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

Mark Bognanni*
Federal Reserve Bank of Cleveland

Edward Herbst†
Federal Reserve Board of Governors

This Draft: July 31, 2014

Abstract

In principle, Bayesians can use Markov chain Monte Carlo (MCMC) algorithms to consistently estimate the posterior moments of a wide range of models. In practice however, using MCMC algorithms to estimate highly-parameterized models requires researchers to use conjugate priors and code model-specific samplers. We use Sequential Monte Carlo (SMC) algorithms, whose effectiveness requires no particular analytical relationship between prior and likelihood, to estimate VAR models without conjugate priors. We highlight two examples of practical interest in which our choice of algorithm is crucial for making correct posterior inference. First, for a constant-parameter, structural VAR we augment a Minnesota prior with a prior derived from a standard Taylor rule. Using SMC methods to sample the posterior, we show that the Taylor Rule prior is consistent witht the data and improves model fit. Second, we show that the normalized posterior of a Markov-Switching SVAR can possess multimodality that may cause MCMC samplers to deliver misleading results.

Keywords: Vector Autoregressions, Sequential Monte Carlo, Regime-Switching Models, Bayesian Analysis
JEL codes: C11, C32, C18, C51, C52

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

†The views expressed in this paper do not reflect those of Federal Reserve Bank of Cleveland, the Federal Reserve Board, or the Federal Reserve System.

1 Introduction

The use of vector autoregressions (VARs) has grown steadily since Sims (1980) and VARs now serve as a vital element of the empirical macroeconomist's toolkit. Among both academics and practitioners, Bayesian methods predominate for VAR estimation for two reasons. Firstly, VARs' large number of parameters relative to available data causes researchers to seek additional discipline on parameter values, which Bayesian priors provide. Secondly, researchers can easily implement Bayesian estimation of VARs, owing to the complementary innovations of priors that yield (conditional) posteriors amenable to direct sampling—e.g., the Minnesota prior—and Markov chain Monte Carlo (MCMC) algorithms for sampling from the full posterior by sampling sequentially from conditional posteriors, namely the so-called Gibbs sampler.

However, not all potentially useful VAR specifications admit straightforward Gibbs samplers. One such VAR specification is the structural VAR with Markov switching parameters (MS-VAR), which Sims and Zha (2006) use in order to make inference about the causes of the "Great Moderation." The Gibbs Sampler used to estimate the model in Sims and Zha (2006), and which Sims, Waggoner, and Zha (2008) describes more fully, requires Metropolis-Hastings steps within the algorithm to handle a subset of parameters whose conditional posterior one cannot sample directly. We demonstrate that an alternative class of algorithms called Sequential Monte Carlo (SMC) is both dramatically simpler and more precise than the current method for estimating MS-VARs. Furthermore, SMC makes the use of alternative priors simple and we show that modifications to the standard structural, which we argue are desirable for reasons apparent a priori, both improve model fit and change substantive interpretation about the sources of the "Great Moderation."

SMC algorithms begin by propagating a set of "particles" from the prior dis-

tribution, where each particle is a vector of values of the model's parameters. The algorithm then moves and reweights the particles to approximate a new distribution that combines the prior and a small amount of information from the likelihood. SMC then uses then updates the particles approximating the previous distribution in order to approximate a new distribution that includes slightly more information from the likelihood than did the previous step. The algorithm iterates on this process until the full likelihood's worth of information is included in the target distribution, which is then the model's posterior. Importantly, the effectiveness of SMC does not rely on any particular analytical convenience of the posterior, but rather relies only on the ability to generate random draws from the prior and to evaluate a posterior kernel pointwise. Thus, SMC dramatically expands the spaces of usable priors and estimable model specifications. SMC's virtues include flexibility across models, ease of parallelization, and simplicity of implementation.

For our first contribution, we demonstrate that SMC algorithms give extremely reliable estimates of VAR marginal data densities 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.

For our second contribution, we use SMC to estimate a structural VAR with Markov-switching parameters (MS-VAR), similar to the model in Sims et al. (2008), and show: 1) substantial multi-modality in the (normalized) posterior that makes estimation with existing methods unreliable, 2) SMC can estimate MDDs of VAR-sized, multi-modal posteriors more accurately than existing combinations of MCMC samplers and MDD estimators. The use of SMC for the estimation of MS-VARs shows particularly large gains relative to the alternatives since MS-VARs lack a pure

Gibbs sampler under any known prior.[1]

Lastly, we want to emphasize that our ability to readily estimate both applications 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, MS-VARs, and VARs with time-varying parameters, each of which relies on a unique posterior sampler when using MCMC for estimation. This fact should not be discounted. As emphasized by Geweke (2004), reliance on model-specific Gibbs samplers for posterior simulation typically involves a lengthy process of tedious algebra and coding, which lends itself well to making difficult-to-detect errors.

With regards to the estimation algorithm, our paper builds on 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.

With regards to applications, our paper builds on the seminal work on structural MS-VARs in Sims and Zha (2006) and Sims et al. (2008), which developed MCMC routines for MS-VAR estimation and augmented the modified harmonic mean estimator to better handle the non-Gaussian posteriors of MS-VARs. However, despite the advances in Sims and Zha (2006) and Sims et al. (2008), few other researchers have tried to estimate MS-VARs, which we suspect owes mainly to the computational difficulty of the estimation process. A notable exception is Hubrich and Tet-

[1] We take the phrase, "pure Gibbs sampler," to refer to a Gibbs sampler for which one can sample directly from each conditional posterior density via known algorithms. Hence, a Gibbs sampler that includes "Metropolis-within-Gibbs" steps is not a pure Gibbs sampler.

low (2013), who use MS-VARs to estimate the effects of financial crises on economic dynamics. Hubrich and Tetlow (2013) also provide the following useful summary of the practical challenges of implementing current estimation methods based on MCMC, "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." Thus, when assessing the efficacy of our estimation methods, one should keep in mind that the benchmark for comparison is an estimation process that can take upwards of one week. Relative to the other tools available for MS-VAR estimation, we find SMC dramatically easier to implement while achieving superior accuracy.

