working paper 14 27

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

Mark Bognanni and Edward Herbst

FEDERAL RESERVE BANK OF CLEVELAND

Working Paper 14-27 November 2014

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

Mark Bognanni and Edward Herbst

Vector autoregressions with Markov-switching parameters (MS-VARs) offer dramatically better data fit than their constant-parameter predecessors. However, computational complications, as well as negative results about the importance of switching in parameters other than shock variances, have caused MS-VARs to see only sparse usage. For our first contribution, we document the effectiveness of Sequential Monte Carlo (SMC) algorithms at estimating MSVAR posteriors. Relative to multi-step, model-specific MCMC routines, SMC has the advantages of being simpler to implement, readily parallelizable, and unconstrained by reliance on convenient relationships between prior and likelihood. For our second contribution, we exploit SMC's flexibility to demonstrate that the use of priors with superior data fit alters inference about the presence of time variation in macroeconomic dynamics. Using the same data as Sims, Waggoner, and Zha (2008), we provide evidence of recurrent episodes characterized by a flat Phillips Curve.

JEL codes: C11, C15, C32, C52, E3, E4, E5.

Keywords: Vector Autoregressions, Sequential Monte Carlo, Regime-Switching Models, Bayesian Analysis.

Suggested citation: Bognanni, Mark, and Edward Herbst, 2014. "Estimating (Markov-Switching) VAR Models without Gibbs Sampling: A Sequential Monte Carlo Approach, Federal Reserve Bank of Cleveland, working paper no. 14-27.

Mark Bognanni is at the Federal Reserve Bank of Cleveland (markbognanni@gmail.com; http://markbognanni.com). Edward Herbst is at the Federal Reserve Board of Governors (edward.p.herbst@frb.gov). The authors thank Todd Clark, James Hamilton, Ron Gallant, Giorgio Primiceri, and Frank Schorfheide for helpful comments and conversations. They also thank Dan Waggoner for a particularly helpful conference discussion and the other participants of various conferences for their feedback.

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 have a large number of parameters relative to data in typical macroeconomic applications, causing researchers to seek the additional parameter discipline that Bayesian priors can provide. Secondly, researchers have developed methods that make Bayesian estimation of VARs a straightforward task. Posterior sampling is often the most challenging aspect of Bayesian inference, but for VARs a family of known priors yields (conditional) posteriors amenable to an efficient posterior sampling algorithm called the Gibbs sampler.

The interests of macroeconometricians have recently moved beyond the basic VAR framework to specifications with time-varying parameters. Sims and Zha (2006) pioneered the extension of the structural VAR framework to include Markov-switching parameters (MS-VARs) and use their model to infer the cause of the "Great Moderation." Based on the results of their MS-VAR estimation Sims and Zha (2006), conclude that "good luck" remains the most parsimonious explanation consistent with the data. As a byproduct of their inquiry, Sims and Zha (2006) also document dramatically superior data fit of MS-VARs when compared to constant parameter specifications.

Yet, despite the ability of MS-VARs to illuminate time-varying relationships in the data, few researchers besides Sims and Zha (2006) and Sims et al. (2008) have used MS-VARs in econometric applications. We suspect that the sparse use of MS-VARs owes to the complicatedness of the estimation process. MS-VARs do not admit Gibbs samplers with the efficiency or simplicity of their constant-parameter predecessors. Sims et al. (2008) describe a four-step process for MS-VAR. First, search (in a high dimensional parameter space) for the posterior mode from which to initialize 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. Fourth and finally, code

a complicated extension of the modified harmonic mean (MHM) algorithm for marginal data density (MDD) estimation. In a recent paper investigating the macroeconomic effects of financial crises (and a notable exception to the hesitance of economists to use MS-VARs), Hubrich and Tetlow (2014) estimate their model using 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 7 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." Of course, even at the end of this process uncertainty would remain about whether or not one has found the true posterior mode in the first step.

Motivated by these difficulties, we use an alternative class of algorithms called Sequential Monte Carlo (SMC) to estimate MS-VARs. We demonstrate that SMC allows for simple and accurate posterior inference for MS-VARs. Furthermore, SMC makes it easy to use alternative priors. In our MS-VAR estimation, we show that using a prior based on that typically applied to the analysis of reduced-form VARs, improves model fit and substantially alters posterior inference about macroeconomic dynamics.

SMC algorithms begin by propagating a set of "particles" from the prior distribution, where each particle is 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 is incorporated. Importantly, the effectiveness of SMC does not rely on any particular analytical convenience of the posterior except for the ability to generate random draws from the prior and to evaluate a posterior kernel pointwise; a far weaker set of restrictions than those required for an effective Gibbs Sampler.

For our first contribution, we demonstrate that SMC algorithms give extremely reliable estimates of VAR marginal data densities, and that one can obtain accu-

rate results when using quantities of particles readily implementable on current computers. We show solid performance by SMC under a variety of choices for the algorithm's tuning parameters, but also highlight a few small changes to existing SMC implementations that yield particularly dramatic performance improvements for VARs. One aspect of particular importance is accounting for the nontrivial correlation structure of parameters typically present in both VAR priors and posteriors.

For our second contribution, we use SMC to estimate an MS-VAR similar to the model in Sims et al. (2008). 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 reduced-form prior, we find nearly 50% posterior weight on a model that features a periodically flattening Phillips Curve, in addition to changing structural shock variances. The results in our paper suggest that prior choice deserves careful attention when working with densely parameterized MS-VARs.

We also want to emphasize that our ability to readily estimate MS-VARs results directly from the genericness of the SMC algorithm. 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. For two reasons, this fact should not be discounted. Firstly, the genericness allows us to explore the implications of using alternative priors for MS-VARs. Secondly, as emphasized by Geweke (2004), 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.

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 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 of interest. In Section 4 we estimate a suite of MS-VAR models and describe the results. In Section 5 we conclude.

2. The Sequential Monte Carlo Algorithm

In this section we describe the details of the general SMC algorithms used in this paper. Chopin (2002), Del Moral, Doucet, and Jasra (2006), Creal (2012) and Herbst and Schorfheide (2014) offer additional details on SMC implementation. The Bayesian researcher is interested in the posterior density $p(\theta|Y)$, which is given by

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

where $p(\theta)$ denotes the prior density and $p(Y|\theta)$ denotes the likelihood of the parameters θ under the observed data Y. The term $p(Y)$ is known as the "marginal data density" (MDD) or "marginal likelihood", an important measure of model fit.[1] For ease of exposition, we will 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

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

[1] Giannone, Lenza, and Primiceri (2013) use the MDD to select among prior densities of a common family for VARs.

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 groups of 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.[2] Unfortunately, MS-VARs lack priors that induce such tractable posteriors. The goal of this paper is to develop an SMC algorithm that robustly overcomes the challenges and opacity of MS-VAR posteriors.

Importance sampling (IS) serves as the keystone of SMC. Indeed, the SMC method we use in this paper is sometimes known as Iterated Batch Importance Sampling. Under IS, the target density f is approximated by another, easy-to-sample density g. Importance sampling is based on the identity

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

Suppose that $\theta^i \overset{iid}{\sim} g(\theta)$, $i = 1, \ldots, N$. 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 which approximates $\pi(\theta)$. The $\tilde{W}^i$'s are known as the (normalized) importance weights assigned to each particle value θ^i. The accuracy of the approximation is driven by the distance between $g(\cdot)$ and $f(\cdot)$ and is reflected in the distribution of the weights. If the distribution of weights is very uneven, 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

[2]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).

$\pi(\theta)$.

