2-16-2017

Interweaving Markov Chain Monte Carlo Strategies for Efficient Estimation of Dynamic Linear Models

Matthew Simpson
University of Missouri

Jarad Niemi
Iowa State University, niemi@iastate.edu

Vivekananda Roy
Iowa State University, vroy@iastate.edu

Interweaving Markov Chain Monte Carlo Strategies for Efficient Estimation of Dynamic Linear Models

Matthew Simpson
Department of Statistics, University of Missouri–Columbia

Jarad Niemi and Vivekananda Roy
Department of Statistics, Iowa State University

October 2, 2015

Abstract

In dynamic linear models (DLMs) with unknown fixed parameters, a standard Markov chain Monte Carlo (MCMC) sampling strategy is to alternate sampling of latent states conditional on fixed parameters and sampling of fixed parameters conditional on latent states. In some regions of the parameter space, this standard data augmentation (DA) algorithm can be inefficient. To improve efficiency, we apply the interweaving strategies of Yu and Meng (2011) to DLMs. For this, we introduce three novel alternative DAs for DLMs: the scaled errors, wrongly-scaled errors, and wrongly-scaled disturbances. With the latent states and the less well known scaled disturbances, this yields five unique DAs to employ in MCMC algorithms. Each DA implies a unique MCMC sampling strategy and they can be combined into interweaving and alternating strategies that improve MCMC efficiency. We assess these strategies using the local level model and demonstrate that several strategies improve efficiency relative to the standard approach and the most efficient strategy interweaves the scaled errors and scaled disturbances. Supplementary materials are available for this article online.

Key Words: Ancillary augmentation; Centered parameterization; Data augmentation; Non-centered parameterization; Reparameterization; Sufficient augmentation; Time series; State-space model

1. INTRODUCTION

The Data Augmentation (DA) algorithm of Tanner and Wong (1987) and the closely related Expectation Maximization (EM) algorithm of Dempster et al. (1977) have become widely used strategies for computing posterior distributions and maximum likelihood estimates. While useful, DA and EM algorithms often suffer from slow convergence and a large literature has grown up around various possible improvements to both algorithms (Meng and Van Dyk, 1997, 1999; Liu and Wu, 1999; Hobert and Marchev, 2008; Yu and Meng, 2011), though much of the work on constructing improved algorithms has focused on hierarchical models (Gelfand et al., 1995; Roberts and Sahu, 1997; Meng and Van Dyk, 1998; Van Dyk and Meng, 2001; Bernardo et al., 2003; Papaspiliopoulos et al., 2007; Papaspiliopoulos and Roberts, 2008). Despite some similarities with some hierarchical models, relatively little attention has been paid to time series models, exceptions include Pitt and Shephard (1999); Frühwirth-Schnatter and Sögner (2003); Frühwirth-Schnatter and Wagner (2006) in the DA literature and Van Dyk and Tang (2003) in the EM literature.

We seek to improve DA schemes in dynamic linear models (DLMs), i.e. linear Gaussian state-space models. The standard DA scheme uses the latent states and alternates between drawing from the full conditionals of the latent states and the model parameters (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994). The existing literature on improving DA algorithms in time series models tends to focus on non-Gaussian state-space models — particularly the stochastic volatility model and derivative models (Shephard, 1996; Frühwirth-Schnatter and Sögner, 2003; Roberts et al., 2004; Bos and Shephard, 2006; Strick-

land et al., 2008; Frühwirth-Schnatter and Sögner, 2008; Kastner and Frühwirth-Schnatter, 2014), but a few work with the class of DLMs we consider (Frühwirth-Schnatter, 2004). One recent development in the DA literature is an "interweaving" strategy for using two separate DAs in a single algorithm (Yu and Meng, 2011). This strategy draws on the strengths of both underlying DA algorithms in order to construct a Markov chain Monte Carlo (MCMC) algorithm which is at least as efficient as the worst of the two DA algorithms and, in some cases, is a dramatic improvement over the best. We implement interweaving algorithms in a general class of DLMs and, in order to do so, we introduce several new DAs for this class of models. We also show that no *practical* sufficient augmentation exists for the DLM, which restricts the interweaving algorithms we can construct. Using the local level model, we assess the relative performance of the various MCMC strategies.

The rest of the paper is organized as follows. In Section 2, we review the relevant DA literature and, in Section 3, we introduce the dynamic linear model and discuss the class of DLMs we consider. In Section 4, we introduce DAs for our class of DLMs and show that any sufficient augmentation is likely to be difficult to use. In Section 5, we discuss the various MCMC strategies available for the DLM while Section 6 applies these algorithms to the local level model. Finally, in Section 7, we interpret these results and suggests directions for further research.

2. VARIATIONS OF DATA AUGMENTATION

Let $p(\phi|y)$ be a probability density, e.g. the posterior distribution of some parameter ϕ given data y. A DA adds a parameter θ with joint distribution $p(\phi, \theta|y)$ such that $\int_\Theta p(\phi, \theta|y)d\theta =$

$p(\phi|y)$ and the associated DA algorithm is a Gibbs sampler for (ϕ, θ). In this DA algorithm, the next draw of ϕ is obtained from the current draw, k, as follows (implicitly conditioning on y):

Algorithm: DA. *Data Augmentation*

$$[\theta|\phi^{(k)}] \rightarrow [\phi^{(k+1)}|\theta]$$

where $[\theta|\phi^{(k)}]$ means a draw of θ from $p(\theta|\phi^{(k)}, y)$ and analogously for $[\phi^{(k+1)}|\theta]$. Though θ may be scientifically interesting, here we view its addition as a computational construct and thus focus our attention on ϕ.

2.1 Alternating and Interweaving

One well known method of improving the efficiency of MCMC samplers is judiciously choosing the DA, an example of reparameterization (see Papaspiliopoulos et al. (2007) and references therein). Often the DA algorithms based on two separate DAs will be efficient in separate regions of the parameter space. This property suggests combining the two such DA algorithms to construct an improved sampler. One intuitive approach is to alternate between the two augmentations within a Gibbs sampler (Papaspiliopoulos et al., 2007). With two DAs θ and γ, the alternating algorithm for sampling from $p(\phi|y)$ is:

Algorithm: Alt. *Alternating Algorithm*

$$[\theta|\phi^{(k)}] \rightarrow [\phi|\theta] \rightarrow [\gamma|\phi] \rightarrow [\phi^{(k+1)}|\gamma].$$

One iteration of the alternating algorithm consists of an iteration of the DA algorithm based on θ followed by one iteration of the DA algorithm based on γ.

Another option is to *interweave* the two DAs together (Yu and Meng, 2011). A global interweaving strategy (GIS) using θ and γ as DAs is:

Algorithm: GIS. *Global Interweaving Strategy*

$$[\theta|\phi^{(k)}] \rightarrow [\gamma|\theta] \rightarrow [\phi^{(k+1)}|\gamma].$$

The GIS algorithm obtains the next iteration of the parameter ϕ in three steps: 1) draw θ conditional on $\phi^{(k)}$, 2) draw γ conditional on θ, and 3) draw $\phi^{(k+1)}$ conditional on γ. The second step of the GIS algorithm is often accomplished by sampling $\phi|\theta$ and then $\gamma|\theta, \phi$. This expanded GIS algorithm is:

Algorithm: eGIS. *Expanded GIS*

$$[\theta|\phi^{(k)}] \rightarrow [\phi|\theta] \rightarrow [\gamma|\theta, \phi] \rightarrow [\phi^{(k+1)}|\gamma].$$

In addition, γ and θ are often, but not always, one-to-one transformations of each other conditional on (ϕ, y), i.e. $\gamma = M(\theta; \phi, y)$ where $M(.; \phi, y)$ is one-to-one, and thus $[\gamma|\theta, \phi]$ is deterministic. The key difference between the GIS and Alt algorithms is in step 3: instead of drawing from $p(\gamma|\phi, y)$, the GIS algorithm draws from $p(\gamma|\theta, \phi, y)$. The interweaving algorithm connects the two DAs together while the alternating algorithm keeps them separate. The weaker the dependence between the two DAs in their joint posterior, the weaker the dependence in the GIS chain and the more efficient the GIS algorithm (Yu and Meng, 2011). In fact with *a posteriori* independent DAs, the GIS algorithm obtains iid draws from ϕ's posterior. Thus, we can control the dependence by choosing the two DAs carefully.

If θ is a DA such that $p(y|\theta, \phi) = p(y|\theta)$, then θ is a *sufficient augmentation* (SA) for ϕ, while if θ is a DA such that $p(\theta|\phi) = p(\theta)$, then θ is an *ancillary augmentation* (AA) for ϕ (Yu and Meng, 2011). In the literature, an SA is sometimes called a centered augmentation or centered parameterization while an AA is sometimes called a non-centered augmentation or non-centered parameterization. A GIS approach where one of the DAs is an SA and the other

is an AA is called an ancillary sufficient interweaving strategy (ASIS). Like Yu and Meng (2011) we prefer the SA and AA terminology because it suggests a connection with Basu's theorem (Basu, 1955). Under the theorem's conditions an SA and an AA are independent conditional on the model parameter, which suggests that the dependence between the two DAs will be limited in the posterior. Yu and Meng (2011) show that when the group structure required to define the optimal PX-DA algorithm (Liu and Wu, 1999) is present, ASIS and optimal PX-DA are equivalent.

In addition to GIS, it is possible to define a componentwise interweaving strategy (CIS) that interweaves within specific steps of a Gibbs sampler as well. A CIS algorithm for $\phi = (\phi_1, \phi_2)$ essentially employs interweaving for each block of ϕ separately, e.g.

Algorithm: CIS. *Componentwise Interweaving Strategy*

$$[\theta_1|\phi_1^{(k)}, \phi_2^{(k)}] \quad \rightarrow [\gamma_1|\phi_2^{(k)}, \theta_1] \quad \rightarrow [\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_1] \quad \rightarrow$$

$$[\theta_2|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_1] \rightarrow [\gamma_2|\phi_1^{(k+1)}, \theta_2] \rightarrow [\phi_2^{(k+1)}|\phi_1^{(k+1)}, \gamma_2]$$

where θ_i and γ_i are distinct data augmentations for $i = 1, 2$, but potentially $\gamma_1 = \theta_2$ or $\gamma_2 = \theta_1$. The first row draws ϕ_1 conditional on ϕ_2 using interweaving in a Gibbs step, while the second row does the same for ϕ_2 conditional on ϕ_1. The algorithm can easily be extended to additional blocks within ϕ. CIS is attractive because it is often easier to find an AA–SA pair of DAs for ϕ_1 conditional on ϕ_2 and another pair for ϕ_2 conditional on ϕ_1 than it is to find an AA–SA pair for $\phi = (\phi_1, \phi_2)$ jointly.

3. DYNAMIC LINEAR MODELS

The general dynamic linear model is well studied (West and Harrison, 1999; Petris et al., 2009; Prado and West, 2010) and is defined as

$$y_t = F_t\theta_t + v_t \qquad v_t \overset{ind}{\sim} N_k(0, V_t) \qquad \text{(observation equation)}$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \overset{ind}{\sim} N_p(0, W_t) \qquad \text{(system equation)}$$

where $N_d(\mu, \Sigma)$ is a d-dimensional multivariate normal distribution with mean μ and covariance Σ. The observation errors, $v_{1:T} \equiv (v_1', v_2', \ldots, v_T')'$, and the system disturbances, $w_{1:T} \equiv (w_1', w_2', \ldots, w_T')'$ are independent. The observed data are $y \equiv y_{1:T} \equiv (y_1', y_2', \cdots, y_T')'$ while the latent states are $\theta \equiv \theta_{0:T} \equiv (\theta_0', \theta_1', \cdots, \theta_T')'$. For each $t = 1, 2, \cdots, T$, F_t is a $k \times p$ matrix and G_t is a $p \times p$ matrix.

The class of DLMs we focus on sets $V_t = V$ and $W_t = W$ and treats F_t and G_t as known for all t. Our results can be extended to time varying V_t or W_t or to when F_t or G_t depend on unknown parameters, but we ignore those cases for simplicity. So $\phi = (V, W)$ is the parameter and we can write the model as

$$y_t|\theta, V, W \overset{ind}{\sim} N_k(F_t\theta_t, V) \qquad \theta_t|\theta_{0:t-1}, V, W \sim N_p(G_t\theta_{t-1}, W) \tag{1}$$

for $t = 1, 2, \cdots T$. We assume the conditionally conjugate priors: θ_0, V, and W independent with $\theta_0 \sim N_p(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ and $W \sim IW(\Lambda_W, \lambda_W)$ where m_0, C_0, Λ_V, λ_V, Λ_W, and λ_W are known hyperparameters and $IW(\Lambda, \lambda)$ denotes the inverse Wishart distribution with degrees of freedom λ and positive definite scale matrix Λ.

The latent states, θ, can be integrated out to obtain the marginal model for the y:

$$y|V, W \overset{ind}{\sim} N_{Tk}(D\tilde{m}, \tilde{V} + \tilde{W} + \tilde{C}). \tag{2}$$

where $\tilde{V} = I_T \otimes V$ where $\otimes$ is the Kronecker product, D is block diagonal with elements $D_1, \ldots, D_T$,

$$\tilde{W}_{Tk\times Tk} = \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix}' W \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix},$$
$$\tilde{C}_{Tk\times Tk} = \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix}' C_0 \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix},$$

$\tilde{m}_{Tp\times 1} = (m_0', m_0', \cdots m_0')'$. D_t, K_t, and H_t are functions of the F_t's and G_t's and their definitions and derivations are provided in Appendix A.

4. AUGMENTING THE DLM

In order to construct an ASIS algorithm, we need to find an SA and an AA for the DLM. Papaspiliopoulos et al. (2007) note that typically the standard DA is an SA for ϕ and an AA can be constructed by creating a pivotal quantity. However, the standard DA for a DLM, θ, is neither an SA nor an AA. In equation (1), V is in the observation equation so that θ is not an SA for (V, W). Similarly W is in the system equation so that θ is also not an AA for (V, W). So to find an SA we need to somehow move V from the observation equation to the system equation. The following lemma suggests that this will be difficult.

