Psi\Sigma^{-1}]\right\} \tag{41}$$

and then derive the implied density of A from the mappings g and g^{-1} described in (12) and (15).

The density for Σ given by (41) is equivalent to specifying a Wishart density for Σ^{-1} with scale matrix Ψ^{-1}. Thus we might just as well write

$$p(\Sigma^{-1}|\Psi^{-1},\nu) = \left[\frac{1}{2^{\nu n/2}|\Psi^{-1}|^{\nu/2}\Gamma_n(\nu/2)}\right] \times|\Sigma^{-1}|^{(\nu-n-1)/2}\exp\left\{-\frac{1}{2}tr[\Psi\Sigma^{-1}]\right\}. \tag{42}$$

Letting

$$g_{\Sigma^{-1}}(A) = \Sigma^{-1} = AA' \tag{43}$$

we then have

$$p_{RFB}(A|\Psi,\nu) = \left[\frac{1}{2^{\nu n/2}|\Psi^{-1}|^{\nu/2}\Gamma_n(\nu/2)}\right] \times|AA'|^{(\nu-n-1)/2}\exp\left\{-\frac{1}{2}tr(\Psi(AA'))\right\}|J(\Sigma^{-1},A)| \tag{44}$$

where $J(\Sigma^{-1}, A)$ denotes the Jacobian of the transformation from A to Σ^{-1}. Magnus and Neudecker (1980) show that, assuming the upper triangularity of A, one can write J as[29]

$$|J(\Sigma^{-1}, A)| = 2^n \prod_{i=1}^{n} A_{ii}^{i} \ . \tag{45}$$

B.2.2 Prior Density for F

The reduced-form parameters on lagged coefficients of the VAR have density

$$p(\Phi|\Sigma) = (2\pi)^{-kn/2}|\Sigma \otimes \Omega^{-1}|^{-1/2} \exp\left\{-\frac{1}{2}(\beta - \beta^*)'(\Sigma \otimes \Omega^{-1})^{-1}(\beta - \beta^*)]\right\} .$$

where $\beta = vec(\Phi)$ and $\beta^* = vec(\Phi^*)$. Recall that the mapping in (12) defines

$$g_\Phi(F|A) = \Phi = FA^{-1} \tag{46}$$

Hence the density of $F|A$ is given by

$$p_{RFB}(F|A) = p(g_\Phi(F|A))\ |J(\Phi, F|A)| \ . \tag{47}$$

Defining

$$V_{A,\Omega} = (AA')^{-1} \otimes \Omega^{-1} \ , \tag{48}$$

we can write

$$\begin{aligned} p_{RFB}(F|A) = {} & (2\pi)^{-kn/2}|V_{A,\Omega}|^{-1/2} \\ & \times \exp\left\{-\frac{1}{2}(vec(FA^{-1}) - \beta^*)'V_{A,\Omega}^{-1}(vec(FA^{-1}) - \beta^*)\right\} \\ & \times |J(\Phi, F|A)| \ , \end{aligned} \tag{49}$$

[29] See Table 6.2 in Magnus and Neudecker (1980). Assuming that A is upper triangular, the relevant row of the table is (vb).

where

$$J(\Phi, F|A) = \frac{dFA^{-1}}{dF} = \frac{dI_m FA^{-1}}{dF} \tag{50}$$

$$= (A^{-1})' \otimes I_m \,. \tag{51}$$

B.3 Relationship Between RFB Prior and SZ Prior

B.3.1 Densities for A

In this appendix we show that there exist choices of hyperparameters (Ψ, ν) for the inverse-Wishart prior in (41) that yield the SZ prior for A as long as the Jacobian of the transformation is excluded.

Letting k_{RFB} denotes a kernel for the density p_{RFB} we can write

$$k_{RFB}(A|\Psi, \nu) = |AA'|^{(\nu-n-1)/2} \exp\left\{-\frac{1}{2} tr(\Psi(AA'))\right\} |J(\Sigma^{-1}, A)| \tag{52}$$

From (28), the density for A in the SZ prior is given by

$$p_{SZ}(A|H_0) = (2\pi)^{-n^2/2} |(I_n \otimes H_0)|^{-1/2} \exp\left\{-\frac{1}{2} a'(I_n \otimes H_0)^{-1} a\right\} \tag{53}$$

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2} a'(I_n \otimes H_0)^{-1} a\right\} \\ &= k_{SZ}(A|H_0) \end{aligned} \tag{54}$$

where $a = vec(A)$.