Unfortunately, constructing "good" importance distributions, g, is difficult for densely parameterized models in which the econometrician does not know the shape of f.[3] We recursively build particle approximations to a *sequence* of distributions, starting from a known distribution (in this case the prior), then slowly adding information from the likelihood until we have obtained the posterior. Specifically, we index the distributions by n,

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

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 $\phi_1 = 0$ implies that the initial target distribution, $\pi_1(\theta)$, is simply the prior, $p(\theta)$. Hence, we can initialize the algorithm by propagating the particles as random draws from the prior. Our algorithm thus requires that one can sample from the prior directly, but this is a far weaker restriction than is restricting oneself to the use of priors currently standard in the estimation of VARs. The choice of $\phi_{N_\phi} = 1$ implies that the final target distribution, $\pi_{N_\phi}(\theta)$, is the posterior.[4] Thus our final particles will approximate the distribution of interest to the researcher. The general form of our algorithm is the same as the one used in Herbst and Schorfheide (2014). We describe the algorithm here for completeness and then follow with discussion.

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). We enter stage n 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

[3] Note that VAR estimation has a history with importance sampling: Leeper, Sims, Zha, Hall, and Bernanke (1996) and Uhlig (1997) developed importance samplers for VAR models.

[4] The "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 $f_n(\theta) = p(Y_{1:n}|\theta)p(\theta)$. This 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 the particles $\{\hat{\theta}_n^i, W_n^i\}$ 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 $\pi_n(\theta)$.

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.

particles have meaningful weight, the particles are rejuvenated using multinomial resampling. This process ensures that 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 particle individually).

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 Markov Chain Monte Carlo (MCMC) techniques (standard for posterior estimation of VARs), 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 at the beginning of the sampler, and more quickly as the sampler moves on. We discuss the role of λ in more detail in Section 3.

Algorithm 1 presents the generic 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 be crucial for 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 point of departure from Herbst and Schorfheide (2014) is our construction of the proposal variance in the block Metropolis-Hasting algorithm, an important determinant of the efficacy of the sampler. Herbst and Schorfheide (2014) use the estimation of the marginal variance for the bth parameter block, that is, the submatrix of $\hat{\Sigma}_n$,

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

We find that this is a suboptimal choice for densely parameterized models, as it ignores the relationship between the b block of parameters and the other "conditioning" parameters. To account for this, we use the multivariate normal approximation to the conditional variance,

$$\hat{\Sigma}_{b,n} = [\hat{\Sigma}_n]_{b,b} - [\hat{\Sigma}_n]_{b,-b}[\hat{\Sigma}_n]^{-1}_{-b,-b}[\hat{\Sigma}_n]_{-b,b}. \tag{9}$$

While this change may appear small, we will show in Section 3 that it can improve the efficacy of the sampler greatly because of the complex correlation structure inherent in (MS)VAR models.

The contribution of this paper is not theoretical, 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). 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

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 $b \in B_n$.

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, though the variances associated with the CLTs have the formulation given in Chopin (2002), the recursive form is, unfortunately, not useful in practice. Instead, we rely on estimates computed across multiple independent runs of the algorithm as in Durham and Geweke (2012).

Finally, a few key questions remain about the use of the algorithm in practice. How should one choose n_ϕ (or λ)? How many particles should one use? How many blocks should one use? While theoretical results on the optimal hyperparameters for estimation are beyond the scope of this paper, in the next section we will exploit the relative transparency of VARs to move beyond the suggestions of Herbst and Schorfheide (2014) and find well-performing choices for tuning parameters.

3. Testing Sequential Monte Carlo in Practice

Before moving to the application of interest, MS-VARs, we first test our SMC algorithm's effectiveness at estimating the MDD of two models for which we know the true MDD in closed-form: 1) a reduced-form VAR with conjugate prior and, 2) as a more challenging test, a mixture of reduced-form VAR posteriors.

3.1 Constant-Parameter VAR

A three-variable, three-lag, reduced-form VAR serves as our starting point. We test the SMC algorithm on two parameterizations of the VAR commonly explored in the literature. The first, given by

$$y_t' = \Phi_0 + y_{t-1}'\Phi_1 + y_{t-2}'\Phi_2 + y_{t-3}'\Phi_3 + u_t', \quad u_t \sim \mathcal{N}(0, \Sigma) \ , \tag{10}$$

is referred to as the reduced-form VAR. Letting $\Phi = [\Phi_0', \Phi_1', \Phi_2', \Phi_3']'$, the reduced-form VAR has parameters $D = \{\Phi, \Sigma\}$. The second, given by

$$y_t'A = y_{t-1}'F_1 + y_{t-2}'F_2 + y_{t-3}'F_3 + \varepsilon_t' \,, \quad \varepsilon_t \sim \mathcal{N}(0, I) \,, \tag{11}$$

is referred to as the structural VAR. Collecting the matrices $F_1, \ldots, F_n$ as $F = [F_1', F_2', F_3']'$, the structural VAR has parameters $S = \{A, F\}$. In the absence of restrictions on the values of either D or S both VAR representations have the same likelihood function.

We estimate both reduced-form and structural versions of the VAR and compare the MDD estimates to the true MDD. Since there exists a conjugate prior for D for which we know the MDD in closed form, the exercise is simple for the reduced-form model.[5] We discuss the conjugate prior's features in Section 4 but for now it suffices to know that the prior is standard in the literature. Unfortunately, the prior for S described in Sims and Zha (1998), which is standard in the structural VAR literature, does not allow a closed-form expression for the MDD. To take advantage of the closed-form expression for the VAR's MDD as a convenient benchmark, we need a prior for S that gives the same MDD as our prior for D. If we assume that $A = (chol(\Sigma)')^{-1}$, where $chol$ refers to the lower-triangular Cholesky factor of Σ, then there exists a one-to-one mapping between D and S given by the pair of functions g and g^{-1} defined as

$$g_1(\Phi, \Sigma) = A = (chol(\Sigma)')^{-1} \tag{12}$$

$$g_2(\Phi, \Sigma) = F = \Phi A \tag{13}$$

and

$$g_1^{-1}(A, F) = \Sigma = (AA')^{-1} \tag{14}$$

$$g_2^{-1}(A, F) = \Phi = FA^{-1} \,. \tag{15}$$

The Cholesky factor identification is common in the literature and yields an

[5] Appendix A gives the details on the expression for the MDD.

exactly identified structural VAR in the sense of Rubio-Ramírez, Waggoner, and Zha (2010). We can then estimate S under the prior

$$p(S) = p_D(g^{-1}(S)) \,|\, \det(J(g^{-1}(S)))|, \tag{16}$$

which is the prior over S induced by the prior on D.

To estimate S with the prior density in (16) we need to be able to both evaluate and sample from $p(S)$. Evaluating (16) requires only the straightforward application of $g^{-1}(S)$ to evaluate $p_D(g^{-1}(S))$ and accounting for the Jacobian term associated with the transformation g^{-1}, which we derive in Appendix A.3. We can sample from the density in (16) by first sampling from $p(D)$ and then applying the transformation $g(D)$.

3.2 Assessing SMC: Baseline Performance

Since the SMC algorithm produces, as a by-product, an estimate for the marginal data density, we use the MDD to gauge the accuracy of the SMC estimator. To assess the algorithm's performance we run a Monte Carlo experiment using SMC on both D and S.[6] 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 will use again when estimating Markov-switching models in Section 4.

Recall that the SMC sampler in Section 2 features a number of hyperparameters 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 $\lambda = 4$, which is approximately our best performing choice of hyperparameters when asessing hyperparameter combinations according to both estimation bias and root mean squared errors. We run 20 Monte Carlo replications of the sampler for each choice of hyperparameters and examine the distribution of $\ln(MDD)$ estimates.

The first row of Table I shows the results, giving the average bias (Avg Bias) and the root mean squared error (RMSE) of the estimates of $\ln p(Y)$ for both the

[6] In both the reduced form and the structural case, the true log MDD is the same, as the structural relationships do not affect the models description of the data.

reduced-form and structural parameterizations. We can see that the sampler is quite accurate under both parameterizations of the algorithm. The mean error of the log marginal data density is 0.01 percent of the true value (1791.9), while the root mean squared error is less than 0.3 in both cases, or less than 0.02 percent. 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 for more effective moves.[7]

3.3 Assessing SMC: Importance of 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 $\lambda = 1, 7$. As before, 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 I 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.

The first set of deviations we consider, line two in Table I, shows when (8)

[7] 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.