Lemma 1. *Suppose η is an SA for the DLM such that conditional on ϕ, η and y are jointly normally distributed, that is*

$$\begin{bmatrix} \eta \\ y \end{bmatrix} \Bigg| \phi \sim N\left(\begin{bmatrix} \alpha_\eta \\ D\tilde{m} \end{bmatrix}, \begin{bmatrix} \Omega_\eta & \Omega_{y,\eta}' \\ \Omega_{y,\eta} & \tilde{V} + \tilde{W} + \tilde{C} \end{bmatrix} \right).$$

Let $A = \Omega_{y,\eta}'\Omega_\eta^{-1}$ and $\Sigma = \tilde{V} + \tilde{W} + \tilde{C} - A\Omega_\eta A'$. Then A, Σ, and α_η are constants with respect to ϕ and if $A'A$ is invertible, then

$$
\begin{aligned}
p(\phi|\eta, y) \propto & p(y|\eta, \phi)p(\eta|\phi)p(\phi) = p(y|\eta)p(\eta|\phi)p(\phi) \propto p(\eta|\phi)p(\phi) \\
= & p(\phi)|(A'A)^{-1}A'(\tilde{V} + \tilde{W} + \tilde{C} - \Sigma)A(A'A)^{-1}|^{-1/2} \\
& \times \exp\left[-\frac{1}{2}(\eta - \alpha_\eta)'[(A'A)^{-1}A'(\tilde{V} + \tilde{W} + \tilde{C} - \Sigma)A(A'A)^{-1}]^{-1}(\eta - \alpha_\eta)\right].
\end{aligned}
$$

The proof of this lemma is in Appendix B. The posterior density we wish to sample from comes from equation (2) and is similar to $p(\phi|\eta, y)$ except less complicated. So what this lemma shows is that in order to use an SA in a GIS algorithm, we must sample from a density that is as hard to sample from as our target posterior. Thus if we cannot draw from the target posterior, then we cannot draw from the full conditional distribution in an SA.

While we cannot find an SA for the DLM, there are several DAs available for the construction of various MCMC algorithms. We now introduce four DAs in addition to the latent states, three of them novel.

4.1 The scaled disturbances

The *scaled disturbances* (SDs) are constructed by creating a pivotal quantity using the system disturbances (Frühwirth-Schnatter, 2004). Let L_W denote the Cholesky decomposition of W, i.e. the lower triangular matrix L_W such that $L_W L_W' = W$. Then we will define the SDs, $\gamma \equiv \gamma_{0:T} \equiv (\gamma_0', \gamma_1', \cdots, \gamma_T')'$, by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$. There are actually $p!$ different versions of the SDs depending on how we order the elements of θ_t but we utilize the natural ordering. The reverse transformation is defined recursively by $\theta_0(\gamma, L_W) = \gamma_0$ and $\theta_t(\gamma, L_W) = L_W\gamma_t + G_t\theta_{t-1}(\gamma, L_W)$ for $t = 1, 2, \cdots, T$. Using the SDs, the model is

$$
y_t|\gamma, V, W \overset{ind}{\sim} N_k\left(F_t\theta_t(\gamma, L_W), V\right), \qquad \gamma_t \overset{iid}{\sim} N_p(0, I_p)
$$

for $t = 1, 2, \cdots, T$ where I_p is the $p \times p$ identity matrix. Since neither V nor W are in the system equation, the SDs are an AA for (V, W).

4.2 The scaled errors

The SDs immediately suggest our first novel augmentation, called the *scaled errors* (SEs), i.e. $v_t = y_t - F_t\theta_t$ scaled by V. Let L_V denote the Cholesky decomposition of V so that $L_V L_V' = V$. We define the SEs as $\psi_t = L_V^{-1}(y_t - F_t\theta_t)$ for $t = 1, 2, \cdots, T$ and $\psi_0 = \theta_0$, although here are $k!$ versions of the SEs depending on how y_t is ordered.

Assume F_t is invertible for all t; see Appendix F of the supplementary materials and Simpson (2015) for examples of how to relax this restriction. Then $\theta_t = F_t^{-1}(y_t - L_V\psi_t)$ for $t = 1, 2, \cdots, T$ while $\theta_0 = \psi_0$. Define $\mu_1 = L_V\psi_1 + F_1G_1\psi_0$ and $\mu_t = L_V\psi_t + F_tG_tF_{t-1}^{-1}(y_{t-1} - L_V\psi_{t-1})$ for $t = 2, 3, \cdots, T$. Then the we can write the model as

$$y_t|V, W, \psi, y_{1:t-1} \sim N_p(\mu_t, F_tWF_t'), \qquad \psi_t \overset{iid}{\sim} N_p(0, I_k)$$

for $t = 1, 2, \cdots, T$ where I_k is the $k \times k$ identity matrix. Since neither V nor W are in the system equation, the SEs are an AA for (V, W). However, both V and W are in the observation equation so that ψ is not an SA for $V|W$ nor for $W|V$.

4.3 The "wrongly-scaled" DAs

Two more novel augmentations can be obtained by scaling the SD and SE by the "wrong" variance so long as F_t is square ($k = p$). Define $\tilde{\gamma}_t = L_V^{-1}(\theta_t - G_t\theta_{t-1})$ and $\tilde{\psi}_t = L_W^{-1}(y_t - \theta_t)$ for $t = 1, 2, \cdots, T$ and $\tilde{\psi}_0 = \tilde{\gamma}_0 = \theta_0$. We call $\tilde{\gamma} \equiv \tilde{\gamma}_{0:T}$ the *wrongly-scaled disturbances* (WSDs) and $\tilde{\psi} \equiv \tilde{\psi}_{0:T}$ the *wrongly-scaled errors* (WSEs). In terms of $\tilde{\gamma}$ the model is

$$y_t|\tilde{\gamma}, V, W \overset{ind}{\sim} N_p\left(F_t\theta_t(\tilde{\gamma}, L_V), V\right), \qquad \tilde{\gamma}_t \overset{ind}{\sim} N_p(0, L_V^{-1}W(L_V^{-1})')$$

for $t = 1, 2, \cdots, T$ where $\theta_t(\tilde{\gamma}, L_V)$ denotes the transformation from $\tilde{\gamma}$ to θ defined by the WSDs. Since L_V is the Cholesky decomposition of V, the observation equation does not contain W, so $\tilde{\gamma}$ is an SA for $W|V$. Since W and L_V are both in the system equation, $\tilde{\gamma}$ is not an AA for $V|W$ nor for $W|V$.

Similarly, we can write the model in terms of $\tilde{\psi}$ as

$$y_t|V, W, \tilde{\psi}, y_{1:t-1} \sim N_p(\tilde{\mu}_t, F_tWF_t'), \qquad \tilde{\psi}_t \overset{iid}{\sim} N_p(0, L_W^{-1}V(L_W^{-1})')$$

for $t = 1, 2, \cdots, T$ where we define $\tilde{\mu}_1 = L_W\tilde{\psi}_1 - F_1G_1\tilde{\psi}_0$ and for $t = 2, 3, \cdots, T$ $\tilde{\mu}_t = L_W\tilde{\psi}_t - F_tG_tF_{t-1}^{-1}(y_{t-1} - L_W\tilde{\psi}_{t-1})$. Since $\tilde{\mu}_t$ only depends on W and not on V, V is absent from the observation equation and thus $\tilde{\psi}$ is an SA for $V|W$. Once again, since both W and V are in the system equation $\tilde{\psi}$ is not an AA for either V or W.

5. MCMC STRATEGIES FOR THE DLM

This section briefly discusses how to construct various MCMC algorithms for approximating the posterior distribution of the DLM. We focus on *what* to do, not *why*, though derivations of the relevant full conditional distributions are available in Appendix D. We occasionally come across a full conditional density that is difficult to sample from — the details about why this happens and how to overcome it are in Appendices G and H.

5.1 Base algorithms

Using any of the DAs introduced in Section 4, we can construct several DA algorithms which we call *base algorithms*. We will call the standard DA algorithm using θ the *state sampler*.

In order to construct this sampler, we need to draw from two densities: $p(\theta|V, W, y)$ and $p(V, W|\theta, y)$. The latter has V and W independent with

$$V|\theta, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right), \qquad W|\theta, y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right),$$

where $v_t = y_t - F_t\theta_t$, and $w_t = \theta_t - G_t\theta_{t-1}$.

The density $p(\theta|V, W, y)$ is multivariate normal and any algorithm to draw from it is called a simulation smoother. FFBS is the most commonly used smoother and it uses the Kalman filter (Frühwirth-Schnatter, 1994; Carter and Kohn, 1994), but there are other options. We use the mixed Cholesky factor algorithm (MCFA) to draw θ (McCausland et al., 2011; Kastner and Frühwirth-Schnatter, 2014). The details of this algorithm are included in Appendix E.

Putting the pieces together, the state sampler is the following DA algorithm:

Algorithm: State. *State Sampler*

$$[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}, W^{(k+1)}|\theta]$$

where the first step uses the MCFA and the second step is independent inverse Wishart draws. It is well known that this Markov chain can mix poorly in some regions of the parameter space, e.g. Frühwirth-Schnatter (2004) and Section 6.

Next, we can use γ in order to construct a DA algorithm called the *scaled disturbance sampler* or SD sampler. In the smoothing step, we need to obtain a draw from $p(\gamma|V, W, y)$. This density is also Gaussian but has a more complex precision matrix. Thus, we use the MCFA to sample $\theta \sim p(\theta|V, W, y)$ and transform from θ to γ. The density $p(V, W|\gamma, y)$ is rather complicated and does not appear easy to draw from, so we draw V and W in separate

Gibbs steps. As a result Algorithm SD has three steps.

Algorithm: SD. *Scaled Disturbance Sampler*

$$[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}|W^{(k)}, \theta] \rightarrow [\gamma|V^{(k+1)}, W^{(k)}, \theta] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma]$$

It is easy to show that $V|W, \gamma, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right)$ where $v_t = y_t - F_t\theta_t$ and θ_t is a function of γ and W. So the first and second steps are the same draws as in Algorithm State while the third step is a transformation from θ to γ. The last step is difficult due to the complexity of $p(W|V, \gamma, y)$, but it can be sampled from with tolerable efficiency in the local level model. In Appendix G of the supplementary materials, we have more detail as well as a rejection sampling algorithm for when W is a scalar. When W is a matrix it is not clear whether drawing from $p(W|V, \gamma, y)$ can be accomplished efficiently.

The DA algorithm based on the SEs is called the *scaled error sampler* or SE sampler (Algorithm SE) and is similar to the SD sampler with a couple of key differences. First, the simulation smoothing step in the SE sampler can be accomplished directly with the MCFA because the precision matrix of the conditional posterior of ψ retains the necessary tridiagonal structure. Second, the full conditional distribution of W is the familiar inverse Wishart density and the full conditional of V is the complicated density. The density of $V|W, \psi, y$ is in the same class as that of $W|V, \gamma, y$. In fact there is a strong symmetry here — the joint conditional posterior of (V, W) given γ is from the same family of densities as that of (W, V) given ψ so that V and W essentially switch places.

Algorithm: SE. *Scaled Error Sampler*

$$[\psi|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}|W^{(k)}, \psi] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi]$$

The third step is the same inverse Wishart draw for W as in Algorithm State. The second step contains the difficult draw.

We can also construct DA algorithms based on the WSDs and the WSEs — the *wrongly-scaled disturbance sampler* and the *wrongly-scaled error sampler*. In Section 6, we show that these samplers perform poorly, so their construction is left to Appendix C. The WSDs and WSEs will ultimately be helpful in the construction of certain CIS algorithms in Section 5.4.

5.2 Alternating algorithms

Using the full conditionals defined in Section 5.1, we can construct several alternating algorithms based on any two of the DAs using Algorithm Alt on page 4. For example, the *State-SD alternating sampler* obtains the $k+1$'st iteration of (V, W) from the k'th as follows:

$$[\theta|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}, W^{(k+0.5)}|\theta] \to$$

$$[\gamma|V^{(k+0.5)}, W^{(k+0.5)}] \to [V^{(k+1)}|W^{(k+0.5)}, \gamma] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

The first line is an iteration of the state sampler while the second line is an iteration of the SD sampler. No work is necessary to link up the two iterations. Each other alternating algorithm is analogous — including algorithms using three or more DAs.

5.3 GIS algorithms

We can use the various DAs of Section 4 to construct GIS algorithms as well, based on Algorithm eGIS on page 5. For example, the *State-SD GIS sampler* is:

$$[\theta|V^{(k)}, W^{(k)}] \to [W^{(k+0.5)}, V^{(k+0.5)}|\theta] \to$$

$$[\gamma|V^{(k+0.5)}, W^{(k+0.5)}, \theta] \to [V^{(k+1)}|W^{(k+0.5)}, \gamma] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

In the first step of the second line, we transform θ to γ using the equations in Section 4.1 which do not depend on V.

There are often improvements that can be made simply by thinking clearly about what the GIS algorithm is doing. For example in the above version of the State-SD GIS sampler, the draw of V in step two of line one and the draw of V in step two of line two are redundant — they come from the same distribution and only the last one is ever used in later steps. The resulting State-SD GIS sampler is as follows:

Algorithm: State-SD GIS. *State-Scaled Disturbance GIS Sampler*

$$[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}, W^{(k+0.5)}|\theta] \rightarrow [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].$$

The first two steps are both steps of Algorithm State, the third step simply transforms from θ to γ, and the final step is the difficult draw from Algorithm SD.

Other GIS algorithms are analogous and we can construct them with three or more DAs without complication.