We rewrite the exponential term in (52), ignoring the $-1/2$, as

$$tr[\Psi(AA')] = tr[A'\Psi A] \tag{55}$$

$$= vec(\Psi A)' vec(A) \tag{56}$$

$$= ((I_n \otimes \Psi) vec(A))' vec(A) \tag{57}$$

$$= vec(A)'(I_n \otimes \Psi) vec(A) \tag{58}$$

$$= a'(I_n \otimes \Psi) a \tag{59}$$

$$= a'(I_n \otimes \Psi^{-1})^{-1} a \,, \tag{60}$$

which matches the exponential term in (54) with $\Psi^{-1} = H_0$. Thus we can write

$$
\begin{aligned}
&k_{RFB}(A|\Psi^{-1} = H_0, \nu) \\
&\qquad = |AA'|^{(\nu-n-1)/2} \exp\left\{-\frac{1}{2} a'(I_n \otimes H_0)^{-1} a\right\} |J(\Sigma^{-1}, A)| \qquad (61) \\
&\qquad = k_{SZ}(A|H_0)|AA'|^{(\nu-n-1)/2}|J(\Sigma^{-1}, A)| \; . \qquad (62)
\end{aligned}
$$

Noting that

$$
|AA'| = |A||A'| = |A|^2 = \prod_{i=1}^{n} A_{ii} \; , \qquad (63)
$$

where the last equality follows from A's triangularity, we can write

$$
k_{RFB}(A|\Psi^{-1} = H_0, \nu) = k_{SZ}(A|H_0) \left(\prod_{i=1}^{n} A_{ii}^{\nu-n-1} \right) |J(g(A), A)| \; . \qquad (64)
$$

Since the expression for the Jacobian in (45) raises each A_{ii} to a unique power, one cannot find a value ν which cancels all of the terms besides k_{sz} on the right hand side of (64). Thus, as stated in Sims and Zha (1998), the only way to align the two kernels is to exclude the Jacobian. Denoting the resulting kernel $k_{RFB/J}$, one can see that setting $\nu = n + 1$ in

$$
k_{RFB/J}(A|\Psi^{-1} = H_0, \nu) = \left(\prod_{i=1}^{n} A_{ii}^{\nu-n-1} \right) k_{SZ}(A|H_0) \; . \qquad (65)
$$

aligns the kernels:

$$
k_{RFB/J}(A|\Psi^{-1} = H_0, \nu = n + 1) = k_{SZ}(A|H_0) \; . \qquad (66)
$$

B.3.2 Densities for $F|A$

Proposition: If $\Omega^{-1} = H_+$ and $\Phi^* = \bar{S}$ then $p_{SZ}(F|A) = p_{RFB}(F|A)$.

Proof:

The two relevant densities are given by

$$p_{SZ}(F|A) = (2\pi)^{-kn/2} \underbrace{|I_n \otimes H_+|^{-1/2}}_{D_{SZ}} \times \exp\left\{ -\frac{1}{2} \underbrace{vec(F - \bar{S}A)'(I_n \otimes H_+)^{-1} vec(F - \bar{S}A)}_{R_{SZ}} \right\}, \tag{67}$$

which is the prior density for $F|A$ from (29) and

$$p_{RFB}(F|A) = (2\pi)^{-kn/2} \underbrace{|V_{A,\Omega}|^{-1/2}|J(F,A)|}_{D_{RFB}} \times \exp\left\{ -\frac{1}{2} \underbrace{(vec(FA^{-1}) - \beta^*)' V_{A,\Omega}^{-1} (vec(FA^{-1}) - \beta^*)}_{R_{RFB}} \right\}, \tag{68}$$

which is the density from (49). We prove the claim by showing that $R_{SZ} = R_{RFB}$ and $D_{SZ} = D_{RFB}$.