TABLE I
SMC ESTIMATES OF $\ln p(Y)$ FOR VAR(3): EFFECTS OF ALGORITHM TUNING PARAMETERS

SMC Tuning Parameters							VAR Parameterization			
							Reduced Form		Structural	
Σ_{prop}	N_{part}	N_{blocks}	Blocking	N_ϕ	M	λ	Avg Bias	RMSE	Avg Bias	RMSE
Cond	2000	3	Random	500	1	4	0.01	0.29	0.01	0.21
Un	-	-	-	-	-	-	0.01	1.37	−0.08	1.90
-	1000	-	-	-	-	-	0.01	0.39	0.02	0.47
-	5000	-	-	-	-	-	0.00	0.19	0.00	0.11
-	-	1	-	-	-	-	0.05	0.61	0.08	0.50
-	-	2	-	-	-	-	0.02	0.39	0.03	0.38
-	-	2	(Φ, Σ)	-	-	-	0.02	0.44	−0.32	3.95
-	-	3	Row	-	-	-	0.01	0.26	−0.02	0.75
-	-	4	(Row, Σ)	-	-	-	0.00	0.18	−0.11	1.51
-	-	-	-	50	10	-	0.00	0.43	−0.02	1.33
-	-	-	-	-	-	1	−0.27	1.87	−0.77	4.02
-	-	-	-	-	-	7	0.00	0.41	0.01	0.37

Notes: The symbol " - " indicates inheritance of the parameter value from the baseline parameterization given in the first line of the table. "Avg Bias" refers to mean error of the estimate of $\ln p(Y)$ and RMSE is the root mean squared error of the estimates. The true value is 1791.9.

is used as the RWMH proposal variance rather the conditional approximation, given by (9). 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 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 I, 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 I, 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 I 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 I, 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.

3.4 Accuracy of SMC for Irregular VAR Posteriors

In Section 4 we apply SMC to the estimation of MS-VARs. Sims et al. (2008) stress that the posterior of MS-VARs "tends to be non-Gaussian" and may well contain "multiple peaks." Indeed, we find evidence of fat-tailed and multipeaked posterior densities in our posterior draws, even after normalizing them. This leads us to ask, does SMC deliver reliable performance for multipeaked posteriors? To answer the question 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 means our simulation has empirical relevance.

When estimating the MDD with draws sampled directly from the target distribution, we test the modified harmonic mean (MHM) method originally considered in Geweke (1989) and the version that Sims et al. (2008) adapt for better performance with non-Gaussian distributions.[8] Since iid draws represent an upper bound on the usefulness of MCMC draws[9], if SMC performs similarly to existing MDD estimators when we supply them with iid draws then we would conclude that SMC performs as well as any MCMC algorithm ever could as long as the researcher had to rely on currently available methods of estimating the MDD from the MCMC output. Thus we implicitly compare SMC to MCMC.

We construct the bimodal target distribution as follows. Let θ be the parameters of a model, $p(\theta)$ be a prior over those parameters, $p(Y_i|\theta)$ be the model's likelihood function for observations Y_i, and

$$p(Y_i) = \int_{\Theta} p(\theta)p(Y_i|\theta)\,. \tag{17}$$

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

[9] In principle, it possible to use Monte Carlo to obtain more precise estimates relative to iid draws (i.e., antithetic variates), in this environment it is impractical

Let L take the form

$$L(Y_1, Y_2|\theta) = \alpha p(Y_2)p(Y_1|\theta) + (1-\alpha)p(Y_1)p(Y_2|\theta)d\theta \,, \tag{18}$$

which we call a pseudo-likelihood. We take α as given and known, so we implicitly condition on this value. We then consider the following pseudo-posterior for θ

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

We can rewrite the denominator as

$$\begin{aligned}
&\int_\Theta p(\theta)[\alpha_1 p(Y_2)p(Y_1|\theta) + (1-\alpha_1)p(Y_1)p(Y_2|\theta)]d\theta \\
&= \alpha_1 p(Y_2)\int_\Theta p(\theta)p(Y_1|\theta)d\theta + (1-\alpha_1)p(Y_1)\int_\Theta p(\theta)p(Y_2|\theta)d\theta \qquad (20) \\
&= p(Y_1)p(Y_2) \qquad (21)
\end{aligned}$$

and hence we know the integrating constant in closed-form as long as we know $p(Y_1)$ and $p(Y_2)$. Some simple algebra (see Appendix C.1) reveals that $p(\theta|Y_1, Y_2)$ equals the distribution that would result from a mixture of posteriors with a common prior of $p(\theta)$ and likelihoods $p(Y_1|\theta)$ and $p(Y_2|\theta)$, hence we can easily sample the distribution directly.[10]

Since we know their MDD in closed-form, we use reduced-form VARs as the mixture components in our example. We generate 50 replications of pseudo-posterior draws and MDD estimates via each of the methods mentioned above. Table II shows the results of our simulation for a VAR($n = 3, p = 5$), from which we arrive at three main conclusions.

Firstly, the MHM extension developed by Sims et al. (2008) performs extraordinarily well (at least compared to traditional MHM), even in the presence of bimodality, when given iid draws from the target distribution. Even though the SWZ estimator constructs its approximating density around one of the distribution's modes, the approximating density has fat enough tails to reliably

[10] See Appendix C.2 for the simple direct sampling algorithm.

TABLE II

ESTIMATES OF $\ln p(Y)$ FOR MIXTURE OF VAR POSTERIORS UNDER DIFFERENT METHODS OF POSTERIOR SAMPLING AND MDD ESTIMATION

Posterior Sampler	MDD Estimator	Avg Bias	RMSE
SMC: $N_{part} = 2000$	SMC	0.122	0.421
SMC: $N_{part} = 5000$	SMC	0.133	0.337
Direct: 10,000 draws	MHM - SWZ	0.187	0.311
	MHM	0.682	1.068
Single Mode	MHM - SWZ	−0.812	0.812
	MHM	−0.829	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.

incorporate information throughout the parameter space.

Secondly, and most central to our interests, SMC estimates the MDD as well as SWZ when we give SWZ an i.i.d. sample of 10,000 draws, despite the fact that SMC must simultaneously sample the posterior of a model whose parameters are treated as unknown. Since researchers typically simulate posterior draws using MCMC algorithms, as with the MS-VARs we estimate in the next section, and since our simulation represents an upper bound on the performance of the SWZ algorithm, we conclude that the SMC algorithm shows superior performance in the presence of substantive multimodality. Furthermore, the presence of substantive multimodality in posteriors gives us little reason for concern when estimating a model with SMC.

Thirdly, bimodality renders MDD estimation via the MHM method of Geweke (1999) 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.

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. However, we do find it worthwhile to document the stakes of proper posterior sampling. The basic concern when using an MCMC sampler with multimodal target distributions is that the sampler may not mix properly in a reasonable amount of time, particularly when the single chain relies on local moves via random walk-MH steps. In the worst-case scenario the MCMC sampler never leaves a neighborhood around the mode nearest to the point from which the algorithm initialized.[11] The rows in Table II labeled "Single Mode" show MDD estimates computed from just such a caricature of a failed MCMC algorithms, i.e. the draws are simulated from only one of the two modes. In such a situation the results are disastrous.

4. Application of Interest: Structural MS-VARs

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 alternative prior specifications. We will show that the use of an off-the-shelf prior, common in the analysis of reduced-form VARs, significantly alters posterior inference about the presence of structural changes to the macroeconomy.

[11] 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 exist MCMC methods able to handle the known-*a priori*-symmetric multimodality of mixture models. Frühwirth-Schnatter (2001) gives the details on effective MCMC methods for mixture models. Unlike the examples in Celeux et al. (2000) and Geweke (2007), our results from MS-VAR estimation in the subsequent section suggest, in addition to the typical symmetric multimodality which can be normalized away, the presence of asymmetric posterior multimodality. The target density in the present section possesses that property by construction.