5.4 CIS algorithms

Next we consider CIS algorithms which have the form of Algorithm CIS on page 6. The advantage of using CIS is that it is sometimes possible to find an AA-SA pair of DAs for each block of the parameter vector even when no such pair of DAs exist for the entire vector. From Section 4, we know that the SDs and the WSDs form an AA-SA pair for $W|V$ while the SEs and the WSEs form an AA-SA pair for $V|W$. A CIS sampler based on these AA-SA

pairs obtains $(V^{(k+1)}, W^{(k+1)})$ from $(V^{(k)}, W^{(k)})$ as follows:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\tilde{\psi}|V^{(k+0.5)}, W^{(k)}, \psi] \to [V^{(k+1)}|W^{(k)}, \tilde{\psi}] \to$$

$$[\tilde{\gamma}|V^{(k+1)}, W^{(k)}, \tilde{\psi}] \to [W^{(k+0.5)}|V^{(k+1)}, \tilde{\gamma}] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \tilde{\gamma}] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

The first line is essentially a Gibbs step for drawing V that interweaves between ψ and $\tilde{\psi}$ while the second line is essentially a Gibbs step for drawing W that interweaves between γ and $\tilde{\gamma}$. In the second line we use the SA before the AA in order to minimize the number of transformations we have to make in every iteration.

Despite the fact that θ, the standard augmentation, is not an SA for $V|W$, each time the WSDs or WSEs appears in the CIS sampler it would make no difference if θ was used instead because $p(V|W, \tilde{\psi}, y) = p(V|W, \theta, y)$ and $p(W|V, \tilde{\gamma}, y) = p(W|V, \theta, y)$. Using this we obtain a slightly different version of the CIS sampler:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\psi|V^{(k+0.5)}, W^{(k)}, \theta] \to [V^{(k+1)}|W^{(k)}, \theta] \to$$

$$[W^{(k+0.5)}|V^{(k+1)}, \theta] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

We show in Appendix I that this algorithm is equivalent to SD-SE GIS in a certain sense so that we expect the mixing and convergence properties of the two algorithms to be very similar, and we confirm this in the local level model in Section 6.

In our original definition of the CIS sampler for the DLM we used the SDs as the AA for W and the SEs as the AA for V. We could have reversed this or used the same AA for both V and W since both the SEs and SDs are AAs for (V, W), or we could have used θ as the AA for V. In each of these cases, the resulting algorithm would reduce to either the state sampler or a *partial CIS* algorithm (Yu and Meng, 2011). Appendix J discusses partial CIS

algorithms in general and in the DLM. In the next section, we will characterize the efficiency of the various available samplers in the local level model (LLM).

6. APPLICATION: THE LOCAL LEVEL MODEL

The local level model is a DLM with $F_t = G_t = 1$ for all t while V and W are scalar. We can write the model as

$$y_t|\theta, V, W \overset{ind}{\sim} N(\theta_t, V), \qquad \theta_t|\theta_{0:t-1}, V, W \sim N(\theta_{t-1}, W)$$

for $t = 1, 2, \cdots, T$. The priors on (θ_0, V, W) from Section 3 become $\theta_0 \sim N(m_0, C_0)$, $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$ with θ_0, V and W mutually independent where $IG(\alpha, \beta)$ is the inverse gamma distribution with shape parameter α and rate parameter β. In this model W is often called the signal, V the noise, and $R = W/V$ is the signal-to-noise ratio.

6.1 DAs for the local level model

We can define the various DAs from Section 4 in the context of the local level model. The latent states are simply θ. From the states we obtain the SDs: $\gamma_0 = \theta_0$ and $\gamma_t = (\theta_t - \theta_{t-1})/\sqrt{W}$ for $t = 1, 2, \cdots, T$. Similarly, the SEs are $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \cdots, T$. The WSDs are then $\tilde{\gamma}_0 = \theta_0$ and $\tilde{\gamma}_t = (\theta_t - \theta_{t-1})/\sqrt{V}$ while the WSEs are $\tilde{\psi}_0 = \theta_0$ with $\tilde{\psi}_t = (y_t - \theta_t)/\sqrt{W}$, both for $t = 1, 2, \cdots, T$.

Most of the full conditional distributions required in the LLM follow straightforwardly from the general case and their derivations can be found in Appendix D. For all algorithms, we use the MCFA to draw the DA except in the case of γ, where we use MCFA to draw θ and then transform to γ. For V and W, their draws are either an inverse gamma draw or a

draw from a difficult full conditional. In Appendix D we derive the difficult density in detail and in Appendix G we show how to obtain random draws from it.

6.2 Simulation setup

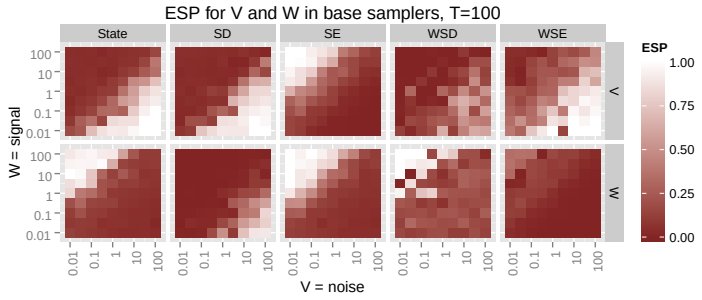
We simulated data from the local level model using a factorial design with V and W each taking the values $10^{i/2}$ where $i = -4, -3, \ldots, 4$ and T taking the values $10, 100,$ & 1000. For each dataset, we fit the model using a variety of the algorithms discussed above. We used the same rule for constructing priors for each model: $\theta_0 \sim N(0, 10^7)$, $V \sim IG(5, 4V^*)$, and $W \sim IG(5, 4W^*)$, mutually independent where (V^*, W^*) are the values used to simulate the time series. The prior means are equal to V^* and W^* so that the prior, likelihood, and thus posterior all roughly agree about the likely values of V and W. This prior allows us to highlight how the behavior of each sampler depends on where the posterior is located.

For each dataset and sampler we obtained $n = 10,500$ posterior draws and threw away the first 500. The chains were started at (V^*, W^*), so they can tell us about mixing but not convergence. Define the effective sample proportion for a scalar component of the chain as the effective sample size (ESS) (Gelman et al., 2013) of the component divided by the number of iterations n ($ESP = ESS/n$). When $ESP = 1$ the chain is behaving as if it obtains iid draws from the posterior. Occasionally $ESP > 1$, if the draws are negatively correlated, but we round it down to one in our plots.

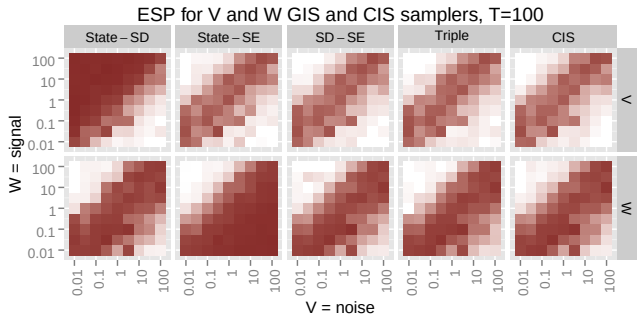
6.3 Simulation results

Figure 1a contains plots of ESP for V and W in each chain of each base sampler for $T = 100$. Let $R^* = V^*/W^*$ denote the true signal-to-noise ratio and note that the likely value of R^* is

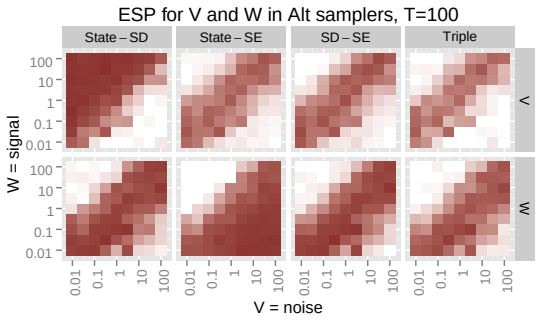
highly application specific. The State sampler tends to have a low ESP for V and high ESP for W when $R^* > 1$ with the behavior switched when $R^* < 1$. The SD sampler has low ESP for both V and W when $R^* > 1$ while the SE sampler has low ESP for both when $R^* < 1$. Table 1 summarizes the results for the base samplers on the top.



(a)



(b)



(c)

Figure 1: Effective sample proportion in the posterior sampler for a time series of length $T = 100$ for V and W in the base sampler (a), GIS and CIS samplers (b), and Alt samplers (c). The axes indicate the true values of V (horizontal) and W (vertical) for the simulated data. The signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

We fit the model using several interweaving (GIS and CIS) samplers as well. Since the

	State	SD	SE	WSD	WSE	State-SD	State-SE	SD-SE	Triple	CIS
V	$R^* < 1$	$R^* < 1$	$R^* > 1$	$R^* < 1$	$R^* < 1$	$R^* < 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$
W	$R^* > 1$	$R^* < 1$	$R^* > 1$	$R^* > 1$	$R^* > 1$	$R^* \not\approx 1$	$R^* > 1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$

Table 1: Rule of thumb for when each sampler has a high ESP for each variable as a function of the true signal-to-noise ratio, $R^* = W^*/V^*$. The right side of the table applies to both the interweaving and alternating algorithms.

wrongly-scaled samplers behaved similarly to the state sampler and neither of the underlying DAs were an SA for (V, W) jointly, we ignored them in the construction of the GIS samplers. Instead, we used the State-SD, State-SE, SD-SE, and Triple (State-SD-SE) GIS samplers, as well as the CIS sampler. Figure 1b has plots of ESP for each of the GIS and CIS algorithms while Figure 1c has plots of ESP for each of the Alt algorithms. Table 1 summarizes the results on the right.

Essentially, each GIS and Alt algorithm has high ESP when at least one of the base algorithms has high ESP. For example, the State-SD GIS and Alt algorithms have high ESP for W except for a narrow band where R^* is near one while ESP is high for W in the state sampler when $R^* > 1$ and in the SD sampler when $R^* < 1$. Similarly in the State-SD GIS and Alt algorithms, mixing for V is identical to the State and SD samplers since neither base sampler improves on the other in any region of the parameter space. Both the State-SD GIS and Alt algorithms take advantage of the fact that the State and SD DA algorithms make a "beauty and the beast" pair for W. However, GIS without an SA-AA pair does not appear to improve on Alt. In Section 5.4 we noted that the CIS and the SD-SE GIS algorithms

consist of the same steps rearranged, which suggests they should perform similarly. In fact The SD-SE GIS algorithm behaves essentially identically to both the CIS and Triple GIS algorithms.

The $T = 10$ and $T = 1000$ plots (Appendix M) are similar, but, as T increases, the region of the parameter space with high ESP shrinks for all samplers. In Appendix K, we discuss how the pattern of correlations between various quantities in the posterior determines the pattern of ESPs in Figure 1.

In Appendix L, we also compare each algorithm based on the time required to adequately characterize the posterior, taking into account both mixing and computational time. GIS and Alt again perform essentially identical in this respect, though there is good reason to expect GIS to sometimes be more efficient. We discuss this in Appendix N and show that for very long time series, GIS does become significantly more efficient than Alt.

7. DISCUSSION

In order to apply the interweaving strategies of Yu and Meng (2011) in DLMs we introduced five DAs, three of them novel. None of these were an SA and we argued through Lemma 1 that it is unlikely that a *useful* SA exists. With available DAs, we constructed several alternating, GIS, and CIS algorithms. In a simulation study using the local level model, we tested these algorithms and found that the true signal-to-noise ratio, $R^* = V^*/W^*$, is important for determining when each algorithm performs well. In addition we found that there appears to be no difference in mixing between a GIS algorithm an its corresponding Alt algorithm for any of the DAs we used. The only caveat is that for very long time

series the GIS version of an algorithm can become cheaper per iteration (Appendix N). Interweaving provides a simple framework to quickly find samplers which perform well, and for this reason we endorse the approach. As one reviewer suggested, a general strategy for constructing interweaving algorithms is as follows: implement the standard DA algorithm for each DA, find the optimal algorithm for each parameter, and combine them with the corresponding SA or AA to construct a CIS sampler. This approach yields our CIS sampler in the LLM, which along with the SD-SE GIS has the best overall performance of all the samplers we consider.

The importance of the signal-to-noise ratio to the properties of various MCMC algorithms has been anticipated in the literature. In the AR(1) plus noise model, Pitt and Shephard (1999) find that the signal-to-noise ratio with the AR(1) coefficient determine the convergence rate of a Gibbs sampler. When Frühwirth-Schnatter (2004) study the dynamic regression model with a stationary AR(1) process on the regression coefficient, they find that the relative behavior of the SD sampler and the State sampler depends on a function of the true signal-to-noise ratio that also depends on the true value of the autocorrelation parameter and the distribution of the covariate. It is likely that a version of the signal-to-noise ratio will determine how well each algorithm performs in the general DLM. This result is probably a consequence of the relevance of the Bayesian (and EM) fraction of missing information to the performance of the DA (and EM) algorithms (Van Dyk and Meng, 2001).

A major computational bottleneck in most of our algorithms occurs when we draw from $p(W|V,\gamma,y)$, $p(V|W,\psi,y)$, $p(V|W,\tilde{\gamma},y)$, or $p(W|V,\tilde{\psi},y)$ as discussed in Appendices G and

H. The densities $p(W|V,\gamma,y)$ and $p(V|W,\psi,y)$ have the form

$$p(x) \propto x^{-\alpha-1} \exp\left[-ax + b\sqrt{x} - c/x\right],$$

while the densities $p(W|V,\tilde{\psi},y)$ and $p(V|W,\tilde{\gamma},y)$ have the form

$$p(x) \propto x^{-\alpha-1} \exp\left[-ax + b/\sqrt{x} - c/x\right]$$

where $\alpha, a, c > 0$ and $b \in \Re$. When $b = 0$ we have a special case of the generalized inverse Gaussian (GIG) distribution, so perhaps the methods used to draw from a GIG can be used here.