We first show that $R_{SZ} = R_{RFB}$. Note that

$$V_{A,\Omega}^{-1} = ((AA')^{-1} \otimes \Omega^{-1})^{-1} \tag{69}$$

$$= ((AA') \otimes \Omega) \tag{70}$$

$$= (A \otimes I_m)(I_n \otimes \Omega)(A' \otimes I_m) . \tag{71}$$

Letting $\beta = vec(FA^{-1})$ and $\Phi = FA^{-1}$ we derive that

$$(\beta - \beta^*)' V_{A,\Omega}^{-1} (\beta - \beta^*)$$

$$= (\beta - \beta^*)'(A \otimes I_m)(I_n \otimes \Omega)(A' \otimes I_m)(\beta - \beta^*) \tag{72}$$

$$= [(A \otimes I_m)'(\beta - \beta^*)]'(I_n \otimes \Omega)[(A' \otimes I_m)(\beta - \beta^*)] \tag{73}$$

$$= vec(I_m(\Phi - \Phi^*)A)'(I_n \otimes \Omega) vec(I_m(\Phi - \Phi^*)A) \tag{74}$$

Setting $\Phi^* = \bar{S}$ we can see that

$$I_m(\Phi - \Phi^*)A = (FA^{-1} - \bar{S})A \tag{75}$$

$$= (F - \bar{S}A) \tag{76}$$

and then substituting (76) into (74) we have

$$vec(I_m(\Phi - \Phi^*)A)'(I_n \otimes \Omega)vec(I_m(\Phi - \Phi^*)A)$$

$$= vec(F - \bar{S}A)'(I_n \otimes \Omega)vec(F - \bar{S}A) \tag{77}$$

$$= vec(F - \bar{S}A)'(I_n \otimes \Omega^{-1})^{-1}vec(F - \bar{S}A)\,. \tag{78}$$

Setting $\Omega^{-1} = H_+$ completes the proof that $R_{SZ} = R_{RFB}$.

We now show that $D_{SZ} = D_{RFB}$. We first note that

$$|V_{A,\Omega}|^{-1/2} = |(AA')^{-1} \otimes \Omega^{-1}|^{-1/2} \tag{79}$$

$$= |(AA') \otimes \Omega|^{1/2} \tag{80}$$

$$= (|AA'|^m|\Omega|^n)^{1/2} \tag{81}$$

$$= (|A|^{2m}|\Omega|^n)^{1/2} \tag{82}$$

$$= |A|^m|\Omega|^{n/2}\,. \tag{83}$$

Next note that

$$|J(F, A)| = |(A^{-1})' \otimes I_m| \tag{84}$$

$$= |(A^{-1})'|^m|I_m|^n \tag{85}$$

$$= |A|^{-m}\,. \tag{86}$$

Finally using (83) and (86) we have that

$$|V_{A,\Omega}|^{-1/2}|J(F,A)| = |A|^m|\Omega|^{n/2}|A|^{-m} \tag{87}$$

$$= |\Omega|^{n/2} \tag{88}$$

$$= (|I_n|^m|\Omega|^n)^{1/2} \tag{89}$$

$$= |I_n \otimes \Omega^{-1}|^{-1/2}\,. \tag{90}$$

Again setting $\Omega^{-1} = H_+$ completes the proof that $D_{SZ} = D_{RFB}$.

B.4 Details for the Minnesota Prior

The reduced-form based prior is a Minnesota-style prior centered at a random walk. The multivariate-normal-inverse-Wishart density parameterization is set via dummy-observations following closely the procedure in Sims and Zha (1998). Their approach requires three sets of hyperparameters $\bar{y}$, $\bar{\sigma}$, and

$$\Lambda = [\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_5, \mu_6] \tag{91}$$

The first parameter λ_0 controls the overall tightness of the prior. The parameter λ_1 functions similarly to λ_0 but it does not affect beliefs about the constant term. The parameter λ_2 should always be set to 1 in this framework. The parameter λ_3 shrinks the prior for the own lags so that prior standard deviation on lag l shrinks by $l^{-\lambda_3}$. The parameter λ_4 controls tightness of beliefs on the constant term in the VAR. The parameter μ_5 controls what is known as the "sums-of-coefficients" dummy. Higher values give more weight to the view that, if an element of the observables has been near its mean $\bar{y}_i$ for sometime, $\bar{y}_i$ will be a good forecast for that observable, regardless of the values of other observables. This induces correlations between "own" lags of Φ. Finally, μ_6 controls the so-called "co-persistence" dummy observations. The observations are similar to the "sums-of-coefficients", but operate jointly on the observables, inducing correlations among columns of Φ.

C. Normalization and Regime Labeling in the MS-VAR

The MS-VAR posterior density is invariant to sign changes on VAR equations and state labeling. To interpret our results economically we thus have to perform normalization in both of these dimensions.