4.1 The Model: Structural MS-VAR

We consider MS-VAR models of the form

$$y_t' A(s_t) = x_t' F(s_t) + \varepsilon_t' \Xi(s_t)^{-1}, \quad \varepsilon_t \sim iid\mathcal{N}(0_n, I_n) \tag{22}$$

$$\Xi(s_t) = diag([\xi_1(s_t), \ldots, \xi_n(s_t)]) \tag{23}$$

$$p(s_t | S_{t-1}, Y_{t-1}, \theta, q) = q_{s_t, s_{t-1}} \tag{24}$$

$$q_{s_t=i, s_{t-1}=j} = q_{i,j} \quad \forall i, j, t, \tag{25}$$

where $\Xi(s_t)$ is an $n \times n$ diagonal matrix, s_t is the joint state of the latent process at time t, 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 $t-1$.

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

$$A = \{A(h)\}_{h \in \{1,\ldots,H\}}, \quad F = \{F(h)\}_{h \in \{1,\ldots,H\}}, \quad \Xi = \{\Xi(h)\}_{h \in \{1,\ldots,H\}}. \tag{26}$$

We then let $\theta = \{A, F, \Xi\}$. Note that we use the set notation in (26) 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. For example, one might restrict the regime-switching so that $A(1)$ and $A(2)$ differ by only their last column, which is one specification that Sims and Zha (2006) and Sims et al. (2008) consider. The state of the latent process at time t may be determined by the joint realization of K independent Markov processes, which determine the state of different sets of parameters. 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 probability model for the data we described in (22)-(25) belongs to the class of models considered in Sims et al. (2008) and, in particular, has the same general form as their empirical application. Our only point of departure from their model 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 will find that not imposing those restrictions allows the model to achieve superior

data fit.

For ease of comparison with Sims et al. (2008), we assume that $\{A, F\}$ and $\{\Xi\}$ follow independent regime-swi tching processes and thus $\theta_1 = \{A, F\}$ and $\theta_2 = \{\Xi\}$. 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.

The MS-VAR has the conditional likelihood

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

and the likelihood

$$
p(y_t|\theta, q, Y_{t-1}) = \sum_{h=1}^{H} p(y_t|\theta, q, s_t, Y_{t-1}) \, p(s_t|s_{t-1}) \, . \tag{28}
$$

and

$$
p(Y_T|\theta, q) = \prod_{t=1}^{T} p(y_t|\theta, q, Y_{t-1}) \tag{29}
$$

To evaluate the likelihood one must filter the sequence of state probabilities in (28). We filter the probabilities using the algorithms derived in Sims et al. (2008).

We estimate MS-VAR models with a variety of choices for H_m and H_v using the data described in Section 3) and lag length ($p = 5$). For each choice of H, we estimate the model under four different priors for (A, F). In each case, the prior on $\{(A(k), F(k))\}_{k=1}^{H_m}$ is identical and independent across k. We now provide additional details on the construction of each prior.

SWZ Prior. This is the prior used in Sims et al. (2008). For each state k, the prior takes the form

$$a(k) \sim \mathcal{N}(0, I_n \otimes H_0) \tag{30}$$

$$f(k)|a(k) \sim \mathcal{N}(vec(\bar{S}A(k)), I_n \otimes H_+) \,, \tag{31}$$

where

$$\bar{S} = \begin{bmatrix} I_n \\ 0_{(n(p-1)+1)\times n} \end{bmatrix} \,. \tag{32}$$

Here $a(k) = vec(A(k))$, $f(k) = vec(F(k))$ and H_0, H_+ are prior parameters. 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 $\bar{\sigma}$, and vector of hyperparameters, $\Lambda = [\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_5, \mu_6]$, 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 $\bar{\sigma}$ as the standard deviations of the residuals from univariate autoregressions for each data series, both of which we follow here.[12] 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}.

The prior has the desirable properties of invariance of density with respect to orthogonal rotations of (A, F) (see Rubio-Ramírez et al. (2010) for proof) and, in constant-parameter VARs, accessibility of conditional posteriors that allow Gibbs sampling via the algorithm in Waggoner and Zha (2003a).[13] However, a less desirable property is the structural prior's strong shrinkage of A towards zero. We find this property unreasonable since it implies $\Sigma = \infty$ at the modal value of A.

Sims and Zha (1998) note that the prior for A in (41) is equivalent to what

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

[13] A well known property of the SVAR's likelihood is invariance with respect to orthogonal rotations of (A, F). It then seems reasonable that the prior density should have the same property.

one would derive from beliefs about $\Sigma = (AA')^{-1}$ that had the inverse Wishart distribution with $n + 1$ degrees of freedom and while also ignoring the Jacobian term for the transformation from $A \to \Sigma$. The modal value of $\Sigma = \infty$ is, in fact, a by-product of the ignored Jacobian term. In practice, one can interpret the resulting beliefs about A as derived from an inverse Wishart distribution with 1 degree of a freedom, a value too low for the distribution to actually exist, since it would imply unbounded density as $\Sigma \to \infty$. We address this shortcoming in the next prior we consider.

Sims-Zha Reduced-Form-Based (RFB-SZ) Prior. We derive a prior for (A, F) by placing a prior distribution over the reduced-form dynamics, summarized by Φ and Σ, then mapping to $(A(k), F(k))$ using equations (12) and (13) as discussed in Section 3. With $n + 1$ degrees of freedom for Σ, this is exactly the prior that Sims and Zha (1998) note would likely give superior performance, but which they could not use for computational reasons. With appropriate choices of hyperparameters, the RFB prior differs from the SWZ prior only in that it includes the Jacobian, which serves to recenter beliefs about A away from 0, while preserving invariance to orthogonal rotations. Given A, the prior on F is equivalent to what one would derive from the RFB prior.

RFB - Constant-Parameter (RFB-CP). The third prior we consider features choices for Λ that are informed by the MDD of a constant-parameter VAR (CP-VAR). When using the RFB prior we know the MDD for the CP-VAR in closed form for any choice of the Λ, making it simple to understand a given Λ's the consequences for model fit. It turns out that Λ_{SWZ} is a particularly poor performing choice, which affects posterior model comparison. We do not fully maximize the MDD of the constant-parameter VAR with respect to Λ because we fear that the resulting prior would be too tight to allow time-varying features of the MS-VAR to come through.[14] One might think of our tighter VAR hyperparameters as an alternative to the proportionality based shrinkage imposed in Sims and Zha (2006) and Sims et al. (2008), but an alternative which is favored by the data.

[14] Indeed, we ran simulations in which excessive shrinkage towards "optimal" Λ for the CP-VAR resulted in lower MS-VAR MDDs.

We change Λ_{SWZ} to have $\lambda_0 = 0.4$, $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.2$, $\lambda_4 = 1.0$, $\mu_5 = 3.0$, and $\mu_6 = 1.0$. These choices make three substantive changes to the basic RFB prior. Note that for λ_i parameters, lower values imply tighter beliefs, while for μ_i parameters, higher values imply tighter beliefs. First, we loosened the prior on the constant term ($\lambda_4 = 1.0$ instead of 0.1). Second, beliefs are otherwise generally tighter ($\lambda_0 = 2.5$ instead of 1.0 and $\mu_5 = 3.0$ instead of 1.0). Third, we increase the degrees of freedom of the inverse Wishart distribution to 7, as would be standard for this model in the reduced-form analysis of VARs.[15] We refer to these hyperparameters as Λ_{CP}

RFB - Constant Parameter, Population Mean (RFB-CP-PM). The dummy observations used to construct the priors draw significant information from the means of the observations that initialize the lagged data in the VAR, $\bar{y}$. The fourth prior we consider uses Λ_{CP}, but forms $\bar{y}$ from the means over the entire sample rather than from only the initial observations. On one hand, this does "use the data twice" and thus violates tenets of strict Bayesian inference. On the other hand, it speaks to the issue that shrinking towards the mean from one specific period (the lagged data) may not be advisable in a setting where means are modelled as potentially changing over time. Moreover, this is similar in spirit to the common practice of choosing prior hyperparameters based on posterior information (i.e., the marginal data density).