This difficulty could be solved by a more judicious choice of priors. We chose inverse Wishart priors for V and W partially because their conditional conjugacy with the states is convenient, but this breaks down when using other DAs. In addition, there are well known inferential problems with the inverse Wishart prior in the hierarchical model literature, e.g. Gelman (2006). An alternative is the conditionally conjugate prior for $\sqrt{W}$ given the SDs. In the LLM this is a Gaussian distribution — strictly speaking this prior is on $\pm\sqrt{W}$. If we use this prior for $\pm\sqrt{V}$ as well, the V step in the SD sampler becomes a draw from the GIG distribution. This prior has been used by Frühwirth-Schnatter and Wagner (2011) and Frühwirth-Schnatter and Tüchler (2008) to speed up computation while using the SDs in hierarchical models and by Frühwirth-Schnatter and Wagner (2010) for time series models with a DA similar to the SDs. We omit the results here, but using this prior on both variances does not alter our mixing results for any of the MCMC samplers.

In the general DLM this prior becomes much more complicated because V and W are matrices. The conditionally conjugate prior for W given γ is now a normal distribution on

L_W, but the full conditional for the other covariance matrix becomes a matrix analogue of the GIG distribution. So no matter which conditionally conjugate prior is used, under the SEs or SDs one of V or W's full conditionals will be intractable. This is not a problem for the DA algorithms necessarily — you have the freedom to use the inverse Wishart prior for V and the normal prior for L_W in the SD sampler, for example. But in any interweaving or alternating algorithm each covariance matrix needs to be drawn from two full conditionals, one of which will be intractable. A Metropolis step is a tolerable solution to the problem, though perhaps we can do better.

8. SUPPLEMENTARY MATERIALS

Appendices: Provides all appendices referenced in the manuscript. (pdf file)

Scripts: Provides R scripts to run the analyses described in the manuscript, please see the `README.txt` for more details. (zip file)

9. ACKNOWLEDGMENTS

The authors thank the participants of the Economics, Finance, and Business workshop at the Bayes 250 conference and of the 2014 Bayesian Young Statisticians meeting for helpful comments, though all errors are our own. The authors would also like to thank three referees, the associated editor, and the editor for valuable comments that improved the manuscript.

REFERENCES

Basu, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhyā: The Indian Journal of Statistics*, 15(4), 377–380.

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (2003), "Non-centered Parameterisations for Hierarchical Models and Data Augmentation," in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, London: Oxford University Press, pp. 307–326.

Bos, C. S., and Shephard, N. (2006), "Inference for Adaptive Time Series Models: Stochastic Volatility and Conditionally Gaussian State Space Form," *Econometric Reviews*, 25(2-3), 219–244.

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81(3), 541–553.

Dempster, A. P., Laird, N. M., Rubin, D. B. et al. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal statistical Society*, 39(1), 1–38.

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15(2), 183–202.

——— (2004), "Efficient Bayesian Parameter Estimation for State Space Models Based on Reparameterizations," in *State Space and Unobserved Component Models: Theory and Applications*, Cambridge, UK: Cambridge University Press, pp. 123–151.

Frühwirth-Schnatter, S., and Sögner, L. (2003), "Bayesian Estimation of the Heston Stochastic Volatility Model," in *Operations Research Proceedings 2002*, eds. A. Harvey, S. J. Koopman, and N. Shephard, Springer, pp. 480–485.

——— (2008), "Bayesian Estimation of the Multi-factor Heston Stochastic Volatility Model," *Communications in Dependability and Quality Management*, 11(4), 5–25.

Frühwirth-Schnatter, S., and Tüchler, R. (2008), "Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models," *Statistics and Computing*, 18(1), 1–13.

Frühwirth-Schnatter, S., and Wagner, H. (2006), "Auxiliary Mixture Sampling for Parameter-driven Models of Time Series of Counts with Applications to State Space Modelling," *Biometrika*, 93(4), 827–841.

——— (2010), "Stochastic Model specification Search for Gaussian and Partial Non-Gaussian State Space models," *Journal of Econometrics*, 154(1), 85–100.

——— (2011), "Bayesian Variable Selection for Random Intercept Modeling of Gaussian and Non-Gaussian Data," in *Bayesian Statistics 9*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford: Oxford University Press, pp. 165–200.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82(3), 479–488.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)," *Bayesian analysis*, 1(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis (3rd ed.)*, New York: CRC press.

Hobert, J. P., and Marchev, D. (2008), "A Theoretical Comparison of the Data Augmentation, Marginal Augmentation and PX-DA Algorithms," *The Annals of Statistics*, 36(2), 532–554.

Kastner, G., and Frühwirth-Schnatter, S. (2014), "Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of Stochastic Volatility Models," *Computational Statistics & Data Analysis*, 76, 408–423.

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94(448), 1264–1274.

McCausland, W. J., Miller, S., and Pelletier, D. (2011), "Simulation Smoothing for State–Space models: A Computational Efficiency Analysis," *Computational Statistics & Data Analysis*, 55(1), 199–212.

Meng, X.-L., and Van Dyk, D. (1997), "The EM Algorithm–An Old Folk-song Sung to a Fast New Tune," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3), 511–567.

——— (1998), "Fast EM-type Implementations for Mixed Effects Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3), 559–578.

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal augmentation," *Biometrika*, 86(2), 301–320.

Papaspiliopoulos, O., and Roberts, G. (2008), "Stability of the Gibbs Sampler for Bayesian Hierarchical Models," *The Annals of Statistics*, 36(1), 95–117.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007), "A General Framework for the Parametrization of Hierarchical Models," *Statistical Science*, 22(1), 59–73.

Petris, G., Campagnoli, P., and Petrone, S. (2009), *Dynamic Linear Models with R*, New York: Springer.

Pitt, M. K., and Shephard, N. (1999), "Analytic Convergence Rates and Parameterization Issues for the Gibbs Sampler Applied to State Space Models," *Journal of Time Series Analysis*, 20(1), 63–85.

Prado, R., and West, M. (2010), *Time Series: Modeling, Computation, and Inference*, London: CRC Press.

Roberts, G. O., Papaspiliopoulos, O., and Dellaportas, P. (2004), "Bayesian Inference for non-Gaussian Ornstein–Uhlenbeck Stochastic Volatility Processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2), 369–393.

Roberts, G. O., and Sahu, S. K. (1997), "Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2), 291–317.

Shephard, N. (1996), *Statistical Aspects of ARCH and Stochastic Volatility*, London: Springer.

Simpson, M. (2015), "Application of Interweaving in DLMs to an Exchange and Specialization Experiment," in *Bayesian Statistics from Methods to Models and Applications: Research from BAYSM 2014*, eds. S. Frühwirth-Schnatter, A. Bitto, G. Kastner, and A. Posekany, Springer.

Strickland, C. M., Martin, G. M., and Forbes, C. S. (2008), "Parameterisation and Efficient MCMC Estimation of Non-Gaussian State Space Models," *Computational Statistics & Data Analysis*, 52(6), 2911–2930.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82(398), 528–540.

Van Dyk, D. A., and Tang, R. (2003), "The One-step-late PXEM Algorithm," *Statistics and Computing*, 13(2), 137–152.

Van Dyk, D., and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10(1), 1–50.

West, M., and Harrison, J. (1999), *Bayesian Forecasting & Dynamic Models (2nd ed.)*, New York: Springer.

Yu, Y., and Meng, X.-L. (2011), "To Center or not to Center: That is not the Question - An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency," *Journal of Computational and Graphical Statistics*, 20(3), 531–570.

Appendices for Interweaving Markov Chain Monte Carlo Strategies for Efficient Estimation of Dynamic Linear Models

Matthew Simpson
Department of Statistics, University of Missouri–Columbia

Jarad Niemi and Vivekananda Roy
Department of Statistics, Iowa State University

October 2, 2015

The following appendices are cited in the main article and included here:

A. Derivation of marginal model for the DLM

B. Proof of lemma 1

C. Construction of the wrongly-scaled DA algorithms

D. Derivations of relevant joint and full conditional distributions

E. MCFA for simulation smoothing

F. Further augmentation for non-invertible F_t

G. Efficiently drawing from $p(W|V, \gamma, y)$ and $p(V|W, \psi, y)$ in the LLM

H. Efficiently drawing from $p(W|V, \tilde{\gamma}, y)$ and $p(V|W, \tilde{\psi}, y)$ in the LLM

I. Equivalence of CIS and SD-SE GIS in the DLM

J. Partial CIS Algorithms in the DLM

K. Using posterior correlations to understand patterns of ESP

L. Computational time for each algorithm

M. Additional plots for other values of T

N. Comparing GIS and Alt in very long time series.

A. MARGINAL MODEL OF THE DLM

The class of DLMs we consider is

$$y_t|\theta, V, W \overset{ind}{\sim} N_k(F_t\theta_t, V) \qquad \theta_t|\theta_{0:t-1}, V, W \sim N_p(G_t\theta_{t-1}, W) \tag{A.1}$$

for $t = 1, 2, \cdots T$ where V and W are unknown covariance matrices. Define $v_t = y_t - F_t\theta_t$ and $w_t = \theta_t - G_t\theta_{t-1}$. Then we can rewrite the model by recursive substitution:

$$y_t = v_t + F_t \left(w_t + G_t w_{t-1} + G_t G_{t-1} w_{t-2} + ... + G_t G_{t-1} \cdots G_2 w_1 + G_t G_{t-1} \cdots G_1 \theta_0\right).$$

Then conditional on $\phi = (V, W)$ each y_t is a linear combination of normal random variables. After marginalizing out θ, $y = (y_1', y_2', \ldots, y_T')$ has a normal distribution such that $\mathrm{E}[y_t|\phi] = F_t H_t m_0$,

$$\mathrm{Var}[y_t|\phi] = V + F_t(K_t W K_t' + H_t C_0 H_t')F_t', \quad \text{and} \quad \mathrm{Cov}[y_s, y_t|\phi] = F_s(K_s W K_t' + H_s C_0 H_t')F_t',$$

where $H_t = G_t G_{t-1} \cdots G_1$ and $K_t = I_p + G_t + G_t G_{t-1} + \cdots + G_t G_{t-1} \cdots G_2$. Next define $D_t = F_t G_t G_{t-1} \cdots G_1$. Then let $\tilde{V} = I_T \otimes V$ and D be block diagonal with elements $D_1, \ldots, D_T$,

$$\tilde{W}_{Tk\times Tk} = \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix}' W \begin{bmatrix} K_1'F_1' & K_2'F_2' & \cdots K_T'F_T' \end{bmatrix},$$
$$\tilde{C}_{Tk\times Tk} = \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix}' C_0 \begin{bmatrix} H_1'F_1' & H_2'F_2' & \cdots H_T'F_T' \end{bmatrix},$$

and $\tilde{m}_{Tp\times 1} = (m_0', m_0', \cdots m_0')'$. Now we have the data model for y without any data augmentation:

$$y|V, W \overset{ind}{\sim} N_{Tk}(D\tilde{m}, \tilde{V} + \tilde{W} + \tilde{C}). \tag{A.2}$$

B. PROOF OF LEMMA 1

First the normality assumption implies

$$y|\eta,\phi \sim N(D\tilde{m} + \Omega'_{y,\eta}\Omega_\eta^{-1}(\eta - \alpha_\eta), \tilde{V} + \tilde{W} + \tilde{C} - \Omega'_{y,\eta}\Omega_\eta^{-1}\Omega_{y,\eta})$$

$$\eta|\phi \sim N(\alpha_\eta, \Omega_\eta).$$

Now for η to be a sufficient augmentation we need $D\tilde{m} + \Omega'_{y,\eta}\Omega_\eta^{-1}(\eta - \alpha_\eta)$ and $\tilde{V} + \tilde{W} + \tilde{C} - \Omega'_{y,\eta}\Omega_\eta^{-1}\Omega_{y,\eta}$ to be functionally independent of ϕ. This requires that

$$D\tilde{m} - \Omega'_{y,\eta}\Omega_\eta^{-1}\alpha_\eta + \Omega'_{y,\eta}\Omega_\eta^{-1}\eta = b + A\eta$$

where $A = \Omega'_{y,\eta}\Omega_\eta^{-1}$ and $b = D\tilde{m} - A\alpha_\eta$ must both be free of ϕ. As a result $A\alpha_\eta$ is also free of ϕ and thus so is α_η.

Then using the second equation, we now require Σ free of ϕ where $\Sigma = \tilde{V} + \tilde{W} + \tilde{C} - A\Omega_\eta A'$. This ensures that $\Omega_{\eta,y}$ is not the zero matrix since $\tilde{V}+\tilde{W}+\tilde{C}$ is not free of ϕ. Rearranging we have $A\Omega_\eta A' = \tilde{V}+\tilde{W}+\tilde{C}-\Sigma$. Consider $\tilde{\eta} = A\eta$, which is also a sufficient augmentation since it is just a linear transformation by a constant matrix. Then we have

$$y|\tilde{\eta},\phi \sim N(b + A\eta, \Sigma)$$

$$\tilde{\eta}|\phi \sim N(A\alpha_\eta, A\Omega_\eta A')$$

in other words

$$y|\tilde{\eta},\phi \sim N(b + \tilde{\eta}, \Sigma)$$

$$\tilde{\eta}|\phi \sim N(A\alpha_\eta, \tilde{V} + \tilde{W} + \tilde{C} - \Sigma).$$

Thus the posterior density of ϕ given $\tilde{\eta}$ can be written as

$$\begin{aligned} p(\phi|\tilde{\eta},y) &\propto p(y|\tilde{\eta},\phi)p(\tilde{\eta}|\phi)p(\phi) \propto p(\tilde{\eta}|\phi)p(\phi) \\ &\propto p(\phi)|\tilde{V} + \tilde{W} + \tilde{C} - \Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\tilde{\eta} - A\alpha_\eta)'(\tilde{V} + \tilde{W} + \tilde{C} - \Sigma)^{-1}(\tilde{\eta} - A\alpha_\eta)\right]. \end{aligned}$$

Now given that $A'A$ is invertible and the properties of multivariate normal distributions, the density of $p(\phi|\eta,y)$ follows from $\eta = (A'A)^{-1}A'\tilde{\eta}$.