From here the rest of the paper proceeds as follows. In Section 2 we describe the estimation algorithm and its place within the larger SMC literature. In Section 3 we demonstrate the algorithm's effectiveness in a setting in which we have closed-form expressions for the objects of interest. In Sections **??** and 4 we estimate VAR models that would be extremely difficult with available MCMC methods. In Section 5 we conclude.

2 The Sequential Monte Carlo Algorithm

In this section we describe the SMC algorithm in detail. The Bayesian researcher is interested in the posterior density $p(\theta|Y)$, which is given by

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

Here, $p(\theta)$ denotes the prior density and $p(Y|\theta)$, the likelihood of the θ under the observed data Y.[2] The term $p(Y)$ is the so-called "marginal data density" (MDD) or "marginal likelihood", an important measure of model fit. Indeed, Giannone, Lenza, and Primiceri (2013) use this measure to select a prior density in the case of VARs. 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

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

For VARs, the literature has previously concentrated on restricted (conjugate) priors which induce a posterior such that $\pi(\theta)$ 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. The goal of this paper is to present a robust algorithm for estimating (MS)VAR models without such restrictions–that is, in models without conjugacy. While such models could, in principal, be estimated by Markov chain monte carlo (MCMC) techniques, the size of the parameter vector precludes efficient estimation. We argue that Sequential Monte Carlo (SMC) techniques can overcome these limitations. Here, we describe the basics of the algorithm. Detailed explanations are in Chopin (2002), Del Moral, Doucet, and Jasra (2006), and Creal (2012).

The keystone of SMC is importance sampling (IS). Under IS, the target density f is approximated by another, easy-to-sample density g. Importance sampling is based on the identity

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

[2]In an (S)VAR context, it is common for the parameter vector θ to be composed of a set of parameters defining the autoregressive coefficients Φ, and variance matrix, Σ. Where appropriate, we will substitute these objects for θ.

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

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

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 particle approximation of $\pi(\theta)$. The $\tilde{W}^i$'s are 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 $\pi(\theta)$.

As we will see, constructing "good" importance distributions g is difficult for densely parameterized models, particularly those without conjugate priors, in which case the econometrician does know the shape of f. [3] The SMC method we use here to sidestep this difficulty is sometimes known as Iterated Batch Importance Sampling (IBIS). 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 constructed the posterior. Specifically, if we index the distributions by n,

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

The scaling parameter ϕ_n is an increasing sequence, such that $\phi_1 = 0$ and $\phi_{N_\phi} = 1$. Then the initial distribution, $\pi_1(\theta)$, equals $p(\theta)$, the prior, while $\pi_{N_\phi}(\theta) = p(\theta|Y)$,

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

the posterior. [4] 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. Starting with a particle approximation to π_{n-1}, we describe the three steps to obtain a particle approximation to π_n, using the nomenclature Chopin (2002) The general form of our algorithm is the same as the one used in Herbst and Schorfheide (2014), which we describe here for completeness. We first present the details of the algorithm and then follow with discussion.

The algorithm can be divided into the three phases listed above. We enter stage n with a particle approximation $\{\theta_{n-1}, \tilde{W}_{n-1}^i\}_{i=1}^{N_{part}}$ of π_{n-1}. In the first component of stage n, the correction step, the particles are reweighted according to π_n. This is an importance sample of π_n using π_{n-1} as the proposal distribution. In the second step, selection, if the sample is unbalanced in the sense that only a few particles have meaningful weight, the particles are rejuvenated using multinomial resampling. This birth and death process ensures that sampler avoids the well-known issue of particle impovereshment. On the other hand, the resampling itself induces noise into the simulation, and so we avoid doing it unless necessary. Finally, in the third step, mutation, particles are moved around the parameter space, using a Metropolis-Hastings algorithm. The step is crucial as it allows particles to move towards areas of higher density of π_n. Without this step, particle values would be fixed at the initial prior draws. Combined with the resampling step, it is likely that few unique values would survive until the final approximation, resulting in a poor approximation to the posterior. An additional point to stress about the mutuation step is that each particle forms its own independent chain. This stands in constrast to standard

[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. Indeed, they have a reduced-form VAR in their example.

Algorithm 1: Simulated Tempering SMC

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

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

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

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

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

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

$$\tilde{h}_{n,N} = \frac{1}{N}\sum_{i=1}^N h(\theta_{n-1}^i)\tilde{W}_n^i. \tag{6}$$

2. **Selection.** Compute the effective sample size $ESS_n = N/\left(\frac{1}{N}\sum_{i=1}^N (\tilde{W}_i^n)^2\right)$.

   If $ESS < N_{part}/2$

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

3. **Mutation.** Propagate the particles $\{\hat{\theta}_i, W_n^i\}$ via M steps of a MH algorithm with transition density $\theta_n^i \sim K_n(\theta_n|\hat{\theta}_n^i; \zeta_n)$ and stationary distribution $\pi_n(\theta)$. We use a block random walk Metropolis-Hastings algorithm with (scaled) proposal given by the current particle estimate of the variance. An approximation of $\mathbb{E}_{\pi_n}[h(\theta)]$ is given by

$$\bar{h}_{n,N} = \frac{1}{N}\sum_{i=1}^N h(\theta_n^i)W_n^i. \tag{7}$$

end

Markov Chain Monte Carlo (MCMC) techniques for posterior estimation of VARs, in which a single chain is generated. This allows us to exploit parallel computations during the mutation step, greatly speeding up the algorithm, a point highlighted by both Durham and Geweke (2012) and Herbst and Schorfheide (2014). Overall, the mutation step is key to the tuning of the sampler.

The contribution of this paper is not theoretical, so we will not go into great detail about the formal arguments proving the strong law of large numbers (SLLN) and central limit theorem (CLT) for the particle approximation. We merely provide high level assumptions and state the theorems without proofs. Interested readers should refer to Chopin (2002) for more details.