Priors on Other Parameters. All the priors share common specifications for the volatilities and transition regimes. For the volatilities, each $p(\xi_j(k))$ independent and identically distributed such that

$$\xi_j^2(k) \sim \mathcal{G}(\bar{\alpha}_j, \bar{\beta}_j) \tag{33}$$

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 to $\xi_j(1) = 1$ 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

[15] See the Appendix for details on this point.

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.

4.1.1 Estimation Details

Under each prior, we estimate MS-VARs for $H_m = 1, 2$ and $H_v = 1, \ldots, 6$ using the SMC algorithm described above. We set $N_{part} = 2000$, $N_{blocks} = 8$ (random) and $M = 1$, using the conditional variance given by the normal approximation for the mutation step. We set the tempering schedule with $\lambda = 4$ and $N_\phi = 2000$.[16] We estimate each model 20 times to assess the stability of the sampler.

The SMC sampler was written in `Fortran` and the calculations were executed on 12-core desktop with an Intel Xeon x5670 CPU. Estimation of a given model takes between one and ten minutes, with likelihood evaluations parallelized across the 12 cores. A `Matlab` version executing on the same machine roughly takes between twenty minutes to six hours, depending on the number of states.

We could, in principle, also simulate from the posteriors using the sampler proposed by Sims et al. (2008) for the Structural Prior and modify the Metropolis-within-Gibbs stages of the sampler to accommodate the RFB prior. However, both we and other researchers have found the MCMC estimation process cumbersome and lengthy. Indeed, experimentation across models indicated difficulties with finding the (a) posterior mode reliably 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.

[16] For a few specifications, we set $N_{part} = 4000$ to increase estimation precision.

TABLE III
COMPARISON OF LOG MDDS OF MS-VARS TO CONSTANT-PARAMETER VARS

	Prior			
Model	SWZ	RFB-SZ	RFB-CP	RFB-CP-PM
Constant Parameter VAR	1759.41	1759.62	1770.47	1772.86
MS-VAR: Best - 1m3v-4v	1872.75	1877.13	1879.92	1882.59
MS-VAR: 2nd worst	1867.14	1873.03	1876.16	1878.73
MS-VAR: Worst - 2m1v	1844.55	1846.32	1852.51	1856.46

Notes: The best fitting MS-VAR is always either 1m3v or 1m4v.

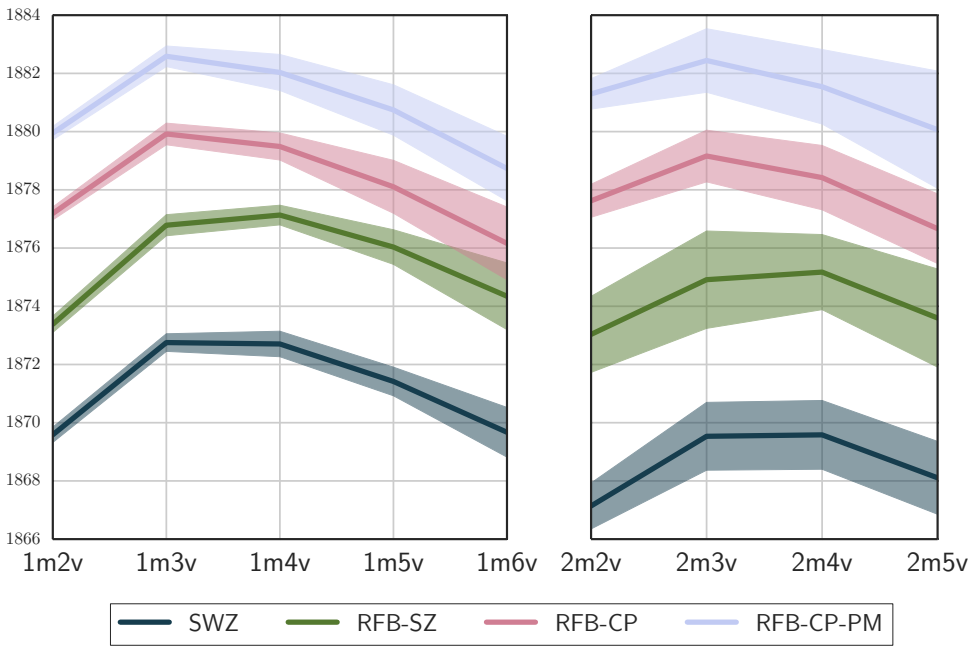
4.2 Estimation Results and Model Comparison

From our estimation results we deduce four main findings. First, 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). Table III shows the point estimates of the log MDDs of a constant parameter VAR and various MS-VARs estimated under each prior. For all models the typical MDD gains exceed a staggering 100 log points.

Second, according to point estimates, the best fitting model under each prior is a only-variances-switch model. One can see this point clearly from Figure 1, which shows the means and standard deviations (across twenty simulations) of the log MDDs from estimating each MS-VAR specification under each prior.[17] Furthermore, regime-switching in structural shock variances is critical for data fit. As is evident from Table III, the 2m1v model is the worst fitting of all MS-VAR specifications by a significant margin regardless of which prior we use. These findings are consistent with Sims and Zha (2006) and Sims et al. (2008) and they form the foundation of their "good luck" narrative of the Great Moderation. However, estimation with successively better fitting priors increases the fragility of the conclusion that a only-variances-change model fits the data best. We return to this point below.

[17] Table A-1, in the Appendix, contains the exact numerical values.

FIGURE 1.—Model log MDDs Across Priors



Notes: The figure shows the point estimate for the log MDD of each model and under each prior (the thin lines). The shaded region shows ± 1.96 times the standard error around each point estimate. We omit the 1m1v and 2m1v because they are the worst fitting models by significant margins.

Third, each successive prior gives higher MDDs for all regime-switching models. This is important. This means that each change to the prior is first and foremost favored by the data rather than any particular specification. Thus the increasing posterior probability If, for example, we had formed a hierarchical model in which we started with discrete, uniform prior beliefs over each of our Λ vectors, then the our final results would be very similar to those of the top line in Figure 1.[18]

Fourth, under successively better-performing priors the 2m models merit increasing attention (and perform impressively well considering their number

[18]Giannone et al. (2013) estimate a constant-parameter VAR in which they form a hierarchical prior for the "overall tightness" hyperparameter, λ_0.

of free parameters). Table IV shows how the posterior probability of on 2m regime changes conditional on each prior, along with the unconditional posterior probability on all specifications for each prior. Under the SWZ Prior, there is negligible probability on changes in the mean parameters; that is, the $2m$ models.[19] However these models garner nearly 50% of the posterior probability both for the best-fitting prior and overall if one takes the model-selection interpretation that we described in above. Furthermore, for the best-fitting prior the 2m3v and 1m3v models are separated by margins well within the range of standard errors for the MDD estimates. This finding contrasts with the results in Sims and Zha (2006) and Sims et al. (2008). 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. The concern expressed in Sims et al. (2008) was that allowing all parameters to 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.[20]

The fact that choices of Λ and $\bar{y}$ matter substantially in both data fit and posterior model probabilities lays bare the delicacy of prior selection for MS-VARs. On the one hand, the dramatic increase in the dimension of the parameter space increases the need for shrinkage. On the other hand, designing good shrinkage schemes becomes more difficult when we estimate a model to discover time-varying features of the data that we may know little about a priori. Since brief episodes may display dynamics that we might consider unreasonable if they occurred for the entire sample, when adapting a prior developed for estimating constant parameter VARs to the estimation of MS-VARs we must take care to ensure that our prior allows for such parameterizations. For example, explosive VAR dynamics seem more reasonable a priori for MS-VARs than VARs since

[19] Relatedly, the SMC algorithm has more difficulty estimating the $2m$ model MDDs precisely under the structural prior, as seen by the larger standard deviations likely owing to its inferior data fit.

[20] We also find such restrictions undesirable on theoretical grounds. As is well-known, and was pointed out particularly starkly in Benati and Surico (2009), all coefficients of the VAR representation of a DSGE model should be expected to change if one changes the DSGE models' policy rule parameters.