C. CONSTRUCTION OF THE WRONGLY-SCALED DA ALGORITHMS

The wrongly-scaled DA algorithms are close analogues to their correctly scaled cousins. Starting with the *wrongly-scaled disturbance sampler* (Algorithm WSD), the simulation smoothing step to draw from $p(\tilde{\gamma}|V, W, y)$ is similar to that of the scaled disturbance sampler — the density is Gaussian, but the precision matrix is not tridiagonal, so we draw θ using the MCFA and transform to obtain a draw of $\tilde{\gamma}$. The density of $V, W|\tilde{\gamma}, y$ is too complicated to draw from directly, as was the case when we used the scaled disturbances. In this case, the full conditional distribution of W is the same as its distribution when we condition on the states while the density of $V|\tilde{\gamma}, y$ is once again difficult to draw from. The density of $V|W, \tilde{\gamma}, y$ is easier to work with, at least in the local level model example in Section 6.

Algorithm: WSD. *Wrongly-Scaled Disturbance Sampler*

1. *Use MCFA to draw* $\theta \sim p(\theta|V, W, y)$.
2. *Transform* θ *to* $\tilde{\gamma}$.
3. *Draw* $V \sim p(V|W, \tilde{\gamma}, y)$.
4. *Draw* $W \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$.

Now the third step is difficult and we demonstrate how to accomplish it in the local level model in Appendix F. We could switch the order in which V and W are drawn in this algorithm so that we can draw W before transforming θ to $\tilde{\gamma}$. This would make each iteration slightly cheaper and probably would not affect the mixing and convergence properties of the algorithm, however we are more interested in comparing the mixing and convergence properties of the various samplers, so we always sample V before W when we cannot sample them jointly.

The *wrongly-scaled error sampler* (Algorithm WSE) is closely related to both the wrongly-scaled disturbance sampler and the scaled error sampler. The density of $\tilde{\psi}|V, W, y$ is Gaussian with a tridiagonal precision matrix, so the simulation smoothing step can be accomplished using the MCFA. The density $p(V, W|\tilde{\psi}, y)$ is from the same class as $p(W, V|\tilde{\gamma}, y)$ so that V and W essentially switch places when we condition on $\tilde{\psi}$

instead of $\tilde{\gamma}$. In particular, $V|W, \tilde{\psi}, y$ has an inverse Wishart density and the density of $W|V, \tilde{\psi}, y$ is from the same class as that of $V|W, \tilde{\gamma}, y$.

Algorithm: WSE. *Wrongly-Scaled Error Sampler*

1. *Use MCFA to draw* $\tilde{\psi} \sim p(\theta|V, W, y)$.
2. *Draw* $V \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right)$.
3. *Draw* $W \sim p(W|V, \tilde{\psi}, y)$

The constructions of Algorithms WSD and WSE in the local level model example from Section 6 require $p(W|V, \tilde{\psi}, y)$ and $p(V|W, \tilde{\gamma}, y)$ respectively. Both densities have the form $p(x) \propto x^{-\alpha-1} \exp\left[-ax + b/\sqrt{x} - c/x\right]$, which is closely related to the difficult density from the correctly scaled samplers. For $p(V|W, \tilde{\gamma}, y)$ we show in Appendix C that $\alpha = \alpha_V$, $a = a_{\tilde{\gamma}} \equiv \frac{1}{2W}\sum_{t=1}^{T} \tilde{\gamma}_t^2$, $b = b_{\tilde{\gamma}} \equiv \sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)\sum_{s=1}^{t} \tilde{\gamma}_s$, and $c = c_{\tilde{\gamma}} \equiv \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)^2$ while for $p(W|V, \tilde{\psi}, y)$ we show that $\alpha = \alpha_W$, $a = a_{\tilde{\psi}} \equiv \frac{1}{2V}\sum_{t=1}^{T} \tilde{\psi}_t^2$, $b = b_{\tilde{\psi}} \equiv \sum_{t=1}^{T} \mathcal{L}\tilde{y}_t \mathcal{L}\tilde{\psi}_t$, and $c = c_{\tilde{\psi}} \equiv \beta_W + \frac{1}{2}\sum_{t=1}^{T} \mathcal{L}\tilde{y}_t^2$. This density is harder to sample from because adaptive rejection sampling does not work very well, so we construct a rejection sampler on the log scale using a t approximation in Appendix G.

D. FULL CONDITIONAL DISTRIBUTIONS IN THE GENERAL DLM FOR VARIOUS DAS

The class of DLMs we consider is defined as follows:

$$y_t = F_t\theta_t + v_t \qquad v_t \overset{ind}{\sim} N_k(0, V) \qquad \text{(observation equation)} \tag{A.3}$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \overset{ind}{\sim} N_p(0, W) \qquad \text{(system equation)} \tag{A.4}$$

for $t = 1, 2, \cdots T$ with the priors $\theta_0 \sim N_p(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ and $W \sim IW(\Lambda_W, \lambda_W)$ with (θ_0, V, W) mutually independent. Then the full joint distribution of (V, W, θ, y) is

$$\begin{aligned} p(V, W, \theta, y) &\propto \exp\left[-\frac{1}{2}(\theta_0 - m_0)' C_0^{-1}(\theta_0 - m_0)\right] \\ &\times |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \\ &\times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right] \end{aligned} \tag{A.5}$$

where tr(.) is the matrix trace operator.

In the following subsections, we provide derivations of the full conditional distributions for when using states, scaled disturbances or scaled errors as the data augmentation.

D.1 States

With the usual DA, the full conditional distributions can be derived from equation (A.5). First, the full conditional distribution of θ is as follows:

$$\begin{aligned} p(\theta | V, W, y) &\propto p(V, W, \theta, y) \propto \exp\left[-\frac{1}{2}(\theta_0 - m_0)' C_0^{-1}(\theta_0 - m_0)\right] \\ &\times \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right]. \end{aligned}$$

It turns out that this density is Gaussian. In Section E, we show how to use the mixed Cholesky factorization algorithm (MCFA) in order to efficiently determine and draw from this distribution.

The full conditional of (V, W) is:

$$\begin{aligned} p(V, W | \theta, y) &\propto p(V, W, \theta, y) \propto |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \\ &\times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right] \\ &\propto |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\left(\Lambda_V + \sum_{t=1}^{T}(y_t - F_t\theta_t)(y_t - F_t\theta_t)'\right) V^{-1}\right)\right] \\ &\times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\left(\Lambda_W + \sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})(\theta_t - G_t\theta_{t-1})'\right) W^{-1}\right)\right]. \end{aligned}$$

In other words, V and W are conditionally independent given y and θ with

$$V|\theta, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right), \qquad W|\theta, y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $v_t = y_t - F_t\theta_t$ and $w_t = \theta_t - G_t\theta_{t-1}$.

In the local level model, the priors on V and W become $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$. The full conditionals then become

$$V|\theta, y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t)^2/2\right), \quad W|\theta, y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2/2\right).$$

D.2 Scaled disturbances

Let L_W denote the Cholesky decomposition of W, i.e. the lower triangle matrix L_W such that $L_W L_W' = W$. Then the scaled disturbances are $\gamma = \gamma_{0:T} = (\gamma_0', \gamma_1', \cdots, \gamma_T')'$ defined by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$. The reverse transformation is defined recursively by $\theta_0 = \gamma_0$ and $\theta_t = L_W\gamma_t + G_t\theta_{t-1}$ for $t = 1, 2, \cdots, T$. Then the Jacobian is block lower triangular with the identity matrix and T copies of L_W along the diagonal blocks, so $|J| = |L_W|^T = |W|^{T/2}$. From equation (A.5) we can write the full joint distribution of (V, W, γ, y) as

$$\begin{aligned} p(V, W, \gamma, y) \propto & \exp\left[-\frac{1}{2}(\gamma_0 - m_0)' C_0^{-1}(\gamma_0 - m_0)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] \\ & \times |W|^{-(\lambda_W + p + 2)/2} |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \\ & \times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right]. \end{aligned} \tag{A.6}$$

where $\theta_t(\gamma, W)$ denotes the recursive back transformation defined by the scaled disturbances. The full conditional distribution of γ is then

$$\begin{aligned} p(\gamma|V, W, y) \propto p(V, W, \gamma, y) \propto & \exp\left[-\frac{1}{2}(\gamma_0 - m_0)' C_0^{-1}(\gamma_0 - m_0)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] \\ & \times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right]. \end{aligned}$$

This density is Gaussian, but difficult to draw from. We use the MCFA to draw from $\theta|V, W, y$ instead, then transform from θ to γ using the definition of γ.

Under this parameterization, the full conditional distribution of (V, W) is

$$p(V, W, |\gamma, y) \propto p(V, W, \gamma, y)|W|^{-(\lambda_W+p+2)/2}|V|^{-(\lambda_V+k+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma, W)\right]' V^{-1}\left[y_t - F_t\theta_t(\gamma, W)\right]\right)\right].$$

The back transformation from θ to γ sets $\theta_0 = \gamma_0$ and for $t = 1, 2, \cdots, T$

$$\begin{aligned}\theta_t &= L_W\gamma_t + G_t\theta_{t-1} \\ &= L_W\gamma_t + \sum_{s=0}^{t-2} G_tG_{t-1}\ldots G_{t-s}L_W\gamma_{t-s-1} + G_tG_{t-1}\ldots G_1\gamma_0 \\ &= \sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} + \tilde{G}_{t,t}\gamma_0\end{aligned}$$

where $\tilde{G}_{s,t} = G_tG_{t-1}\cdots G_{t-s+1}$ for $s > 0$ and $\tilde{G}_{0,t} = I_p$, the $p \times p$ identity matrix.. Then we can rewrite the conditional distribution of (V, W) as

$$p(V, W, |\gamma, y) \propto p(V, W, \gamma, y) \propto |W|^{-(\lambda_W+p+2)/2}|V|^{-(\lambda_V+k+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]' V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right].$$

This density is fairly complicated, so we resort to the full conditionals of V and W separately. The full conditional of V is familiar:

$$p(V|W, \gamma, y) \propto p(V, W|\gamma, y) \propto |V|^{-(\lambda_V+k+T+2)/2} \times \exp\left[-\frac{1}{2}\left(\operatorname{tr}\left[\Lambda_V + \sum_{t=1}^{T} v_tv_t'\right] V^{-1}\right)\right]$$

where $v_t = y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0 = y_t - F_t\theta_t$. This implies that

$$V|W, \gamma, y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_tv_t', \lambda_V + T\right)$$

which is the same distribution as for $V|\theta, y$. In the local level model this reduces to

$$V|W, \gamma, y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t(\gamma))^2/2\right)$$

which is again the same density if we conditioned on θ.

The full conditional density of W is more complicated:

$$p(W|V,\gamma,y) \propto p(V,W,\gamma,y) \propto |W|^{-(\lambda_W+p+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]' V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right].$$

In the local level model, the density is even simpler:

$$p(W|V,\gamma,y) \propto W^{-\alpha_W-1}\exp\left[-\frac{1}{W}\beta_W\right]\exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - \sum_{s=0}^{t}\gamma_{t-s}\sqrt{W}\right]' V^{-1}\left[y_t - \sum_{s=0}^{t-1}\gamma_{t-s}\sqrt{W}\right]\right)\right]$$
$$\propto W^{-\alpha_W-1}\exp\left[-a_\gamma W + b_\gamma\sqrt{W} - \frac{\beta_W}{W}\right].$$

where $a_\gamma = \sum_{t=1}^{T}(\sum_{s=1}^{t}\gamma_j)^2/2V$ and $b_\gamma = \sum_{t=1}^{T}(y_t-\gamma_0)(\sum_{s=1}^{t}\gamma_j)/V$. In Section G we show how to efficiently obtain a random draw from this density.

D.3 Scaled errors

Let L_V denote the Cholesky decomposition of V, that is $L_V L_V' = V$, then we can define the scaled errors as $\psi_t = L_V^{-1}(y_t - F_t\theta_t)$ for $t = 1, 2, \cdots, T$ and $\psi_0 = \theta_0$. Here we assume that $k = p$ and that F_t is invertible for all t. Then the back transformation is $\theta_t = F_t^{-1}(y_t - L_V\psi_t)$ for $t = 1, 2, \cdots, T$ and $\theta_0 = \psi_0$. The Jacobian of this transformation is block diagonal with a single copy of the identity matrix along with the $F_t^{-1}L_V$'s along the diagonal, so $|J| = (\prod_{t=1}^{T}|F_t|^{-1})|V|^{T/2}$. Then from equation (A.5) we can write the joint distribution of (V, W, ψ, y) as

$$p(V,W,\psi,y) \propto \exp\left[-\frac{1}{2}(\psi_0 - m_0)'C_0^{-1}(\psi_0 - m_0)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t'\psi_t\right]$$
$$\times |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \times |W|^{-(\lambda_W+p+T+2)/2}$$
$$\exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_W W^{-1}\right) + \sum_{t=1}^{T}(y_t-\mu_t)'(F_t W F_t')^{-1}(y_t-\mu_t)\right)\right] \tag{A.7}$$

where we define $\mu_1 = L_V\psi_1 + F_1G_1\psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = L_V\psi_t + F_tG_tF_{t-1}^{-1}(y_{t-1} - L_V\psi_{t-1})$. The $|F_t|^{-1}$'s have been absorbed into the normalizing constant, but if they depended on some unknown parameter then we could not do this and as a result would have to take them into account in the Gibbs step or steps for the model parameters.

The full conditional distribution of ψ is

$$p(V, W, \psi, y) \propto \exp\left[-\frac{1}{2}(\psi_0 - m_0)' C_0^{-1}(\psi_0 - m_0)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \psi_t' \psi_t\right]$$
$$\exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}(y_t - \mu_t)'(F_t W F_t')^{-1}(y_t - \mu_t)\right)\right]$$

where note that μ_t depends on ψ. This density is Gaussian and like with γ, we can use the MCFA from Section E to draw from the full conditional of θ and then transform from θ to ψ. However it turns out the precision matrix of ψ's full conditional distribution has the necessary block tridiagonal structure, so we use the MCFA directly on ψ.