First, we place the same assumptions on the prior and likelihood function as in Herbst and Schorfheide (2014):

Assumption 1. *Suppose that (i) the prior is proper:* $\int p(\theta)d\theta < \infty$; *(ii) the likelihood function is uniformly bounded* $\sup_{\theta\in\Theta} p(Y|\theta) < M < \infty$*; and (iii)* $\int [p(Y|\theta)^{\theta_2} d\theta > 0$.

The first assumption is important given the recursive nature of the algorithm. We are further implicitly assuming that one can exactly sample from the prior. Moreover, as we have argued previously, the existing SVAR literature has satisfied itself with an even more restrictive class of priors. The second assumption, while difficult to verify in practice, ensures that the weights are bounded. The third assumption merely says that the prior has nonzero measure where it coincides with the likelihood function–or that the marginal data density is nonzero at the observed data.

Under these assumptions, one can show that for every function $h(\theta)$, that if $E_p[|h(\theta)|^{1+\delta}] < \infty$ for some $\delta > 0$[5], then for every stage in the sampler, $n = 2, \ldots, N_\phi$, that the particle estimate of $h(\theta)$ converges to $E_{\pi_{n-1}}[h(\theta)]$ almost surely as the number of particles increases, provided that this convergence held at the

[5]This expectation is under the prior.

initial stage (i.e., under the prior). This holds after each of the correction, selection, and mutation steps.

Under the more restrictive moment conditions that $E_p[|h(\theta)|^{2+\delta}] < \infty$, one can show that a CLT also applies, again assuming that it first holds under the prior. Moreover, the variances associated with the CLT, when applied after each of the correction, selection, and mutation steps has a recursive form detailed in Chopin (2002). Unfortunately, the recursive form is not useful in the implementation of the algorithm. Instead, we rely on estimates computed across multiple independent runs of the algorithm as in Durham and Geweke (2012).

Algorithm 1 presents the generic algorithm for estimating Bayesian models, but does not specify the exact nature of the Metropolis-Hastings (MH) kernel used for particle mutation. As we show in Section 3, the form of the MH 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)). In general, blocking helps by reducing the dimensionality of the target density, making it easier to well approximated by the proposal density. When blocking, it is important to select parameters which are highly correlated into the same block; that is, to make the blocks "as independent as possible." The more highly correlated the blocks, the smaller the moves around the parameter space are likely to be (expand on this.) In our context, this limits the ability of the particles to mutate into particles representing higher posterior density. We find, somewhat surprisingly, that the use of random blocks has large effects on the accuracy of simulator, even for moderately correlated posteriors well approximated by normal distributions. In particular, we find that randomization avoids selecting bad block configurations.

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 just before the mutation phase (i.e., just after selection),

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

Let $\hat{\Sigma}_{b,n}$ denote the conditional variance (using a normal approximation) of the bth block of parameters.

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 $b \in B_n$.

A few key questions might be raised about the settings of the algorithm 1) How should one choose n_ϕ (or the schedule $\{\phi_i\}_{i=1}^{n_\phi}$)? and, 2) How many particles, n_{part}, should one use? 3) 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 Documenting the SMC Algorithm's Effectiveness

We document the effectiveness of the algorithm by demonstrating its ability to generate estimates of the MDD close to the truth in a setting in which we know the answer closed-form, reduced-form VARs.

3.1 The Model: VAR(3)

The three-variable, three-equation reduced-form VAR serves as our starting point,

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \Phi_3 y_{t-3} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma) \ . \tag{8}$$

Letting

$$\Phi = [\Phi_0', \Phi_1', \Phi_2', \Phi_3']' \tag{9}$$

$$\theta = \left[vec(\Phi)', vech(\Sigma)\right]' \ , \tag{10}$$

the VAR is parameterized by the vector θ. Reduced-form VARs with conjugate priors have a known closed-form expression for the marginal data density, which allows us to verify the ability of the algorithm to reproduce the correct results and to assess its effectiveness under different choices of algorithm hyperparameters. For

the data, we use the dataset used in the empirical example of Sims et al. (2008): observations from 1959Q1 to 2005Q4 of the output gap, GDP-deflator inflation, and the federal funds rate. A complete model specification requires a prior distribution for θ. We use a random-walk prior of the familiar multivariate-normal-inverse-Wishart family which is both conjugate for the parameterization in (8) and standard in the literature.

This VAR has a structural representation,

$$y_t' A_0 = y_{t-1}' A_1 + y_{t-2}' A_2 + y_{t-3}' A_3 + \varepsilon_t' \ , \quad u_t \sim \mathcal{N}(0, I). \tag{11}$$

The matrices $A_1, \ldots, A_n$ are often collected into a vector $A_+ = [A_1, A_2, A_3]$ in this example. As we have previously noted in most cases the econometrician is interested in $[A_0, A_+]$ rather than $[\Phi, \Sigma]$. However, there is not a unique mapping between these two objects. The practice of choosing this mapping is known as identifying the VAR. For this exercise we consider a popular one-to-one transformation of the parameters Φ and Σ, known as an exactly-identified structural VAR,

$$A_0 = [chol(\Sigma)]^{-1} \text{ and } A_+ = \Phi A_0 \ . \tag{12}$$

The operator *chol* defines the lower Cholesky factorization. The relationships in (12) define a one-to-one mapping (call it g) between the reduced form parameters and structural parameters: $g : (\Phi, \Sigma) \longrightarrow (A_0, A_+)$. Using this relationship, one can derive the prior density for the structural parameters and construct draws from the prior by applying g to draws from the reduced-form prior.[6] Thus, it is straightforward to apply the SMC sampler to the structural parameters.

As noted above, the SMC algorithm produces as a by-product an estimate for the

[6]See Appendix A for the pdf of (A_0, A_+) induced by applying g to the prior used in Section **??**.

marginal data density–we will use this to gauge the accuracy of the SMC algorithm. In both the reduced form and the structural case, the true log marginal data density will be the same, as the structural relationships don't affect the models description of the data.

To assess the algorithm's performance we run a Monte Carlo experiment using the sampler on both the reduced form and structural parameters. 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 hyperparameters. We run 20 Monte Carlo replications of the sampler, and examine the distribution of the estimates of $\ln p(Y)$.

The first row of Table 1 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 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.