TABLE IV
POSTERIOR MS-VAR PROBABILITIES OF 2M MODEL FOR EACH PRIOR

Prior	$P(2m\|Prior, Y)$	$P(Prior\|Y)$
SWZ	0.04	0.00
RFB-SZ	0.12	0.00
RFB-CP	0.30	0.05
RFB-CP-PM	0.46	0.94

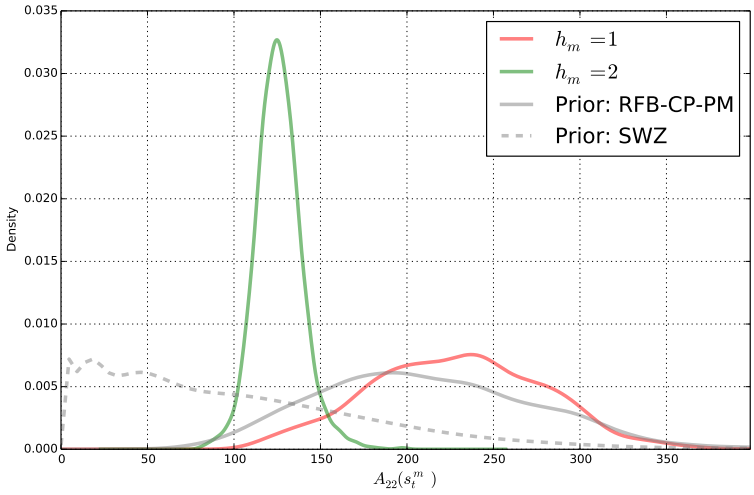
Notes: The second column gives the posterior probability of the 2m models, conditional on the prior. The third column gives the posterior probability on all models for a given particular prior.

they may approximate dynamics that occur for only brief periods.[21] One might interpret the proportionality restrictions imposed in Sims et al. (2008) as an attempt at striking this balance. However, the MDDs we achieve without these restrictions clearly show the suboptimality of the dogmatic shrinkage based on proportionality.

It's worth going into a bit more detail on how the prior can affect inference in these kind of models. We focus on the extreme cases: the SWZ prior and the RFB-CP-PM prior. Figure 2 plots kernel density estimates of the posterior for $A_{22}(s_t^m)$ for 2m3v model under the RFB-CP-PM prior. In our model comparison exercise, this model receives substantial probability, indicating a role for changing mean dynamics, the economics of which we discuss in the next section. The green line traces the posterior density for $A_{22}(s_t^m)$ when $h_m = 2$, while the the red line is the posterior density for $A_{22}(s_t^m)$ when $h_m = 1$. The solid grey line is the prior density of the RFB-CP-PM (which the same for both regimes), while the grey dashed line is the prior density under the SWZ prior. It turns out that centering A_{22} at zero (i.e., shrinking towards $\Sigma = \infty$) is, in fact, one key reason for the different model rankings across prior choice. From the plot of A_{22} one can see that the SWZ prior assigns nearly zero probability mass to the portion

[21] For a theoretical justification of this intuition, consider a DSGE model with a monetary policy rule that evolves as in Davig and Leeper (2007).

FIGURE 2.—Priors and Reduced Form Posterior on $A_{22}(s_t^m)$



Notes: Figure shows density estimates for SWZ and RFB-CP-PM for elements of $A_{22}(s_t)$ (the dashed and solid grey line, respectively) as well as the posterior under the RFB-CP-PM prior for both $h_m = 1$ (the red line) and $h_m = 2$ (the green line).

of the parameter space inhabited by what we call the flat Phillips Curve regime (described in the next section), thus ruling out the presence of such a regime *a priori*. Indeed, this shrinkage is so severe that when estimating the 2m3v model under the SWZ prior, the model throws away the second regime for $\{A, F\}$, since the prior density does not allow their values to move to locations in the parameter space that most improve data fit.

Finally, it is, of course, difficult to know for certain that our posterior estimates are correct. However, this difficulty is only more severe with the alternative MCMC posterior samplers. We feel that the results in Section 3 leave us with compelling reasons to trust our results in this section. Recall that we can execute multiple independent runs of our algorithm and compare the precision of results. Indeed, Table A-1 includes standard errors from independent runs of the algorithm for each model, rather than from different subsets of draws along a single MCMC chain. Since we initialize each run by an independent draw of initial particles, there is no risk of the precision resulting from influential initial conditions, i.e.,

it is highly unlikely that we would get similar estimates across runs by random chance unless they were, in fact, near the true answer. Hence, we take the precision of our estimates to reflect accuracy. Recall also that, SMC performs well in the example based on mixtures in Section 3.4, which we take as evidence that the multipeaked posteriors that other MS-VAR users have raised concerns about, pose little danger for SMC.[22]

In the remainder of this section we focus on describing the important features of the $2m3v$ model estimated under the RFB-CP-PM prior (the best fitting one), which, as mentioned above, fits the data well and has not received much attention in the literature, to the best of our knowledge. To be sure, our estimation exercise indicates substantial weight should be placed on $m1$ models as well; these have been studied before in, for example, Sims and Zha (2006).

4.3 Examining the 2m3v Model: Time-Series of Regimes and Economic Interpretation

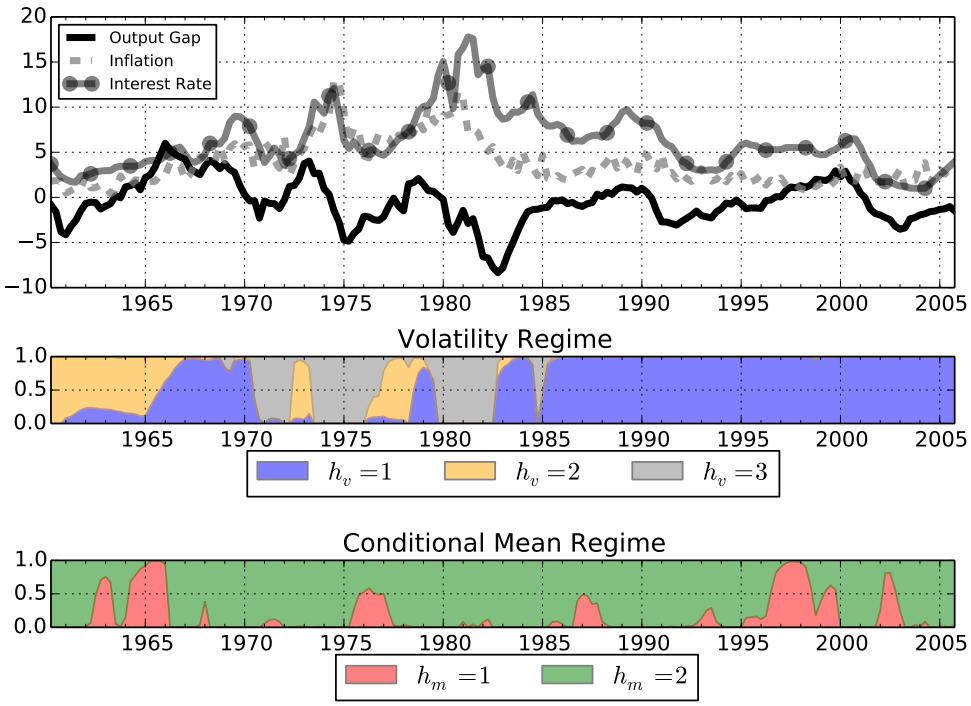
Figure 3 summarizes the estimated regime-occurrences for the 2m3v model. The first plot shows the data, the second plot shows the probabilities of $h_v = 1$ and $h_v = 2$ with $p(s_t^v = 2)$ stacked on top of $p(s_t^v = 1)$, and the third plot shows the probability of being in regime $h_m = 1$. We computed the probabilities at the highest density parameter from the SMC sampler.

4.3.1 Time-Varying Volatilities

Turning to the shock volatility regimes, shown in the middle panel of 3, 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. 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

[22]Note that our posterior distributions for the MS-VAR parameters also show evidence of irregularities and multipeakedness, corroborating the characterizations described in other papers. From Figure 2 one can see that even after normalizing regime labels, the posterior density for the values of $A_{33}(h^m = 1)$ tend to exhibit multimodality. As we alluded to in Section 3.4, the dangers that such irregularities pose for MCMC remains a subject a debate.