C.1 Sign Normalization

For each state of $\{A, F\}$, we first normalize each column of the $A((h_m), F(h_m))$ system by sign, forcing nonnegativity of $A(h_m)$'s diagonal elements. When we change the sign of the A_{ii} element to satisfy nonnegativity, we also change the sign of all elements in the ith column of $(A(h_m), F(h_m))$. With the Cholesky identification employed in this paper, this method of sign-normalization implements the "likelihood-preserving" normalization of Waggoner and Zha (2003b).

C.2 Regime Labeling

After normalizing signs we still need to assign regime labels in each draw. To do so, we implement a version of the algorithm described in Stephens (2000) for clustering inference. This algorithm seeks to minimize the the expected loss from reporting a sequence of state probabilities $Q(\theta)$, when the loss function is the Kullback-Leibler divergence of $Q(\theta)$ from the true state probabilities, $P(\theta)$. Hence, the algorithm selects state labels using a rule that has a reasonable decision theoretic foundation. A similar approach used in the population genetics literature is that of Jakobsson and Rosenberg (2007), who minimize a different notion of average distance between $Q(\theta)$ across draws. Both approaches give very similar results to the posteriors reported in the text. We take this approach because it tends to leave less severe multimodality in the posterior after regime labeling.

D. Bimodal Example

D.1 Direct Sampler for Mixture of Posteriors

To generate n_{sim} draws, execute:

Algorithm 3: Direct Sampler for Mixture of Posteriors

for $i = 1, \ldots, n_{sim}$ **do**

1. Draw latent state s_i according to

$$p(s_i = 1) = \alpha$$
$$p(s_i = 2) = 1 - \alpha$$

2. Draw $\Sigma_i | s_i, \Phi_i, Y_{s_i}$, which is a draw from $p(\Sigma_i | \Phi_i, Y_{s_i})$. Under the conjugate prior this is simply $p(\Sigma_i | Y_{s_i})$.

3. Draw $\Phi_i | s_i, \Sigma_i, Y_{s_i}$, which is a draw from $p(\Phi_i | \Sigma_i, Y_{s_i})$.

end

E. RFB-Hierarchical: Hyperparameter Posteriors

Tables A-2, A-3, and A-4 show the posterior mean and 90 percent credible set for the estimated hyperparameters under the RFB-Hierarchical prior from one run of the SMC sampler.

TABLE A-2
POSTERIOR OF Λ

	Mean	[05, 95]		Mean	[05, 95]
			λ_0		
1m1v	0.969	[0.631, 1.405]	2m1v	0.938	[0.708, 1.225]
1m2v	1.226	[0.809, 1.763]	2m2v	1.319	[0.862, 1.820]
1m3v	1.046	[0.725, 1.448]	2m3v	1.025	[0.704, 1.438]
1m4v	1.119	[0.767, 1.566]	2m4v	1.147	[0.816, 1.553]
1m5v	1.057	[0.721, 1.481]	2m5v	0.890	[0.654, 1.203]
			λ_4		
1m1v	0.325	[0.018, 1.083]	2m1v	0.125	[0.012, 0.350]
1m2v	0.174	[0.012, 0.613]	2m2v	0.122	[0.012, 0.307]
1m3v	0.220	[0.012, 0.771]	2m3v	0.162	[0.017, 0.438]
1m4v	0.189	[0.010, 0.613]	2m4v	0.127	[0.014, 0.316]
1m5v	0.183	[0.012, 0.595]	2m5v	0.145	[0.019, 0.348]
			μ_5		
1m1v	1.794	[0.022, 6.278]	2m1v	0.597	[0.001, 2.821]
1m2v	1.824	[0.020, 5.984]	2m2v	0.366	[0.000, 1.785]
1m3v	1.838	[0.025, 6.094]	2m3v	0.405	[0.000, 2.044]
1m4v	1.756	[0.023, 5.654]	2m4v	0.534	[0.000, 2.435]
1m5v	1.726	[0.027, 5.588]	2m5v	0.467	[0.001, 2.103]
			μ_6		
1m1v	3.015	[0.894, 7.131]	2m1v	2.646	[0.852, 5.305]
1m2v	4.619	[1.665, 9.561]	2m2v	3.533	[1.660, 6.275]
1m3v	4.175	[1.300, 9.270]	2m3v	3.256	[1.371, 6.117]
1m4v	4.056	[1.307, 8.655]	2m4v	2.971	[1.232, 5.419]
1m5v	3.709	[1.214, 8.052]	2m5v	2.654	[1.098, 5.198]