The full conditional distribution of (V, W) is complicated, like the case of the scaled disturbances, so we find the full conditionals of V and W separately instead. The full conditional of W is

$$p(W|V, \psi, y) \propto |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\left(\text{tr}\left(\left[\Lambda_W + \sum_{t=1}^{T} F_t^{-1}(y_t - \mu_t)(y_t - \mu_t)'(F_t^{-1})'\right] W^{-1}\right)\right)\right],$$

in other words

$$W|V, \psi, y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $w_t = F_t^{-1}(y_t - \mu_t) = \theta_t - G_t\theta_{t-1}$. In the local level model, this becomes

$$W|V, \psi, y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^{T}(\theta_t(\psi) - \theta_{t-1}(\psi))^2/2\right).$$

The full conditional distribution of V is more complicated:

$$p(V|W, \psi, y) \propto p(V, W, \psi, y) \propto |V|^{-(\lambda_V + p + 2)/2} \exp\left[-\frac{1}{2}\,\text{tr}\left(\Lambda_V V^{-1} + \sum_{t=1}^{T}(y_t - \mu_t)'(F_t W F_t')^{-1}(y_t - \mu_t)\right)\right]$$

with μ_t a function of V, defined above. In the local level model with an $IG(\alpha_V, \beta_V)$ prior on V, this density is simpler:

$$p(V|W, \psi, y) \propto V^{-\alpha_V - 1} \exp\left[-\frac{\beta_V}{V} + \frac{1}{W}\sum_{t=1}^{T}(y_t - \mu_t)'(y_t - \mu_t)\right]$$

where $\mu_1 = \sqrt{V}\psi_1 + \psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = \sqrt{V}(\psi_t - \psi_{t-1}) + y_{t-1}$. Thus

$$p(V|W, \psi, y) \propto V^{-\alpha_V - 1} \exp\left[-a_\psi V + b_\psi\sqrt{V} - \frac{\beta_V}{V}\right]$$

where $a_\psi = \sum_{t=1}^T (\mathcal{L}\psi_t)^2/2W$ and $b_\psi = \sum_{t=1}^T (\mathcal{L}\psi_t \mathcal{L}y_t)/W$, and we define $\mathcal{L}y_t = y_t - y_{t-1}$ for $t = 2, 3, \cdots, T$, $\mathcal{L}y_1 = y_1 - \psi_0$, $\mathcal{L}\psi_t = \psi_t - \psi_{t-1}$ for $t = 2, 3, ..., T$ and $\mathcal{L}\psi_1 = \psi_1 - 0$. In other words, the form of $p(V|W, \psi, y)$ is the same as $p(W|V, \gamma, y)$. The general form of these two densities is $p(x) \propto x^{-\alpha-1} \exp[-ax + b\sqrt{x} - c/x]$. In Section G we show how to efficiently sample from this distribution.

D.4 The wrongly-scaled disturbances

The wrongly-scaled disturbances are defined as $\tilde{\gamma} = \tilde{\gamma}_{0:T} = (\tilde{\gamma}_0', \tilde{\gamma}_1', \cdots, \tilde{\gamma}_T')'$. The wrongly-scaled disturbances are related to the scaled disturbances by $\tilde{\gamma}_t = L_V^{-1} L_W \gamma_t$ for $t = 1, 2, \cdots, T$ and $\tilde{\gamma}_0 = \gamma_0$. The reverse transformation is $\gamma_t = L_W^{-1} L_V \tilde{\gamma}_t$ and the Jacobian is block diagonal with a copy of the identity matrix and T copies of $L_W^{-1} L_V$ along the diagonal. Thus $|J| = |L_W|^{-T} |L_V|^T = |W|^{-T/2} |V|^{T/2}$. Then from equation (A.6) we can write the joint distribution of $(V, W, \tilde{\gamma}, y)$ as

$$\begin{aligned} p(V, W, \tilde{\gamma}, y) \propto & \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1} (\tilde{\gamma}_0 - m_0)\right] |V|^{-(\lambda_V + p + 2)/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \\ & \times \exp\left[-\frac{1}{2} \sum_{t=1}^T \left(y_t - F_t \theta_t(\tilde{\gamma}, L_V)\right)' V^{-1} \left(y_t - F_t \theta_t(\tilde{\gamma}, L_V)\right)\right] \\ & \times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2} \sum_{t=1}^T \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \end{aligned} \tag{A.8}$$

where $\theta_t(\tilde{\gamma}, L_V)$ denotes the transformation from $\tilde{\gamma}$ to θ defined by the wrongly-scaled disturbances.

Now from equation (A.8), we can write the full conditional density of $\tilde{\gamma}$ as

$$\begin{aligned} p(\tilde{\gamma}|V, W, y) \propto & \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1} (\tilde{\gamma}_0 - m_0)\right] \exp\left[-\frac{1}{2} \sum_{t=1}^T \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \\ & \times \exp\left[-\frac{1}{2} \sum_{t=1}^T \left(y_t - F_t \theta_t(\tilde{\gamma}, L_V)\right)' V^{-1} \left(y_t - F_t \theta_t(\tilde{\gamma}, L_V)\right)\right]. \end{aligned}$$

This density is Gaussian but difficult to draw from, so we use the MCFA to draw $\theta|V, W, y$ instead, then transform from θ to $\tilde{\gamma}$.

Then full conditional density of (V, W) is complicated, but their separate full conditionals are easier to work with. The full conditional density of W is

$$p(W|V, \tilde{\gamma}, y) \propto |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2} \operatorname{tr}\left(\left[\Lambda_W + \sum_{t=1}^T L_V \tilde{\gamma}_t \tilde{\gamma}_t' L_V'\right] W^{-1}\right)\right],$$

i.e.

$$W|V,\tilde{\gamma},y \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$$

where $w_t = L_V\tilde{\gamma}_t = \theta_t - G_t\theta_{t-1}$. In the local level model, this density becomes

$$W|V,\tilde{\gamma},y \sim IG\left(\alpha_W + T/2, \beta_W + \sum_{t=1}^{T}(\theta_t(\tilde{\gamma}) - \theta_{t-1}(\tilde{\gamma}))^2/2\right).$$

The full conditional density of V is more complicated, from equation (A.8):

$$p(V|W,\tilde{\gamma},y) \propto |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\gamma}_t'(L_V^{-1}W(L_V^{-1})')^{-1}\tilde{\gamma}_t\right]$$
$$\times\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(y_t - F_t\theta_t(\tilde{\gamma},L_V)\right)'V^{-1}\left(y_t - F_t\theta_t(\tilde{\gamma},L_V)\right)\right].$$

In the local level model with an $IG(\alpha_V,\beta_V)$ prior on V, this density becomes simpler. Since in that case $\theta_t = \sqrt{V}\sum_{s=1}^{t}\tilde{\gamma}_s + \tilde{\gamma}_0$, we have

$$p(V|W,\tilde{\gamma},y) \propto V^{-\alpha_V-1}\exp\left[-a_{\tilde{\gamma}}V + b_{\tilde{\gamma}}/\sqrt{V} - c_{\tilde{\gamma}}/V\right]$$

where $a_{\tilde{\gamma}} = \frac{1}{2W}\sum_{t=1}^{T}\tilde{\gamma}_t^2$, $b_{\tilde{\gamma}} = \sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)\sum_{s=1}^{t}\tilde{\gamma}_s$, and $c_{\tilde{\gamma}} = \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)^2$. We show in Section H how to efficiently obtain a random draw from this density.

D.5 The wrongly-scaled errors

The wrongly-scaled errors are denoted by $\tilde{\psi} = \tilde{\psi}_{0:T} = (\tilde{\psi}_0', \tilde{\psi}_1', \cdots, \tilde{\psi}_T')'$. They are related to the scaled errors by $\tilde{\psi}_t = L_W^{-1}L_V\psi_t$ for $t = 1,2,\cdots,T$ and $\tilde{\psi}_0 = \psi_0$. Then $\psi_t = L_V^{-1}L_W\tilde{\psi}_t$ and the Jacobian is block diagonal with a copy of the identical matrix and T copies of $L_V^{-1}L_W$ along the diagonal. So $|J| = |V|^{-T/2}|W|^{T/2}$ and from equation (A.7) we can write the joint distribution of $(V,W,\tilde{\psi},y)$ as

$$\begin{aligned}
p(V,W,\tilde{\psi},y) &\propto \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)'C_0^{-1}(\tilde{\psi}_0 - m_0)\right] \\
&\times |V|^{-(\lambda_V+p+T+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right] \\
&\times |W|^{-(\lambda_W+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]\exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\mu}_t)'(F_t W F_t')^{-1}(y_t - \tilde{\mu}_t)\right] \qquad \text{(A.9)}
\end{aligned}$$

where we define $\tilde{\mu}_1 = L_W\tilde{\psi}_1 - F_1G_1\tilde{\psi}_0$ and for $t = 2,3,\cdots,T$ $\tilde{\mu}_t = L_W\tilde{\psi}_t - F_tG_tF_{t-1}^{-1}(y_{t-1} - L_W\tilde{\psi}_{t-1})$.

From equation (A.9) the full conditional distribution of $\tilde{\psi}$ is

$$p(\tilde{\psi}|V,W,y) \propto \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)'C_0^{-1}(\tilde{\psi}_0 - m_0)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right]$$
$$\times \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\mu}_t)'(F_tWF_t')^{-1}(y_t - \tilde{\mu}_t)\right].$$

This density is again Gaussian and it can be shown that the precision matrix is tridiagonal, so the MCFA can be directly applied. The full conditional density of V is the familiar inverse Wishart:

$$p(V|W,\tilde{\psi},y) \propto |V|^{-(\lambda_V+p+T+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right].$$

So $V|W,\tilde{\psi},y \sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right)$ where $v_t = L_W\tilde{\psi}_t = y_t - F_t\theta_t$. In the local level model, this becomes

$$V|W,\tilde{\psi},y \sim IG\left(\alpha_V + T/2, \beta_V + \sum_{t=1}^{T}(y_t - \theta_t(\tilde{\psi}))^2/2\right).$$

The full conditional density of W is more complicated, but has the same form as the full conditional density of V given $\tilde{\gamma}$:

$$p(W|V,\tilde{\psi},y) \propto |W|^{-(\lambda_W+p+2)/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\tilde{\psi}_t'(L_W^{-1}V(L_W^{-1})')^{-1}\tilde{\psi}_t\right]$$
$$\times \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\mu}_t)'(F_tWF_t')^{-1}(y_t - \tilde{\mu}_t)\right].$$

In the case of the local level model with a $IG(\alpha_W, \beta_W)$ prior on W, this density simplifies to

$$p(W|V,\tilde{\psi},y) \propto W^{-\alpha_W-1} \exp\left[-a_{\tilde{\psi}}W + b_{\tilde{\psi}}/\sqrt{W} - c_{\tilde{\psi}}/W\right]$$

where $a_{\tilde{\psi}} = \frac{1}{2V}\sum_{t=1}^{T}\tilde{\psi}_t^2$, $b_{\tilde{\psi}} = \sum_{t=1}^{T}\mathcal{L}\tilde{y}_t\mathcal{L}\tilde{\psi}_t$, and $c_{\tilde{\psi}} = \beta_W + \frac{1}{2}\sum_{t=1}^{T}\mathcal{L}\tilde{y}_t^2$. Here we define $\mathcal{L}y_t = y_t - y_{t-1}$ for $t = 2,3,\cdots,T$ while $\mathcal{L}y_1 = y_1 - \tilde{\psi}_0$, and $\mathcal{L}\tilde{\psi}_t = \tilde{\psi}_t - \tilde{\psi}_{t-1}$ for $t = 2,3,\cdots,T$ while $\mathcal{L}\tilde{\psi}_1 = \tilde{\psi}_1 - 0$. This is the same family of densities as $p(V|W,\tilde{\gamma},y)$, and in Section H we show how to efficiently obtain random draws.

E. MIXED CHOLESKY FACTORIZATION ALGORITHM (MCFA) FOR SIMULATION SMOOTHING

Traditionally in DLMs, forward filtering, backward sampling (FFBS) is used in order to draw from the latent states $\theta_{0:T}$. This requires running the Kalman filter in order to determine the marginal distribution of θ_T, then drawing $\theta_t|\theta_{t+1:T}$ for $t = T-1, T-2, \cdots, 1$ (Carter & Kohn 1994; Frühwirth-Schnatter 1994). The mixed Cholesky factorization algorithm (MCFA) determines the joint distribution of $\theta_{0:T}$ and draws from it using a backward sampling step as in FFBS. The idea comes from Rue (2001), which introduces a Cholesky factorization algorithm (CFA) for drawing from a Gaussian Markov random field and notes that the conditional distribution of $\theta_{0:T}$ given $y_{1:T}$ in a Gaussian linear statespace model is a special case, called the AWOL smoother in Kastner & Frühwirth-Schnatter (2014). The algorithm exploits the fact that the full conditional distribution of $\theta_{0:T}$ is Gaussian with a block tridiagonal precision matrix in order to quickly compute its Cholesky decomposition. McCausland, Miller & Pelletier (2011) improves the idea by implicitly computing this Cholesky decomposition through a backward sampling strategy by mixing the substeps of the factorization and sampling steps – hence the name – starting with sampling from the marginal distribution of θ_T.