FIGURE 3.—Observables and Regime Probabilities



Notes: The figure shows the data used in estimation (top panel) together with the smoothed posterior probabilities for the volatily regimes (middle panel) and the conditional mean regimes (bottom panel).

as a once-and-for-all decrease in shock volatilities in line with a "good luck" explanation. The "good luck" regime also prevailed during the late 1960s.

4.3.2 Time-Varying Means

From the second panel of Figure 3 one can see that the $h_m = 1$ regime dominated the mid-1960s as well as well as the late 1990s, with some marginally likely occurrences in second half of the 1970s, second half of 1980s, and 2003. The $h_m = 2$ regime prevails for the remainder of the sample.

The key feature of both periods in which $h_m = 1$ occurs with certainty is that the output gap is closing while inflation and the nominal interest rate remain relatively stable. Thus the model points to repeated episodes of diminished

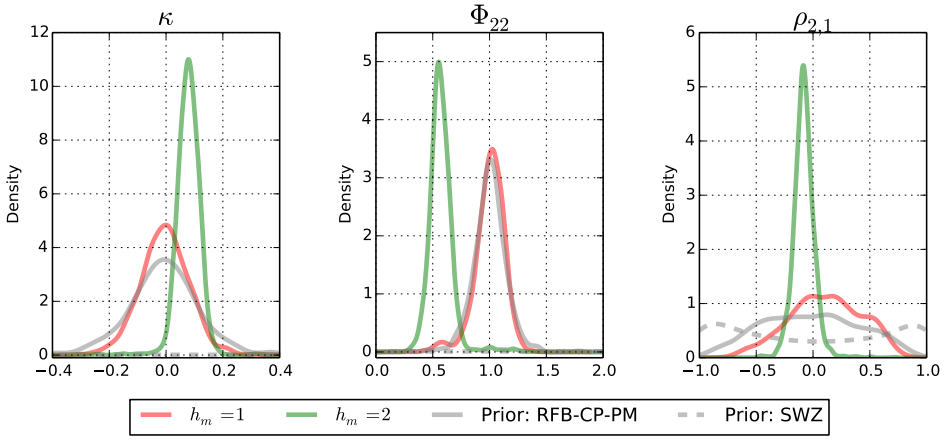
economic slack without significant upward pressure on inflation. While a $1m$ (or constant coefficient DSGE) model would ascribe these outcomes to a particular sequence of shock realizations (or perhaps volatility regimes), the $2m$ model finds substantial shifts structural economic dynamics. Both explanations are equally plausible, according to our posterior probabilities.

To further investigate this result, we concentrate on the joint dynamics of the output gap and inflation under the two mean regimes. Specifically, we compute estimates of the slope of the Phillips Curve, which we call κ, under each regime.[23] The left panel of Figure 4 shows the posterior of κ under each regime.[24] One can see that the posterior estimates of κ under $h_m = 1$ (the red line) are centered at 0, while under $h_m = 2$ (the green line) the estimates are centered near 0.1. Hence, we refer to $h_m = 1$ as the flat Phillips Curve regime and point to one interpretation our results as evidence of a periodic flattening of the Phillips Curve. One can also see from the the plot that the $h_m = 1$ estimate of κ is similar–but not identical–to the its prior. The structural prior is much more diffuse on the object. It is flat over a wide range of possible values for κ. The middle panel of 4 shows similar kernel density estimates for Φ_{22}, the reduced-form weight given to π_{t-1} when predicting π_t. Under $h_m = 1$, inflation exhibits random-walk behavior, with Φ centered at 1, while the inflation is markedly less persistent when $h_m = 2$. Finally, density estimates for the reduced-form correlations of the innovations to the output gap and inflation, ρ_{21} are shown in the right panel. Again we see a marked difference between states $h_m = 1$ and $h_m = 2$.

[23] We compute κ by first transforming $\{A, F\}$ into the reduced-form parameters and then summing the coefficients on lags of the output gap in the equation that predicts inflation. We have also computed κ as a ratio of the cumulative impulse responses over eight quarters of inflation and the output gap to a monetary policy shock. The results from the second exercise are virtually the same as the results shown in the paper.

[24] 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, which is a well-known annoyance and is 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. We handle label switching by trying a variety of methods proposed in the literature on mixture-models; all methods give similar results and Appendix B contains the details.

FIGURE 4.—Priors and Reduced Form Posterior



Notes: Figure shows posterior density estimates under the RFB-CP-PM prior for κ, Φ_{22}, and ρ for both $h_m = 1$ (the red line) and $h_m = 2$ (the green line). The RFB-CP-PM and SWZ prior densities for these objects are also shown in the solid and dashed grey lines, respectively.

4.4 Connection to Other Literature

Reduced form approaches have yielded nonlinearity in the Phillips curve. Stock and Watson (2010) and many references therein document a nonlinear relationship between the traditional gap measures and inflation. They show that the strongest Phillips curve relationship occurs in recessions which is roughly consistent with the finding here: that the relationship between inflation and economic slack deteriorates during (some) periods of quickly diminishing slack.

One can also examine economic dynamics during the $h_m = 1$ period in a fixed coefficient structural general equilibrium model. That these periods represent a structural change in economic dynamics is, in a sense, visible from the historical decompositions implied by relatively rich NK-DSGE models, such as the model of Smets and Wouters (2007). For example, 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 and 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 these innovations suggests a dimension of model misspecification and our results suggest that economic dynamics may well have changed.

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 reduced-form prior, we find nearly 50% posterior weight on a model that features a periodically flattening Phillips Curve, in addition to changing structural shock variances.

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.

A. VAR Priors

A.1 Reduced-Form Prior Parameterization

The standard conjugate prior for a VAR gives Inverse-Wishart beliefs about Σ and Gaussian beliefs about $\Phi|\Sigma$.

$$\Sigma \sim \mathcal{IW}(\Psi, d) \tag{34}$$

$$vec(\Phi)|\Sigma = \mathcal{N}(vec(\Phi^*), \Sigma \otimes \Omega^{-1}) \tag{35}$$

where Ψ,d,Φ^*, and Ω are (matrices of) 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

$$\Omega = X^{*'}X^* \tag{36}$$

$$\Phi^* = \left(X^{*'}X^*\right)^{-1} X^{*'}Y^* \tag{37}$$

$$\Psi = \left(Y^{*'}Y^*\right) - \left(\Phi^{*'}\Omega\Phi^*\right) \tag{38}$$

$$d = T^* - m\,, \tag{39}$$

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

Given the data and choices of prior hyperparameters, the MDD of the VAR is given in closed form by the expression

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

A.2 Structural Prior Parameterization

The reference prior of Sims and Zha (1998) for structural VARs, as described in Rubio-Ramírez et al. (2010), takes the form

$$a \sim \mathcal{N}(0, I_n \otimes H_0) \tag{41}$$

$$f|a \sim \mathcal{N}(vec(\bar{S}A), I_n \otimes H_+) \,, \tag{42}$$

where

$$\bar{S} = \begin{bmatrix} I_n \\ 0_{(n(p-1)+1)\times n} \end{bmatrix} \,. \tag{43}$$

and $a = vec(A)$, $f = vec(F)$ and H_0, H_+ are prior parameters.

A.3 RFB Prior for Structural VAR Parameters

A.3.1 Prior Density for A

Our reduced-form-based prior for A is derived from the fact that, in the absence of changing shock variances, $(AA')^{-1} = \Sigma$. 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.

$$p(\Sigma|\Psi, \nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu n/2}\Gamma_n(\nu/2)}|\Sigma|^{-(\nu+n+1)/2} \exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\Psi]\right\} \tag{44}$$

and then derive the implied density of A. Let the function g map a symmetric positive definite matrix into the inverse of the transpose of its Cholesky decomposition matrix. So

$$g(\Sigma) = \left(chol(\Sigma)^{-1}\right)' = A \tag{45}$$

$$g^{-1}(A) = (AA')^{-1} = \Sigma \tag{46}$$

We can then define a density directly over A as

$$h(A) = p\left(g^{-1}(A)\right)|J(A)| \tag{47}$$

$$J(A) = dg^{-1}(A). \tag{48}$$

Note that g is one-to-one since the Cholesky decomposition, inverse, and transpose each yields a unique matrix. The mapping will necessarily be onto as well, as long as we restrict our interest to A matrices that are the Cholesky factors of some Σ.