3.2 Assessing the Tuning Parameters

To assess the importance of each of the algorithm parameters individually, we change each component holding the rest of the hyperparameters at the baseline case. This gives a rough "partial derivative" of each the parameter's contribution to the effectiveness of the algorithm. In particular, we 1) vary the number of particles to 1000 and 5000, 2) gauge the trade-off between the number of bridge distributions and intermediate Metropolis-Hastings steps while keeping the number of likelihood evaluations fixed but setting $(N_\phi, M) = (50, 10)$, 3) examine the role of blocking by varying the number of blocks and the mechanism for selecting them, 4) consider the use of the proposal distribution for the mutation steps as described in Herbst and Schorfheide (2014), and 5) finally, test the importance of the ϕ schedule by setting $\lambda = 1, 7$ and by determining the ϕ schedule via an algorithm based on the Hellinger distance between distributions, which we will discuss in more depth in Section **??**. 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 1 about here.

Table 1 show 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, rows 2 and 3 of Table 1, shows the effects of changing the quantity of particles. As one would expect, RMSEs fall as the number of particles increases.

The second set of deviations we consider, the fourth through eighth rows of Table 1, examines the role of the configurations of the blocking of the parameter vector for the mutation phase. First, we allow for two blocks instead of the baseline number, three. These two blocks are chosen either randomly or by slicing the

parameter vector θ, in half, deterministically setting the first half in one block and second half in the other. We call this a "fixed" block scheme. We also allow for a three block fixed scheme (dividing the parameter vector into thirds), and one we call the "natural" 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 A_+ and A_0, the RMSE of the log marginal data density is 3.95, more then ten times the size when randomizing the blocks.

The third set of deviations we consider, the ninth and tenth rows of Table 1, concern the number of ϕ stages and mutation step. The ninth row of Table 1 shows the results when the number of stages N_ϕ is reduced to 50 but the number of intermediate MH steps is increased to 10. 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 random blocks are fixed *within* the MH step, so that if a there even a few "bad" configurations of blocks can deteriorate performance despite a large number of MH steps. Under the next set of expirements we see more clearly the important role of

blocks.

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

Overall, the SMC algorithm works well across a wide range of values for the hyperparameters under both the reduced form and structural parameterizations of the VAR.

4 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 show that the SMC method overcomes three important shortcomings of current MCMC-based methods of inference in MS-VAR models: 1) MS-VAR posteriors routinely possess

multi-modality which may severely constrain the reliability of the currently-used MCMC algorithms for reasonable amounts of draws, 2) MHM-based methods of estimating MDDs struggle for reliability with reasonable amounts of posterior draws from models as highly-parameterized and multi-modal as MS-VARs, and 3) standard structural VAR priors place probability mass in a priori undesirable places in order to preserve invariance to sign changes, which affects inference about the best fitting model.

4.1 MS-VAR Posteriors: SMC vs. MCMC

In this section we compare the characterizations of MS-VAR posteriors that result from estimation with SMC and the MCMC algorithm proposed in Sims et al. (2008). We consider MS-VAR models of the form

$$y_t' A(s_t) = x_t' F(s_t) + \varepsilon_t' \Xi(s_t)^{-1} \tag{13}$$

$$diag(\Xi(s_t)) = [\xi_1(s_t), \ldots, \xi_n(s_t)] \tag{14}$$

$$\varepsilon_t | Y_{t-1}, S_t, \theta, q \sim \mathcal{N}(0_n, I_n) \tag{15}$$

$$p(s_t | S_{t-1}, Y_{t-1}, \theta, q) = q_{s_t, s_{t-1}} \tag{16}$$

$$q_{s_t=i, s_{t-1}=j} = q_{i,j} \quad \forall i, j, t, \tag{17}$$

where $\Xi(k)$ 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 $t-1$.

As a motivating example, we estimate a 3 equation system with 2 regimes for the parameters that determine the conditional mean of y_t (i.e. A and F) and 2 regimes for the shock standard deviations (i.e. Ξ). We use the shorthand 2m2v to denote such a model. We assume that the two processes evolve independently

and thus there are 4 joint states. We use the same data (output gap, inflation, and the nominal interest rate), prior hyperparameters ($\lambda_0 = 1.0$, $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.2$, $\lambda_4 = 0.1$, $\mu_5 = 1.0$, and $\mu_6 = 1.0$), and lag length ($p = 5$) as Sims et al. (2008) use in their empirical exercise. Our only point of departure from their models is that we do not restrict the parameters multiplying variable i in equation j to change only proportionally across regimes.

For most parameters, the output of the samplers is very similar. However, Figure 1 shows scatter plots of posterior draws from each sampler and each regime for the loadings on inflation and its first lag ($A_{2,3}$ and $F_{2,3}$) and the posteriors are very different across samplers. The portions of the posterior not reflected in the MCMC sampler have substantial density, and, while the reduced form dynamics implied by each are roughly similar, the impulse responses to a structural shock are markedly different.

When interpreting the scatter plots one must keep in mind the distinction between two types of multimodality, a distinction discussed in slightly different terms in Geweke (2007). The first, which we call normalization multimodality, is the *a priori* symmetric multimodality of a posterior that necessarily exists because of invariance to sign changes or the arbitrariness of state labeling. If a posterior exhibits only normalization multimodality then complete knowledge of one of the symmetric modes implies complete knowledge of the others. The second, which we call substantive multimodality, is multimodality present in a posterior such that knowledge of one mode does not imply complete knowledge of all other modes, i.e. the multimodality is not known to be symmetric *a priori*. The draws in Figure 1 have already been normalized, so the multimodality of their distribution is substantive multimodality.

The upshot of Figure 1 is that the two methods deliver substantially different

posterior inference. While it is hard to know which characterization is correct, we offer two pieces of evidence that the SMC characterization is more likely to be correct. In the first, we show comparably high values for $\ln p(Y)$ in regions not visited by MCMC as for those that it does visit. In the second, which we differ to the next section, we consider an example with substantive multimodality and known MDD and document that draws from the SMC algorithm accurately reflect the target distribution.