TABLE A-3
POSTERIOR OF $\bar{y}$

	Mean	[05, 95]		Mean	[05, 95]
			$\bar{y}_1$		
1m1v	-0.002	[-0.117, 0.110]	2m1v	0.000	[-0.081, 0.079]
1m2v	-0.003	[-0.148, 0.136]	2m2v	-0.007	[-0.117, 0.113]
1m3v	0.000	[-0.136, 0.143]	2m3v	-0.007	[-0.115, 0.113]
1m4v	-0.002	[-0.132, 0.137]	2m4v	0.007	[-0.104, 0.113]
1m5v	-0.004	[-0.131, 0.127]	2m5v	-0.019	[-0.115, 0.097]
			$\bar{y}_2$		
1m1v	0.010	[-0.161, 0.163]	2m1v	0.005	[-0.141, 0.143]
1m2v	0.008	[-0.160, 0.169]	2m2v	0.002	[-0.122, 0.128]
1m3v	0.003	[-0.167, 0.166]	2m3v	0.005	[-0.131, 0.128]
1m4v	0.008	[-0.162, 0.168]	2m4v	-0.030	[-0.131, 0.104]
1m5v	0.011	[-0.157, 0.173]	2m5v	0.009	[-0.120, 0.127]
			$\bar{y}_3$		
1m1v	0.015	[-0.185, 0.198]	2m1v	-0.002	[-0.169, 0.161]
1m2v	0.007	[-0.206, 0.211]	2m2v	0.040	[-0.154, 0.182]
1m3v	0.010	[-0.201, 0.207]	2m3v	0.041	[-0.142, 0.172]
1m4v	0.014	[-0.194, 0.215]	2m4v	-0.045	[-0.177, 0.151]
1m5v	0.009	[-0.202, 0.204]	2m5v	-0.024	[-0.164, 0.157]

TABLE A-4
POSTERIOR OF $\bar{s}$

	Mean	[05, 95]		Mean	[05, 95]
			$\bar{s}_1$		
1m1v	0.023	[0.018, 0.031]	2m1v	0.022	[0.018, 0.026]
1m2v	0.035	[0.024, 0.048]	2m2v	0.032	[0.024, 0.042]
1m3v	0.031	[0.021, 0.045]	2m3v	0.030	[0.022, 0.041]
1m4v	0.031	[0.021, 0.044]	2m4v	0.031	[0.023, 0.040]
1m5v	0.028	[0.019, 0.040]	2m5v	0.024	[0.019, 0.030]
			$\bar{s}_2$		
1m1v	0.028	[0.020, 0.037]	2m1v	0.025	[0.020, 0.031]
1m2v	0.038	[0.025, 0.053]	2m2v	0.032	[0.023, 0.042]
1m3v	0.032	[0.021, 0.046]	2m3v	0.029	[0.020, 0.041]
1m4v	0.032	[0.021, 0.044]	2m4v	0.029	[0.020, 0.040]
1m5v	0.033	[0.021, 0.049]	2m5v	0.026	[0.020, 0.034]
			$\bar{s}_3$		
1m1v	0.023	[0.017, 0.031]	2m1v	0.018	[0.015, 0.022]
1m2v	0.035	[0.023, 0.049]	2m2v	0.028	[0.020, 0.039]
1m3v	0.024	[0.017, 0.035]	2m3v	0.026	[0.019, 0.035]
1m4v	0.026	[0.018, 0.037]	2m4v	0.024	[0.019, 0.032]
1m5v	0.025	[0.018, 0.034]	2m5v	0.021	[0.017, 0.027]

References

BENATI, L. AND P. SURICO (2009): "VAR Analysis and the Great Moderation," *American Economic Review*, 99, 1636–52.

CELEUX, G., M. HURN, AND C. P. ROBERT (2000): "Computational and inferential difficulties with mixture posterior distributions," *Journal of the American Statistical Association*, 95, 957–970.

CHIB, S. AND S. RAMAMURTHY (2010): "Tailored Randomized Block MCMC Methods with Application to DSGE Models," *Journal of Econometrics*, 155, 19–38.

CHOPIN, N. (2002): "A Sequential Particle Filter for Static Models," *Biometrika*, 89, 539–551.