Applying results from Magnus and Neudecker (1988) we derive $J(A)$ as follows

$$\frac{d(AA')^{-1}}{dA} = \left(\frac{d(AA')^{-1}}{d(AA')}\right)\left(\frac{dAA'}{dA}\right) \tag{49}$$

$$\frac{d(AA')^{-1}}{d(AA')} = -\left((AA')'\right)^{-1} \otimes (AA')^{-1} \tag{50}$$

$$\frac{dAA'}{dA} = 2N_n(A \otimes I_n) \tag{51}$$

$$N_n = \frac{1}{2}\left(I_{n^2} + K_{nn}\right) \tag{52}$$

where K_{nn} is the commutation matrix between $vec(A)$ and $vec(A')$. So we have

$$\frac{d(AA')^{-1}}{dA} = \left[-\left(AA'\right)^{-1} \otimes \left(AA'\right)^{-1}\right] \times 2N_n\left(A \otimes I_n\right) \tag{53}$$

$$= \left[-\left(AA'\right) \otimes \left(AA'\right)\right]^{-1} \times 2N_n\left(A \otimes I_n\right) \tag{54}$$

Now accounting for the fact that Σ and A each have only $n(n+1)/2$ unique elements, we derive

$$J(A) = D^{+}_{n,\Sigma}\left(\frac{d(AA')^{-1}}{dA}\right) M_A \tag{55}$$

where M_A is such that

$$M_A vech(A) = vec(A)\,. \tag{56}$$

Summing up we have

$$h(A) = p\left((AA')^{-1}\right)\left|\det\left(D^{+}_{n,\Sigma}\left(\frac{d(AA')^{-1}}{dA}\right)M_A\right)\right| \tag{57}$$

where $d(AA')^{-1}/dA$ is given in (54).

A.3.2 Prior Density for F

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

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

The mapping of reduced-form to structural parameters is given by

$$g(\Phi|A) = \Phi A = F \tag{58}$$

$$g^{-1}(F|A) = FA^{-1}. \tag{59}$$

Hence the density of $F|A$ will be given by

$$h(F|A) = p(g^{-1}(F|A))\,|J| \tag{60}$$

where

$$J = \frac{dFA^{-1}}{dF} = \frac{dI_m FA^{-1}}{dF} \tag{61}$$

$$= (A^{-1})' \otimes I_m \tag{62}$$

Summing up we have

$$h(F|A) = p\left(FA^{-1}\right)\left|\det\left(\frac{dFA^{-1}}{dF}\right)\right| \tag{63}$$

and $d(FA^{-1})/dF$ is given in (63).

A.4 Relationship Between RFB and Structural Priors

It turns out that priors for $F|A$ described in (42) and (64) are equivalent for $H_+ = \Omega^{-1}$. Thus, when moving back and forth between the RFB prior and the structural prior the only inherent difference is the inclusion of the Jacobian term in the prior for A.

A.5 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 λ.

$$\Lambda = [\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \mu_5, \mu_6] \tag{64}$$

The first parameter λ_0 controls the overall tightness of the prior. The parameter λ_1 functions similarly to $lambda_0$ 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.

B. Normalization 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.

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).

After normalizing signs, we assign regime labels via an ordering by the size of a particular coefficient in $\{A, F\}$ that most reduces the multimodality in the posterior densities of parameter values. In the 2m models we use the A_{22} element to assign regime labels. Assigning labels to the volatility regimes works basically the same way, where we order on ξ_2.

We have also implemented 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.

C. Bimodal Example

C.1 Equivalence of "Pseudo-Posterior" and Mixture of Posteriors

Consider the mixture of posteriors

$$\widetilde{p}(\theta|Y_1, Y_2) = \alpha_1 \left(\frac{p(\theta)p(Y_1|\theta)}{p(Y_1)} \right) + (1 - \alpha_1) \left(\frac{p(\theta)p(Y_2|\theta)}{p(Y_2)} \right) . \tag{65}$$

We can rewrite the mixture as

$$\widetilde{p}(\theta|Y_1, Y_2) = \frac{p(Y_2)\alpha_1 p(\theta)p(Y_1|\theta) + p(Y_1)(1-\alpha_1)p(\theta)p(Y_2|\theta)}{p(Y_1)p(Y_2)} \tag{66}$$

$$= \frac{p(\theta)[\alpha_1 p(Y_2)p(Y_1|\theta) + (1-\alpha_1)p(Y_1)p(Y_2|\theta)]}{p(Y_1)p(Y_2)}\,, \tag{67}$$

which matches (19).

C.2 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

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.

DAVIG, T. AND E. M. LEEPER (2007): "Generalizing the Taylor Principle," *American Economic Review*, 97, 607–635.

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

DEL NEGRO, M. AND F. SCHORFHEIDE (2011): "Bayesian Macroeconometrics," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford University Press, chap. 7, 293–389.

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.

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

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

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

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

GIANNONE, D., M. LENZA, AND PRIMICERI (2013): "Prior Selection for Vector Autoregressions," *Review of Economics and Statistics*, Forthcoming.

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

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

HUBRICH, K. AND R. J. TETLOW (2014): "Financial Stress and Economic Dynamics: The Transmission of crises," *Journal of Monetary Economics*, Forthcoming.

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.

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 (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons.

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.

D. Tables

TABLE A-1
SMC ESTIMATES OF log(MDD) FOR MARKOV-SWITCHING VAR($n = 3, p = 5$) MODELS.

		Prior							
		SWZ		RFB		RFB-CP		RFB-CP-PM	
Model		log(MDD)	(S.E.)	log(MDD)	(S.E.)	log(MDD)	(S.E.)	log(MDD)	(S.E.)
1m	1v	1759.41	(0.08)	1759.62	(0.10)	1770.47	(0.10)	1772.86	(0.08)
1m	2v	1869.59	(0.13)	1873.38	(0.14)	1877.19	(0.11)	1879.95	(0.12)
1m	3v	1872.75	(0.15)	1876.78	(0.18)	1879.92	(0.19)	1882.59	(0.18)
1m	4v	1872.70	(0.22)	1877.13	(0.17)	1879.49	(0.24)	1882.03	(0.32)
1m	5v	1871.42	(0.25)	1876.03	(0.30)	1878.10	(0.46)	1880.74	(0.44)
1m	6v	1869.67	(0.44)	1874.34	(0.58)	1876.16	(0.64)	1878.73	(0.56)
2m	1v	1844.55	(2.30)	1846.32	(1.75)	1852.51	(0.33)	1856.46	(0.44)
2m	2v	1867.14	(0.40)	1873.03	(0.67)	1877.62	(0.29)	1881.30	(0.27)
2m	3v	1869.53	(0.59)	1874.91	(0.85)	1879.16	(0.45)	1882.44	(0.56)
2m	4v	1869.58	(0.60)	1875.17	(0.66)	1878.42	(0.56)	1881.54	(0.65)
2m	5v	1868.11	(0.64)	1873.59	(0.86)	1876.66	(0.61)	1880.07	(1.03)

Notes: log($p(Y)$) estimates are means from 20 independent runs of the algorithm for each model, from which we also compute standard errors, given in *italics*. The values for the SMC algorithm hyperparameters are $N_{part} = 2000$, $\lambda = 4$, N_{blocks} varies across models to keep about 10 parameters per block, $N_{\phi} = 2000$, and $M = 1$. The hyperparameters we choose give stable estimates of the log(MDD) across multiple runs of the algorithm on the 2v2m model.