4.2 Comparing Accuracy of MDD Estimators for Multimodal Posteriors

In this section we assess the accuracy of SMC estimates of MDDs for distributions with substantive multimodality and compare the SMC estimator to relevant alternatives. We conduct our assessment in the context of a bimodal distribution with the key properties that 1) we know its integrating constant in closed-form, thus providing an absolute benchmark for success, 2) we can sample from it directly and then apply existing MDD estimation techniques, thus allowing a comparison of relative success across estimators, and 3) the distribution is similar to the SMC-estimated posterior of the MS-VAR, thus providing an example of empirical relevance. When estimating the MDD with draws sampled directly from the target distribution, we use the modified harmonic mean (MHM) method originally considered in Geweke (1989) and the version that Sims et al. (2008) adapt for better performance with fat-tailed distributions.[7] Since iid draws represent an upper bound on the usefulness of MCMC draws, 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 could ever hope to. At least in the setting of interest to us, we thus implicitly compare SMC to MCMC.

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

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

$$p(Y_i) = \int_\Theta p(\theta)p(Y_i|\theta) \ . \tag{18}$$

Let g be a function of the form

$$g(Y_1, Y_2|\theta) = \alpha p(Y_2)p(Y_1|\theta) + (1-\alpha)p(Y_1)p(Y_2|\theta)d\theta \ , \tag{19}$$

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

$$p(\theta|Y_1, Y_2) = \frac{p(\theta)g(Y_1, Y_2|\theta)}{\int_\Theta p(\theta)g(Y_1, Y_2|\theta)d\theta} \ . \tag{20}$$

We can rewrite the denominator as

$$\begin{aligned}
&\int_\Theta p(\theta)[\alpha_1 p(Y_2)p(Y_1|\theta) + (1-\alpha_1)p(Y_1)p(Y_2|\theta)]d\theta \\
&\qquad = \alpha_1 p(Y_2)\int_\Theta p(\theta)p(Y_1|\theta)d\theta + (1-\alpha_1)p(Y_1)\int_\Theta p(\theta)p(Y_2|\theta)d\theta \qquad (21)\\
&\qquad = p(Y_1)p(Y_2) \qquad (22)
\end{aligned}$$

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 prior $p(\theta)$ and likelihoods $p(Y_i|\theta)$.

Since we know their MDD in closed-form, we are able to 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 2 shows the results of our simulation for both an AR(1) and 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 incorporate information throughout the parameter space.

Secondly, and most central to our interests, SMC performs as well as SWZ, the best-performing MDD estimator, when we feed iid draws to SWZ. If researchers could always sample directly from the posteriors of multi-modal distributions, then our algorithm would offer little improvement over existing techniques. However, since posterior draws are typically simulated using MCMC algorithms, as with the MS-VAR, and 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.

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.[8] 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 50% of its true value.

Since we indirectly compare SMC to MCMC, one should note that the efficacy of MCMC methods for sampling from general multimodal distributions remains a topic of debate; a debate whose waters we do not care to wade into more deeply

[8] See Appendix C.2 for the details of the direct sampling algorithm.

than necessary. Having said that, 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 relying on local moves via RW-MH steps. In the worst-case scenario the MCMC sampler never leaves a neighborhood around the mode nearest to where the algorithm was initialized. The rows in Table 2 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. Celeux, Hurn, and Robert (2000) document degeneracies of this nature and general poor performance of posterior sampling with simple MCMC for mixture models. However, Geweke (2007) shows that there exist MCMC methods able to handle the normalization multimodality of mixture models and argues that Celeux et al. (2000) show little evidence of MCMC breakdowns resulting from substantive multimodality.[9] But unlike the examples in Geweke (2007), the results in the previous section indicate the presence of, and our example in the present section possesses by construction, substantive multimodality. We consider both SMC's documented effectiveness and the fact that leading researchers remain split on the effectiveness of MCMC to be compelling reasons to favor the SMC algorithm. We leave other researchers to discern whether or not MCMC can match SMC's success.

4.3 Prior Choice and How it Affects Model Selection

In this section we document some important shortcomings of the structural version of the Minnesota prior and show how SMC allows access to better performing priors.

The typical Minnesota prior for structural VARs, as described in Rubio-Ramrez,

[9]See also Frühwirth-Schnatter (2001) for details on MCMC methods for mixture models.

Waggoner, and Zha (2010), takes the form

$$a_0 \sim \mathcal{N}(0, I_n \otimes H_0) \tag{23}$$

$$a_+|a_0 \sim \mathcal{N}(vec(\bar{S}A_0), I_n \otimes H_+). \tag{24}$$

In addition to centering the dynamics of each series at a random walk, this prior has the desirable properties of invariance with respect to rotation of (A_0, A_+) by orthogonal matrices and accessibility of conditional posteriors that allow for Gibbs sampling via the algorithms in Waggoner and Zha (2003).[10] However, we show that it also has the undesirable property of poor data fit relative to the reduced-form implementation of the Minnesota prior and the difference in fit is severe enough to alter model selection when using the prior for MS-VARs.

We consider an alternative prior based on an inverse-Wishart density for $(AA')^{-1}$, standard in the analysis of reduced-form VARs, and which preserves the property of invariance with respect to orthogonal rotation.[11] We call this a reduced-form prior and Appendix A contains the details of its derivation. We find two subtle, but important, points about this approach worth mentioning. First, we can (and do) implement this prior while maintaining the restriction that structural shock variances change regimes according to a separate process from the VAR slope coefficients, i.e. we are not simply estimating a reduced-form MS-VAR. Second, We also derive the implied prior for $A_+|A_0$ under reduced-form prior beliefs. However, one could also use the prior for $A_+|A_0$ given in (24) if preferred and the prior would still be invariant to orthogonal rotations.

[10] As has been well-documented, the likelihood of the SVAR is invariant to orthogonal rotations of (A_0, A_+), so it seems reasonable that the prior density should be as well. See the discussion in Rubio-Ramrez et al. (2010) for details on this property of the structural version of the Minnesota prior.