CREAL, D. (2012): "A Survey of Sequential Monte Carlo Methods for Economics and Finance," *Econometric Reviews*, 31, 245–296.

DEL MORAL, P., A. DOUCET, AND A. JASRA (2006): "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society, Series B*, 68, 411–436.

DURHAM, G. AND J. GEWEKE (2012): "Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments," *Unpublished Manuscript*.

FRÜHWIRTH-SCHNATTER, S. (2001): "Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models," *Journal of the American Statistical Association*, 96, 194–209.

——— (2004): "Estimating Marginal Likelihoods for Mixture and Markov Switching Models Using Bridge Sampling Techniques," *The Econometrics Journal*, 7, 143–167.

GELMAN, A. AND D. B. RUBIN (1992): "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, 7, 457–472.

GEWEKE, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1399.

——— (2004): "Getting it right: Joint Distribution Tests of Posterior Simulators," *Journal of the American Statistical Association*, 99, 799–804.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*, vol. 537, Wiley-Interscience.

——— (2007): "Interpretation and inference in mixture models: Simple MCMC works," *Computational Statistics & Data Analysis*, 51, 3529 – 3550.

GIANNONE, D., M. LENZA, AND G. E. PRIMICERI (2015): "Prior selection for vector autoregressions," *Review of Economics and Statistics*, 97, 436–451.

HERBST, E. AND F. SCHORFHEIDE (2014): "Sequential Monte Carlo Sampling for DSGE Models," *Journal of Applied Econometrics*, 29, 1073–1098.

——— (2015): "Bayesian Estimation of DSGE Models," *Mimeo*.

HERBST, E. P. (2012): "Gradient and Hessian-based MCMC for DSGE Models," Unpublished Manuscript, Federal Reserve Board.

HUBRICH, K. AND R. J. TETLOW (2015): "Financial stress and economic dynamics: The transmission of crises," *Journal of Monetary Economics*, 70, 100 – 115.

JAKOBSSON, M. AND N. A. ROSENBERG (2007): "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure," *Bioinformatics*, 23, 1801–1806.

JASRA, A., C. HOLMES, AND D. STEPHENS (2005): "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling," *Statistical Science*, 20, 50–67.

KADIYALA, K. R. AND S. KARLSSON (1997): "Numerical Methods for Estimation and Inference in Bayesian VAR-Models," *Journal of Applied Econometrics*, 12, 99–132.

LEEPER, E. M., C. A. SIMS, T. ZHA, R. E. HALL, AND B. S. BERNANKE (1996): "What Does Monetary Policy Do?" *Brookings Papers on Economic Activity*, 1996, pp. 1–78.

LITTERMAN, R. (1986): "Forecasting with Bayesian Vector Autoregressions: Five Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38.

MAGNUS, J. R. AND H. NEUDECKER (1980): "The Elimination Matrix: Some Lemmas and Applications," *SIAM Journal on Algebraic Discrete Methods*, 1, 422–449.

RUBIO-RAMÍREZ, J. F., D. F. WAGGONER, AND T. ZHA (2010): "Structural Vector Autoregressions: Theory of Identification and Algorithms for Inference," *The Review of Economic Studies*, 77, 665–696.

SIMS, C. AND T. ZHA (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*, 39, 949–68.

SIMS, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48, 1–48.

SIMS, C. A., D. F. WAGGONER, AND T. ZHA (2008): "Methods for inference in large multiple-equation Markov-switching models," *Journal of Econometrics*, 146, 255 – 274.

SIMS, C. A. AND T. ZHA (2006): "Were There Regime Switches in U.S. Monetary Policy?" *The American Economic Review*, 96, 54–81.

SMETS, F. AND R. WOUTERS (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

STEPHENS, M. (2000): "Dealing with label switching in mixture models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 795–809.

STOCK, J. H. AND M. W. WATSON (2010): "Modeling inflation after the crisis," Tech. rep., National Bureau of Economic Research.

UHLIG, H. (1997): "Bayesian Vector Autoregressions with Stochastic Volatility," *Econometrica*, 65, pp. 59–73.

WAGGONER, D. AND T. ZHA (2003a): "A Gibbs sampler for structural vector autoregressions," *Journal of Economic Dynamics and Control*, 28, 349–366.

WAGGONER, D. F. AND T. ZHA (2003b): "Likelihood preserving normalization in multiple equation models," *Journal of Econometrics*, 114, 329 – 347.