Suppose our model is as follows:

$$y_t = F_t\theta_t + v_t$$

$$\theta_t = G_t\theta_{t-1} + w_t$$

with $v_t \overset{ind}{\sim} N(0, V_t)$ independent of $w_t \overset{ind}{\sim} N(0, W_t)$ for $t = 1, 2, \cdots, T$ and $\theta_0 \sim N(m_0, C_0)$. This is the usual DLM except now we allow for time dependent variances for illustrative purposes. Then $(y_{1:T}, \theta_{0:T})$ is joint Gaussian conditional on $(V_{1:T}, W_{1:T})$ (in this section, everything is conditional on $V_{1:T}$ and $W_{1:T}$, so we will not make this conditioning explicit). So we can write $p(\theta_{0:T}|y_{1:T})$ as

$$\log p(\theta_{0:T}|y_{1:T}) = -\frac{1}{2}g(\theta_{0:T}, y_{1:T}) + K$$

where K is some constant with respect to $\theta_{0:T}$ and

$$g(\theta_{0:T}, y_{1:T}) = \theta'_{0:T}\Omega\theta_{0:T} - 2\omega'\theta_{0:T}.$$

However, we also have

$$\log p(\theta_{0:T}|y_{1:T}) = \log p(\theta_{0:T}, y_{1:T}) - \log p(y_{1:T}).$$

This means that

$$\begin{aligned} g(\theta_{0:T}, y_{1:T}) &= (\theta_0 - m_0)C_0^{-1}(\theta_0 - m_0) + K' \\ &+ \sum_{t=1}^{T}(y_t - F_t\theta_t)'V_t^{-1}(y_t - F_t\theta_t) \\ &+ \sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})'W_t^{-1}(\theta_t - G_t\theta_{t-1}). \end{aligned}$$

where K' is another constant that does not depend on $\theta_{0:T}$.

So now we can identify blocks of Ω with the cross product terms of the θ_t's and blocks of ω with the single product terms. Specifically, Ω is a banded diagonal matrix with

$$\Omega = \begin{bmatrix} \Omega_{00} & \Omega_{01} & 0 & \ddots & 0 & 0 \\ \Omega_{10} & \Omega_{11} & \Omega_{12} & \ddots & 0 & 0 \\ 0 & \Omega_{21} & \Omega_{22} & \ddots & 0 & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \ddots & \Omega_{T-1,T-1} & \Omega_{T-1,T} \\ 0 & 0 & 0 & \ddots & \Omega_{T,T-1} & \Omega_{TT} \end{bmatrix}$$

and $\omega = (\omega_0', \omega_1', \cdots, \omega_T')'$ where the Ω_{st}'s and ω_t's defined below:

$$\begin{aligned}
\Omega_{00} &= C_0^{-1} + G_1' W_1^{-1} G_1 \\
\Omega_{tt} &= F_t' V_t^{-1} F_t + W_t^{-1} + G_{t+1}' W_{t+1}^{-1} G_{t+1} && \text{for } \; t = 1, 2, \cdots T-1 \\
\Omega_{TT} &= F_T' V_T^{-1} F_T + W_T^{-1} \\
\Omega_{t,t-1} &= -W_t^{-1} G_t && \text{for } \; t = 1, 2, \cdots T \\
\Omega_{t-1,t} &= -G_t' W_t^{-1} = \Omega_{t,t-1}' && \text{for } \; t = 1, 2, \cdots T \\
\omega_0 &= C_0^{-1} m_0 \\
\omega_t &= F_t' V_t^{-1} y_t && \text{for } \; t = 1, 2, \cdots T.
\end{aligned}$$

Together, Ω and a determine the Gaussian distribution from which $\theta_{0:T}$ should be drawn. Rue (2001) shows how to take advantage of the sparsity of Ω in order to quickly compute its Cholesky factorization and in order to find the mean vector from ω and this factorization. McCausland et al. (2011) shows that instead of computing these quantities directly, you can draw θ_T and $\theta_t | \theta_{t+1:T}$ iteratively, which ultimately reduces the number of linear algebra operations which must be performed and typically speeds up the computation despite taking advantage of essentially the same mathematical technology.

The resulting algorithm requires a couple more intermediate quantities. Let $\Sigma_0 = \Omega_{00}^{-1}$, $\Sigma_t = (\Omega_{tt} - \Omega_{t,t-1}\Sigma_{t-1}\Omega_{t-1,t})^{-1}$ for $t = 1, 2, \cdots, T$, $h_0 = \Sigma_0 \omega_0$, and $h_t = \Sigma_t(\omega_t - \Omega_{t,t-1} h_{t-1})$ for $t = 1, 2, \cdots, T$. Then

$$\begin{aligned}
\theta_T &\sim N(h_T, \Sigma_T) \\
\theta_{t|t+1:T} &\sim N(h_t - \Sigma_t \Omega_{t,t+1} \theta_{t+1}, \Sigma_t) && \text{for } \; t = T-1, T-2, \cdots, 0.
\end{aligned}$$

McCausland et al. (2011) shows how to quickly compute the required linear algebra operations and finds that this method is often faster than simply doing the Cholesky factorization. This algorithm can also be applied to drawing the scaled errors, $\psi_{0:T}$, and the wrongly-scaled errors, $\tilde{\psi}_{0:T}$.

F. FURTHER AUGMENTATION FOR NON-INVERTIBLE F_T

Throughout the paper we assumed that F_t is square and invertible for all t which made the construction of the SE sampler and other samplers that use the scaled errors easier. However, most DLMs do not have

F_t's which are square, let alone invertible. The samplers we constructed can often still be used in this case with one tweak: an additional DA is required in order to ensure that F_t is square and invertible for all t. The basic strategy is to add elements to y_t until F_t is invertible, then add an additional step to the sampler in order to draw the new augmentation. A second issue is that often G_t or F_t or both depend on some unknown parameter which must also be sampled from in the various MCMC samplers. The second case is easily dealt with simply by adding another sampling step for the unknown parameters in F_t and G_t. The following example illustrates how to deal with the first case. See Simpson (2015) for another example.

Consider the dynamic regression model

$$y_t = \alpha_t + x_t\beta_t + v_t$$

$$\alpha_t = \alpha_{t-1} + w_{1,t}$$

$$\beta_t = \beta_{t-1} + w_{2,t}$$

for $t = 1, 2, \cdots, T$ with $v_{1:T}$ independent of $w_{1:T} = (w_1', w_2', \cdots, w_T')'$ where $w_t = (w_{1,t}, w_{2,t})'$, $v_t \overset{iid}{\sim} N(0, V)$ and $w_t \overset{iid}{\sim} N_2(0, W)$. Here the latent state in period t is $\theta_t = (\alpha_t, \beta_t)'$. The problem is that $F_t = [1, x_t]$ is neither square nor invertible. But notice that the matrix

$$F_t^* = \begin{bmatrix} 1 & x_t \\ 0 & 1 \end{bmatrix}$$

is invertible. Now we add an additional DA z_t to y_t to construct $y_t^* = (y_t, z_t)'$ so that now the model is

$$y_t^* = F_t^*\theta_t + v_t^*$$

$$\theta_t = \theta_{t-1} + w_t$$

where $v_t^* = (v_t, u_t)$ where $u_{1:T}$ is independent of $(v_{1:T}, w_{1:T})$ and $u_t \overset{iid}{\sim} N(0, 1)$. By construction $v_t^* \overset{iid}{\sim} N_2(0, V^*)$ where V^* is a diagonal matrix with the vector $(V, 1)$ along the diagonal and the full conditional distribution of z_t is $N(\beta_t, 1)$. Then we define the scaled errors as $\psi_0 = \theta_0$ and $\psi_t = L_{V^*}^{-1}(y_t^* - F_t^*\theta_t)$. Let $z = z_{1:T}$ and $y^* = y_{1:T}^*$ for brevity.

In terms of θ, the likelihood is

$$\begin{aligned} p(y,z,\theta|V,W) \propto & |V^*|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t^* - F_t^*\theta_t)'(V^*)^{-1}(y_t^* - F_t^*\theta_t)\right] \\ & \times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1})\right] \\ \propto & V^{-T/2} \exp\left[-\frac{1}{2V}\sum_{t=1}^{T}(y_t - \alpha_t - x_t\beta_t)^2\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{t}(z_t - \beta_t)^2\right] \\ & \times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - \theta_{t-1})'W^{-1}(\theta_t - \theta_{t-1})\right] \end{aligned}$$

Then by transforming to ψ, the back transformation is $\theta_t = (F_t^*)^{-1}(y_t^* - L_{V^*}\psi_t)$ so the Jacobian is block diagonal with T copies of $(F_t^*)^{-1}L_{V^*}$ along with a single copy of the identity matrix along the diagonal. So the determinant of the Jacobian is $|J| = |V^*|^{T/2}$ and the likelihood can be written in terms of ψ as

$$p(y,z,\theta|V,W) \propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t'\psi_t\right] |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t^* - \mu_t)'(F_t^*W(F_t^*)')^{-1}(y_t^* - \mu_t)\right]. \quad \text{(A.10)}$$

where we define $\mu_1 = L_{V^*}\psi_1 + F_1^*\psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = L_{V^*}\psi_t + F_t^*(F_{t-1}^*)^{-1}(y_{t-1}^* - L_{V^*}\psi_{t-1})$.

Now in order to construct a sampler that uses ψ, we simply add a new step to sampler to draw z from its full conditional just before transforming to ψ. In the GIS and alternating algorithms, we now have to draw an updated z every time we change the DA. When using the states, $z_t|V,W,\theta,y \overset{iid}{\sim} N(\beta_t, 1)$, so it is easiest to transform to θ before drawing z. So for example in the SD-SE GIS sampler with V, W, α_0, and β_0 independent in the prior, an $IG(\alpha_V, \beta_V)$ prior on V, and an $IW(\Lambda_W, \lambda_W)$ prior on W, the algorithm becomes

Algorithm: SD-SE GIS for dynamic regression. *Scaled Disturbance-Scaled Error GIS Sampler for the dynamic regression model*

1. *Use the MCFA to sample* $\theta \sim p(\theta|V,W,y)$.
2. *Sample* $V \sim IG\left(\alpha_V + T/2, \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \alpha_t - \beta_t)^2\right)$.
3. *Transform* θ *to* γ.
4. *Sample* $W \sim p(W|V,\gamma,y)$.

5. *Transform* γ *to* θ.

6. *Sample* $z_t \overset{iid}{\sim} N(\beta_t, 1)$ *and form* y^*.

7. *Transform* θ *to* ψ.

8. *Sample* $V \sim p(V|W, \psi, y^*)$.

9. *Sample* $W \sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right)$.

Step 8 is particularly tricky since V is a component of V^*, and V^* has the same density $p(V|W, \psi, y)$ that shows up in the usual case of the scaled disturbances, except now the lower right diagonal element is set to one. So while we can write down the various algorithms in the non-invertible F case, the density $p(V|W, \psi, y^*)$ is tricky to work with. In step 8 V is drawn conditional on y^*, but another option is to draw V conditional on y but not on z. This would require integrating z out of the likelihood, equation (A.10). It is not clear which of these is easier or faster, though it is likely that the changing the prior for V and W will have an impact.

G. EFFICIENTLY DRAWING FROM $P(W|V, \gamma, Y)$ AND $P(V|W, \psi, Y)$ IN THE LLM

From Appendix D.2, the full conditional distribution of W given γ is

$$p(W|V,\gamma,y) \propto p(V,W,\gamma,y) \propto |W|^{-(\lambda_W+p+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\sum_{t=1}^{T}\left[y_t - F_t\sum_{s=0}^{t}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]' V^{-1}\left[y_t - F_t\sum_{s=0}^{t-1}\tilde{G}_{s,t}L_W\gamma_{t-s} - F_t\tilde{G}_{t,t}\gamma_0\right]\right)\right]$$

where L_W is the Cholesky factor of W defined so that $L_W L_W' = W$. We can write this density as

$$p(W|V,\gamma,y) \propto |W|^{-(\lambda_W+p+2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right]$$
$$\times \exp\left[-\frac{1}{2}\left(\operatorname{vec}(L_W)' A_W \operatorname{vec}(L_W) - 2B_W \operatorname{vec}(L_W)\right)\right]$$

where

$$A_W = \sum_{t=1}^{T}\sum_{s=0}^{t}\left(\gamma_{t-s}\gamma_{t-s}' \otimes \tilde{G}_{s,t}' F_t' V^{-1} F_t \tilde{G}_{s,t}\right)$$

and

$$B_W = \sum_{t=1}^{T}\sum_{s=0}^{t}\left(\gamma'_{t-s} \otimes (y_t - F_t\tilde{G}_{t,t}\gamma_0)'V^{-1}F_t\tilde{G}_{s,t}\right)$$

can be found using the properties of the vec and tr operators.

Similarly from Appendix D.3, the full conditional distribution of V given ψ is

$$p(V|W,\psi,y) \propto p(V,W,\psi,y) \propto |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\left(\operatorname{tr}\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}(y_t-\mu_t)'(F_tWF_t')^{-1}(y_t-\mu_t)\right)\right]$$

where $\mu_1 = L_V\psi_1 + F_1G_1\psi_0$ and for $t = 2,3,\cdots,T$, $\mu_t = L_V\psi_t + F_tG_tF_{t-1}^{-1}(y_{t-1} - L_V\psi_{t-1})$. This density can be written in a familiar form:

$$\begin{aligned} p(V|W,\psi,y) \propto p(V,W,\psi,y) \propto |V|^{-(\lambda_V+p+2)/2}\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \\ \times \exp\left[-\frac{1}{2}\left(\operatorname{vec}(L_V)'A_V\operatorname{vec}(L_V) - 2B_V\operatorname{vec}(L_V)\right)\right] \end{aligned}$$

where

$$\begin{aligned} A_V =& \sum_{t=1}^{T}\psi_t\psi_t' \otimes (F_tWF_t')^{-1} + \sum_{t=2}^{T}\psi_{t-1}\psi_{t-1}' \otimes (G_tF_{t-1}^{-1})'W^{-1}G_tF_{t-1}^{-1} \\ & - \sum_{t=2}^{T}\psi_t\psi_{t-1}' \otimes (WF_t')^{-1}G_tF_{t-1}^{-1} - \sum_{t=2}^{T}\psi_{t-1}\psi_t' \otimes (G_tF_{t-1}^{-1})'(F_tW)^{-1} \end{aligned}$$

and

$$\begin{aligned} B_V =& \psi_1' \otimes (y_1 + F_1G_1\psi_0)'(F_1WF_1')^{-1} + \sum_{t=2}^{T}\psi_t' \otimes (y_t - F_tG_tF_{t-1}^{-1}y_{t-1})'(F_tWF_t')^{-1} \\ & - \sum_{t=2}^{T}\psi_{t-1}' \otimes (y_t - F_tG_tF_{t-1}^{-1}y_{t-1})'(WF_t')^{-1}G_tF_{t-1}^{-1} \end{aligned}$$

can again be found using the properties of the vec and tr operators. Both of these densities are of the form

$$p(X) \propto |X|^{-(\lambda+p+2)/2}\exp\left[-\frac{1}{2}\left(\operatorname{tr}(\Lambda X^{-1}) + \operatorname{vec}(L_X)'A\operatorname{vec}(L_X) - 2B\operatorname{vec}(L_X)\right)\right]$$

where X is a $p \times p$ symmetric and positive definite random matrix, L_X is the Cholesky factor of X so that $L_XL_X' = X$, $\lambda > 0$, Λ is a $p \times p$ symmetric and positive definite matrix, A is a $p^2 \times p^2$ matrix, and B is a $1 \times p^2$ matrix.