[11] Sims and Zha (1998) also note that one could derive prior beliefs about A from Inverse-Wishart beliefs about Σ and suggest that doing so is probably superior to treating each column of A as independent.

Table **??** shows MDDs from estimating the parameters of numerous structural MS-VARs under each type of prior. Two key results emerge from this exercise. First, the choice of prior has a substantial impact on the fit of any particular model. The reduced-form prior gives a higher MDD for every model we estimate.[12] Second, the choice of the best-fitting model changes across choice of prior. Under the prior derived from the reduced-form, the best-fitting model becomes one with both changing shock variances and changing means.

To understand the reason for the general difference in model fit we examine point estimates of the implied reduced-form VAR parameters under each regime and compare them to the prior. Figure ?? shows the point estimates of a_{11} and the implied σ_{11} for a constant parameter VAR, along with its MLE value. Clearly the structural prior is more severely at odds with the data. One should note that the reduced-form prior is not some arbitrary construction to match the peaks of the likelihood, but rather is a standard conjugate prior in the reduced-form VAR literature.

5 Conclusion

We conclude here.

[12]The extent to which this is true would, of course, depend on the choice of hyperparameters for the reduced-form prior. In case of a constant-parameter VAR we know the MDD in closed-form when using the reduced-form prior allowing us to easily check sensitivity to choice of prior hyperparameters. The MDD under the reduced-form prior is very robustly greater than the MDD under the structural prior.

A Density of a Reduced-Form Prior Over Structural Parameters

A.1 Prior Density for A_0

Our reduced-form-based prior for A_0 is derived from the fact that, in the absence of changing shock variances, $(A_0 A_0')^{-1} = \Sigma$. 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.

$$f(\Sigma|\Psi,\nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu n/2}\Gamma_n(\nu/2)}|\Sigma|^{-(\nu+n+1)/2}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\Psi]\right\} \tag{25}$$

and then derive the implied density of A_0. Let the function g map a symmetric positive definite matrix into the inverse of the transpose of its Cholesky decomposition matrix. So

$$g(\Sigma) = \left(chol(\Sigma)^{-1}\right)' = A \tag{26}$$

$$g^{-1}(A) = (AA')^{-1} = \Sigma \tag{27}$$

We can then define a density directly over A as

$$h(A) = f\left(g^{-1}(A)\right)|J(A)| \tag{28}$$

$$J(A) = dg^{-1}(A). \tag{29}$$

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

$$\frac{d(AA')^{-1}}{dA} = \left(\frac{d(AA')^{-1}}{d(AA')}\right)\left(\frac{dAA'}{dA}\right) \tag{30}$$

$$\frac{d(AA')^{-1}}{d(AA')} = -\left((AA')'\right)^{-1} \otimes (AA')^{-1} \tag{31}$$

$$\frac{dAA'}{dA} = 2N_n(A \otimes I_n) \tag{32}$$

$$N_n = \frac{1}{2}\left(I_{n^2} + K_{nn}\right) \tag{33}$$

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

$$\frac{d(AA')^{-1}}{dA} = \left[-\left(AA'\right)^{-1} \otimes \left(AA'\right)^{-1}\right] \times 2N_n\left(A \otimes I_n\right) \tag{34}$$

$$= \left[-\left(AA'\right) \otimes \left(AA'\right)\right]^{-1} \times 2N_n\left(A \otimes I_n\right) \tag{35}$$

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

$$J(A) = D^{+}_{n,\Sigma}\left(\frac{d(AA')^{-1}}{dA}\right) M_A \tag{36}$$

where M_A is such that

$$M_A vech(A) = vec(A)\ . \tag{37}$$

Summing up we have

$$h(A) = f\left((AA')^{-1}\right)\left|\det\left(D^{+}_{n,\Sigma}\left(\frac{d(AA')^{-1}}{dA}\right) M_A\right)\right| \tag{38}$$

where $d(AA')^{-1}/dA$ is given in (35).

A.2 Prior Density for A_+

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

$$f(\Phi|\Sigma) = (2\pi)^{-kn/2}|\Sigma \otimes \Omega^{-1}|^{-1/2} \exp\left\{-\frac{1}{2} tr[\Sigma^{-1}(\Phi - \Phi^*)'\Omega(\Phi - \Phi^*)]\right\} . \quad (39)$$

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

$$g(\Phi|A) = \Phi A = A_+ \quad (40)$$

$$g^{-1}(A_+|A) = A_+ A^{-1} . \quad (41)$$

Hence the density of $A_+|A$ will be given by

$$h(A_+|A) = f(g^{-1}(A_+|A))\ |J| \quad (42)$$

where

$$J = \frac{dA_+ A^{-1}}{dA_+} = \frac{dI_m A_+ A^{-1}}{dA_+} \quad (43)$$

$$= (A^{-1})' \otimes I_m \quad (44)$$

Summing up we have

$$h(A_+|A) = f\left(A_+ A_0^{-1}\right) \left| \det\left(\frac{dA_+ A^{-1}}{dA_+}\right)\right| \quad (45)$$

and $d(A_+ A^{-1})/dA_+$ is given in (44).

B Reduced-form VAR Normalizing Constants of Bridge Distributions

Here we derive a closed-form expression for the value in the denominator of

$$p(\Phi, \Sigma|Y) = \frac{p(Y|\Phi, \Sigma)^{\phi} p(\Phi, \Sigma)}{\int p(Y|\Phi, \Sigma)^{\phi} p(\Phi, \Sigma) d\Phi d\Sigma} \tag{46}$$

when $p(Y|\Phi, \Sigma)$ is a VAR likelihood and $p(\Phi, \Sigma)$is a conjugate prior for the VAR likelihood. Defining

$$\widetilde{\Phi}(\phi) = \left(\phi X'X + \Omega\right)^{-1} \left(\phi X'Y + \Omega\Phi^*\right) \tag{47}$$

$$\widetilde{S}(\phi) = \left(\phi Y'Y + \Phi^{*'}\Omega\Phi^*\right) - \left(\phi X'Y + \Omega\Phi^*\right)' \left(\phi X'X + \Omega\right)^{-1} \left(\phi X'Y + \Omega\Phi^*\right) \tag{48}$$

one can write $p(Y|\Phi, \Sigma)^{\phi} p(\Phi, \Sigma)$ as

$$\Sigma|Y \sim \mathcal{IW}\left(\widetilde{S}(\phi) + \Psi,\ \phi T + d\right) \tag{49}$$

$$\text{vec}(\Phi)|Y, \Sigma \sim \mathcal{N}\left(\text{vec}\left(\widetilde{\Phi}(\phi)\right),\ \Sigma \otimes \left(\phi X'X + \Omega\right)^{-1}\right)\ . \tag{50}$$

And thus the integral and normalizing constant in the denominator of (46) is

$$\begin{aligned} C(\phi) = &(2\pi)^{-\phi T/2} \left(\frac{|\phi X'X + \Omega|^{-n/2}}{|\Omega|^{-n/2}}\right) \left(\frac{|\widetilde{S}(\phi) + \Psi|^{-(\phi T+d)/2}}{|\Psi|^{-d/2}}\right) \\ &\times \left(\frac{2^{(\phi T+d)n/2}}{2^{dn/2}}\right) \left(\frac{\Gamma_n\left((\phi T + d)/2\right)}{\Gamma_n\left(d/2\right)}\right)\ . \end{aligned} \tag{51}$$

Since $\phi = 1$ gives the MDD for a reduced-form VAR, (51) generalizes the expressions found elsewhere in the literature.

C Bimodal Example

C.1 Equivalence of (20) and Mixture of Posteriors

Consider the mixture of posteriors

$$\widetilde{p}(\theta|Y_1, Y_2) = \alpha_1 \left(\frac{p(\theta)p(Y_1|\theta)}{p(Y_1)} \right) + (1 - \alpha_1) \left(\frac{p(\theta)p(Y_2|\theta)}{p(Y_2)} \right) . \tag{52}$$

We can rewrite the mixture as

$$\widetilde{p}(\theta|Y_1, Y_2) = \frac{p(Y_2)\alpha_1 p(\theta)p(Y_1|\theta) + p(Y_1)(1 - \alpha_1)p(\theta)p(Y_2|\theta)}{p(Y_1)p(Y_2)} \tag{53}$$

$$= \frac{p(\theta)[\alpha_1 p(Y_2)p(Y_1|\theta) + (1 - \alpha_1)p(Y_1)p(Y_2|\theta)]}{p(Y_1)p(Y_2)} , \tag{54}$$

which matches (20).

C.2 Direct Sampler for Mixture of Posteriors

The direct sampling algorithm is as follows:

1. Draw latent state s according to

$$p(s = 1) = \alpha \tag{55}$$

$$p(s = 2) = 1 - \alpha \tag{56}$$

2. Draw $\Sigma|s, \Phi, Y$, which is a draw from $p(\Sigma|\Phi, Y_s)$ and under the conjugate prior this is simply $p(\Sigma|Y_s)$.

3. Draw $\Phi|s, \Sigma, Y$, which is a draw from $p(\Sigma|\Phi, Y_s)$.

D Additional Examples

In this section we present a number of additional examples of applications in which the use of SMC algorithms can either simplify or increase the accuracy of estimation.

D.1 Reduced-Form VAR with Steady-State Prior

Lastly, we examine the ability of the SMC algorithm to match MDD estimates for a reduced-form VAR with steady-state prior, in the style of Villani (2009). To facilitate clear comparison we use the data and prior of Fuentes-Albero and Melosi (2013), who estimate MDDs of variety of sizes of VARs with steady-state priors, up to a maximum size of six-variables and four lags.[13] The steady-state prior formulation of the VAR writes the dynamic system in terms of a mean-adjusted form

$$\tilde{y}_t = \tilde{x}_t \Phi + \varepsilon_t, \tag{57}$$

where $\tilde{x}_t$ is vector which includes lags of $\tilde{y}_t$ and (potentially) a constant term. The original series y_t is "demeaned" via

$$y_t = d_t' \Gamma + \tilde{y}_t. \tag{58}$$

Here d is a vector containing $[1, t, \ldots, t^l]$ and Γ is an $(l+1) \times n$ matrix of coefficients. The maximum order of the trend, l, is like p, specified *a priori* by the user. Setting $l = 0$ results in demeaning, $l = 1$ removes a linear trend, $l = 2$ removes a quadratic trend, and so on.

[13] Thank you to authors of that paper for providing us with their exact data and choice of hyperparameters.

D.2 Over-identified SVARs

Waggoner and Zha (2003) describes a method for estimating over-identified SVARs by Gibbs sampling which relies on the use of priors of a particular form and on generating random numbers from a special distribution. Our algorithm can estimate VARs of this type without the need to code a separate sampling algorithm. What's further, our leading example shows a situation in which a researcher's true prior beliefs about the world would not take the form required in order to use the Gibbs sampler in Waggoner and Zha (2003).

to be completed.

D.3 Sign-Restricted VARs

Following the work of Uhlig (2005), sign restrictions have become a popular method of identifying particular types of shocks, while remaining agnostic about the impulse responses under study. However Bayesian analysis of these models remains challenging owing to the fact that priors over impulse responses are analytically intractable unless they are implemented in the form of an indicator function. By treating our prior distribution a target distribution of interest, we can estimate models that generalize the sign restrictions framework to allow for prior densities over impulse response values directly, allowing for significantly greater flexibility and transparency of prior restrictions.

Let $p(\theta)$ be a density over the model parameters from which we can sample directly. Let $f(\theta)$ be some function of the model parameters, possibly the impulse response function of the VAR. Let $g(f(\theta))$ be a density over the values of f. Now consider a version of (5) in which the target distribution is the prior formed from

the product of p and g

$$\pi_n(\theta) = \frac{f_n(\theta)}{Z_n} = \frac{[g(f(\theta))]^{\phi_n} p(\theta)}{\int [g(f(\theta))]^{\phi_n} p(\theta) d\theta}, \quad n = 1, \ldots, N_\phi. \tag{59}$$

References

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

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

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

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

Del Moral, P., A. Doucet, and A. Jasra (2006): "Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society, Series B*, 68, 411–436.

Durham, G. and J. Geweke (2012): "Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments," *Unpublished Manuscript*.

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

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

Fuentes-Albero, C. and L. Melosi (2013): "Methods for Computing Marginal Data Densities From the Gibbs Output," *Journal of Econometrics*, 175, 132 – 141.

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," *Unpublished Manuscript, Universitè Libre de Bruxelles, European Central Bank, and Northwestern University*.

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. Tetlow (2013): "Financial stress and economic dynamics: The transmission of crises," *European Central Bank, Federal Reserve Board and IMF, Mimeo*.

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.

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

Rubio-Ramrez, 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.

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

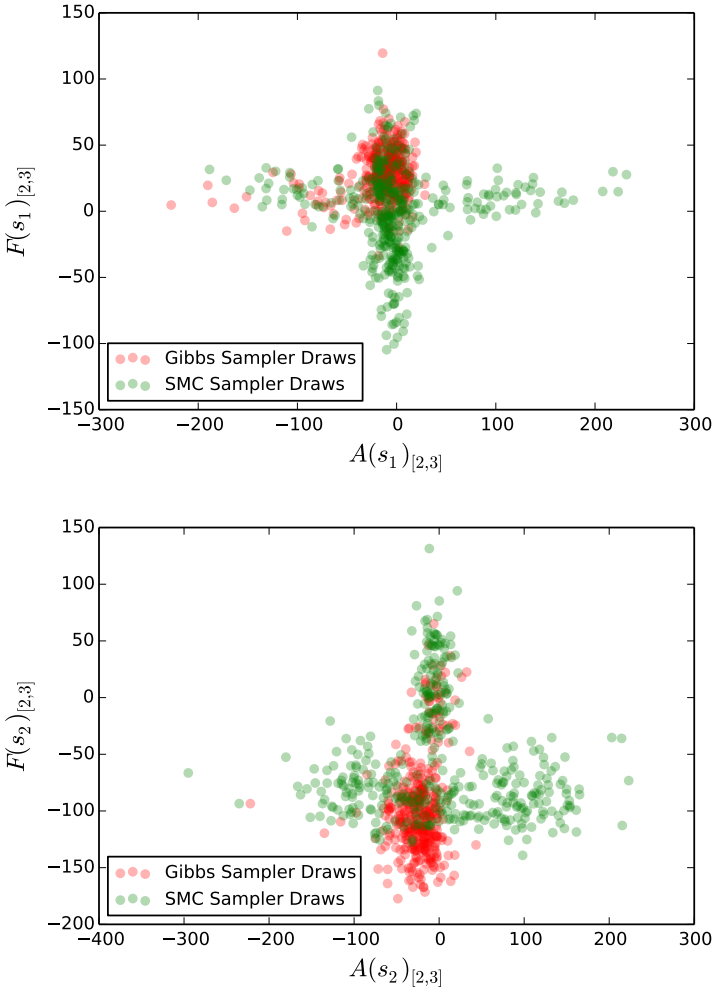
——— (2005): "What are the effects of monetary policy on output? Results from an agnostic identification procedure," *Journal of Monetary Economics*, 52, 381 – 419.

Villani, M. (2009): "Steady-state priors for vector autoregressions," *Journal of Applied Econometrics*, 24, 630–650.

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

E Figures

Figure 1: Draws from the (Normalized) Posterior of the MS-VAR



F Tables

Table 1: SMC Estimates of $\ln p(Y)$ for VAR(3): Effects of Algorithm Tuning Parameters

SMC Algorithm Parameters							VAR Paremeterization			
							Reduced Form		Structural	
N_{part}	N_{blocks}	Blocking	N_ϕ	N_{mut}	Σ_{prop}	λ	Avg Bias	RMSE	Avg Bias	RMSE
2000	**3**	**Random**	**500**	**1**	**Cond**	**4**	**0.01**	**0.29**	**0.01**	**0.21**
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
-	-	-	-	-	Un	-	0.01	1.37	−0.08	1.90
-	-	-	-	-	-	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 bold 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.

Table 2: AR(1): Estimates of $\ln p(Y)$ for Mixture Distribution Under Different Methods of Posterior Sampling and MDD Estimation

Model	Posterior Sampler	MDD Estimator	Avg Bias	RMSE
AR(1) $\ln p(Y) = 614.504$	SMC: $N_{part} = 2000$	SMC	-0.008	0.059
	Direct	SWZ	-0.000	0.038
		MHM	0.410	0.372
	Single Mode	SWZ	-0.702	0.021
		MHM	-0.699	0.012
VAR($n = 3, p = 5$) $\ln p(Y) = 1725.289$	SMC: $N_{part} = 2000$	SMC	0.090	0.457
	SMC: $N_{part} = 5000$	SMC	-0.018	0.368
	Direct	SWZ	0.123	0.255
		MHM	0.640	0.941
	Single Mode	SWZ	-0.817	0.019
		MHM	-0.842	0.074

Notes: Values are based on 50 replications. AR(1) algorithm settings: SMC sampler uses $\lambda = 4$, $n_\phi = 50$, $N_{blocks} = 1$; direct sampling and MCMC both use 10,000 draws; and MHM estimate uses $p = 0.9$ for truncation. VAR algorithm settings: SMC sampler uses $\lambda = 4$, $n_\phi = 500$, $N_{blocks} = 8$; direct sampling and MCMC both use 10,000 draws; and MHM estimate uses $p = 0.9$ for truncation.

Table 3: SMC estimates of $\log(p(Y))$ for Markov-switching VAR($n = 3, p = 5$) models.

Model		Prior	
		Reduced Form	Structural
1v	1m	1772.75	1760.31
2v	1m		1869.73
3v	-	1882.54	1873.03
		0.13	*0.13*
4v	-	1879.49	1872.63
5v	-	1880.07	1871.87
6v	-	1874.74	
1v	2m	1844.04	
2v	2m	1878.34	1869.63
3v	2m	1880.67	1869.87
4v	2m	1879.89	

Notes: $\log(p(Y))$ estimates are means from 5 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.