The complexity of this density is caused by the interaction between the inverse Wishart prior and the augmented data likelihood in terms of the scaled disturbances for W or for the scaled errors for V. In the local level model, the density still is not a known form and is difficult to sample from, but sampling from it is possible. In this case the log density is

$$\log p(x) = -(\alpha+1)\log x - ax + b\sqrt{x} - c/x + C$$

for $x > 0$ where C is some constant, $\alpha > 0$ and $c > 0$ are the hyperparameters for x, and $a > 0$ and $b \in \Re$ are parameters that depend on the data, y, the relevant data augmentation (ψ or γ), and the other variable (W or V). We provide two different rejection sampling strategies below that work well under different circumstances, and combine them into a single strategy.

G.1 Adaptive rejection sampling

One nice strategy is to use adaptive rejection sampling, e.g. Gilks & Wild (1992). This requires $\log p(x)$ to be concave, which is easy enough to check. The second derivative of $\log p(x)$ is:

$$\frac{\partial^2 \log p(x)}{\partial x^2} = -\frac{1}{4}bx^{-3/2} + (\alpha+1)x^{-2} - 2cx^{-3}.$$

Then we have

$$\frac{\partial^2 \log p(x)}{\partial x^2} < 0 \qquad \Longleftrightarrow \qquad -\frac{b}{4}x^{3/2} + (\alpha+1)x - 2c < 0$$

which would imply that $\log p(x)$ is concave. We can maximize the left hand side of the last equation very easily. When $b \leq 0$ the max occurs at $x = \infty$ such that $LHS > 0$, but when $b > 0$:

$$\frac{\partial LHS}{\partial x} = -\frac{3}{8}bx^{1/2} + \alpha + 1 = 0 \qquad \Longrightarrow \qquad x^{max} = \frac{(\alpha+1)^2}{b^2}\frac{64}{9}.$$

Then we have

$$LHS \leq LHS|_{x=x^{max}} = \frac{(\alpha+1)^3}{b^2}\frac{64}{27} - 2c$$

so that

$$LHS|_{x=x^{max}} < 0 \iff \frac{(\alpha+1)^3}{b^2}\frac{64}{27} < 2c \qquad \Longleftrightarrow \qquad b > \left(\frac{(\alpha+1)^3}{c}\right)^{1/2}\frac{4\sqrt{2}}{3\sqrt{3}}.$$

This last condition is necessary and sufficient for $\log p(x)$ to be globally (for $x > 0$) concave since $b < 0$ forces $LHS > 0$ for some x. When the condition is satisfied, we can use adaptive rejection sampling — which is already implemented in the `R` package `ars` (Rodriguez 2009). We input the initial evaluations of $\log p(x)$ at the mode x^{mode} and at $2x^{mode}$ and $0.5x^{mode}$ in order to get the algorithm going.

G.2 Rejection sampling on the log scale

When $b \leq \left(\frac{(\alpha+1)^3}{c}\right)^{1/2} \frac{4\sqrt{2}}{3\sqrt{3}}$, which happens often — especially for small T — we need to rely on a different method to sample from $p(x)$. A naive approach would be to construct a normal or t approximation to $p(x)$ and use that as a proposal in a rejection sampler. It turns out that this is often very inefficient, but for $z = \log(x)$ the approach works well. Note that

$$p_z(z) = p_x(e^z)e^z$$

so that we can write the log density of z as (dropping the subscripts):

$$\log p(z) = -ae^z + be^{z/2} - \alpha z - ce^{-z}.$$

The mode of this density z^{mode} can be easily found numerically, and the second derivative is:

$$\frac{\partial^2 \log p(z)}{\partial z^2} = -ae^z + \frac{b}{4}e^{z/2} - ce^{-z}.$$

The t approximation then uses the proposal distribution p

$$t_v\left(z^{mode}, \left[-\left.\frac{\partial^2 \log p(z)}{\partial z^2}\right|_{z=z^{mode}}\right]^{-1}\right).$$

In practice choosing degrees of freedom $v = 1$ works very well over the region of the parameter space where adaptive rejection sampling cannot be used. We can easily use this method when adaptive rejection sampling does not work, then transform z back to x. It remains to check that the tails of t distribution dominate the tails of our target distribution. Let $\log q(z)$ denote the log density of the proposal distribution. Then we need

$$\log p(z) - \log q(z) \leq M$$

for some constant M, i.e.

$$-ae^{z} + be^{z/2} - \alpha z - ce^{-z} - \left(\frac{v+1}{2}\right)\log\left[1 + \frac{1}{v}\left(\frac{z-\mu}{\sigma}\right)^2\right] \leq M$$

where $a > 0$, $c > 0$, $\alpha > 0$, $v > 0$, $\sigma > 0$, and $b, \mu \in \Re$. We can rewrite the LHS as

$$e^{z/2}(b - ae^{z/2}) - \alpha z - ce^{-z} - \left(\frac{v+1}{2}\right)\log\left[1 + \frac{1}{v}\left(\frac{z-\mu}{\sigma}\right)^2\right].$$

So as $z \to \infty$ this quantity goes to $-\infty$ since the first term will eventually become negative no matter the value of b, and all other terms are always negative. Similarly as $z \to -\infty$ this quantity goes to $-\infty$. Now pick any interval (z_1, z_2) such that outside of the interval, $LHS < \epsilon$. Since treated as a function of z the LHS is clearly continuous, it attains a maximum on this interval, and thus is bounded.

G.3 Intelligently choosing a rejection sampler

In practice, adaptive rejection sampling is relatively efficient for $p_x(x)$ but inefficient for $p_z(z)$ — so much so that rejection sampling with the t approximation for $p_z(z)$ is more efficient. To minimize computation time, it is best to use adaptive rejection sampling for $p_x(x)$ when the concavity condition is satisfied. When it is not, the t approximation works well enough.

H. EFFICIENTLY DRAWING FROM $P(W|V, \tilde{\gamma}, Y)$ AND $P(V|W, \tilde{\psi}, Y)$ IN THE LLM

Both the density of $\log(W)|V, \tilde{\gamma}, y$ and the density of $\log(V)|W, \tilde{\psi}, y$ have the following form:

$$p(z) \propto \exp\left[-\alpha z - ae^{-z} + be^{-z/2} - ce^{z}\right].$$

where $\alpha > 0$, $a > 0$, $c > 0$, and $b \in \Re$. The log density is:

$$\log p(z) = -\alpha z - ae^{-z} + be^{-z/2} - ce^{z} + C$$

where C is some constant. We only provide one strategy for rejection sampling from this density: the t approximation. Similar reasoning to the previous subsection above shows that we can use a t distribution as a proposal in a rejection sampler for this density. Now we choose the location parameter by maximizing $\log p(z)$ in z numerically to find the mode, z^{mode}. Next the second derivative of $\log p(z)$ is given by

$$\frac{\partial^2 \log p(z)}{\partial z^2} = -ae^{-z} + \frac{b}{4}e^{-z/2} - ce^{z}.$$

We then set the scale parameter to be

$$-\left[\left.\frac{\partial^2 \log p(z)}{\partial z^2}\right|_{z=z^{mode}}\right]^{-1}$$

as in the normal approximation, and the degrees of freedom parameter to $v = 1$. This rejection sampler is tolerably efficient for our purposes, but there is much room for improvement.

I. EQUIVALENCE OF CIS AND GIS IN THE DLM

The CIS algorithm consists of the following steps:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\tilde{\psi}|V^{(k+0.5)}, W^{(k)}, \psi] \to [V^{(k+1)}|W^{(k)}, \tilde{\psi}] \to$$

$$[\tilde{\gamma}|V^{(k+1)}, W^{(k)}, \tilde{\psi}] \to [W^{(k+0.5)}|V^{(k+1)}, \tilde{\gamma}] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \tilde{\gamma}] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

In the fourth step of line one and the second step of line two, each of those densities would be unchanged if we conditioned on θ instead of $\tilde{\psi}$ on the first line or $\tilde{\gamma}$ on the second line. So the CIS algorithm above is equivalent to the following:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\theta|V^{(k+0.5)}, W^{(k)}, \psi] \to [V^{(k+1)}|W^{(k)}, \theta] \to$$

$$[W^{(k+0.5)}|V^{(k+1)}, \theta] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

Now since V and W are conditionally independent given θ and y, the last step of line one and the first step of line 2 can be switched:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [\theta|V^{(k+0.5)}, W^{(k)}, \psi] \to [W^{(k+0.5)}|V^{(k+0.5)}, \theta] \to$$

$$[V^{(k+1)}|W^{(k+0.5)}, \theta] \to [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

Next V's conditional density is the same whether we condition on θ or γ, so we can do the V step between the γ step and the W step in line two. Similarly we can move the W step to between the V step and the θ step in line one. This yields:

$$[\psi|V^{(k)}, W^{(k)}] \to [V^{(k+0.5)}|W^{(k)}, \psi] \to [W^{(k+0.5)}|V^{(k+0.5)}, \psi] \to$$

$$[\gamma|V^{(k+0.5)}, W^{(k+0.5)}, \psi] \to [V^{(k+1)}|W^{(k+0.5)}, \gamma] \to [W^{(k+1)}|V^{(k+1)}, \gamma].$$

This is actually a SE-SD GIS algorithm, so the CIS sampler we started with is equivalent to SE-SD GIS. Since we do not expect the order in which the DAs appear in a GIS algorithm to matter, CIS should have the same mixing and convergence properties as the SD-SE GIS algorithm we constructed.

J. PARTIAL CIS ALGORITHMS IN THE DLM

In addition to the GIS and CIS algorithms discussed in the main body of the article, Yu & Meng (2011) also introduce *partial CIS* algorithms. While a CIS algorithm interweaves in separate Gibbs steps for each sub-vector of the parameter, a partial CIS algorithm has a usual Gibbs step for at least one of the parameter vectors. For example, suppose that the model parameter is $\phi = (\phi_1, \phi_2)$, and γ_1, γ_2, and θ are available DAs. Then a partial CIS algorithm using these DAs is

Algorithm: partial CIS. *Partial Componentwise Interweaving Strategy*

$$[\gamma_1|\phi_1^{(k)}, \phi_2^{(k)}] \quad \rightarrow \quad [\phi_1^{(k+0.5)}|\phi_2^{(k)}, \gamma_1] \quad \rightarrow \quad [\gamma_2|\phi_1^{(k+0.5)}, \phi_2^{(k)}, \gamma_1] \quad \rightarrow \quad [\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_2] \quad \rightarrow$$
$$[\theta|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_2] \quad \rightarrow \quad [\phi_2^{(k+1)}|\phi_1^{(k+1)}, \theta].$$

The first line is an interweaving step for ϕ_1 while the second line is a standard Gibbs step for ϕ_2. Partial CIS algorithms are easier to construct than full CIS algorithms at the cost of slower convergence (Yu & Meng 2011).

In the DLM we can construct two partial CIS algorithms using the wrongly-scaled DAs in much the same way they were used to construct the full CIS algorithm. The first algorithm interweaves for W using the scaled disturbances, γ, and the wrongly-scaled disturbances, $\tilde{\gamma}$:

$$[\theta|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}|W^{(k)}, \theta] \rightarrow$$
$$[W^{(k+0.5)}|V^{(k+1)}, \theta] \rightarrow [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].$$

As in the construction of the full CIS algorithm, we use θ instead of $\tilde{\gamma}$ in the second line since $p(W|V, \tilde{\gamma}) = p(W|V, \theta)$. Using an argument similar to that used in Appendix I, we can show that this partial CIS algorithm is equivalent to the SD-State GIS algorithm.

Analogously, we can use the scaled errors, ψ, and the wrongly-scaled errors, $\tilde{\psi}$, to construct a partial CIS

algorithm that interweaves for V:

$$[\psi|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+0.5)}|W^{(k)}, \psi] \rightarrow [\theta|V^{(k+0.5)}, W^{(k)}, \psi] \rightarrow [V^{(k+1)}|W^{(k)}, \theta] \rightarrow$$

$$[W^{(k+1)}|V^{(k+1)}, \theta].$$

This algorithm is equivalent to the SE-State GIS algorithm.

K. USING POSTERIOR CORRELATIONS TO UNDERSTAND PATTERNS OF ESP

Most of the patterns in Figures M.3, M.4, and M.5 in the next section can be explained by Figure K.1, which contains the estimated posterior correlations between various functions of parameters estimated using the simulations from the Triple-Alt sampler for a time series with $T = 100$. We omit a similar analysis for $T = 10$ and $T = 1000$. The state sampler consists of two steps — a draw of θ given V and W, and a draw of (V, W) given θ. From Section D.1 we have that conditional on θ, V and W are independent in the posterior and each has an inverse gamma distribution that depends on the states only through the second parameter:

$$b_V \equiv \beta_V + \sum_{t=1}^{T} (y_t - \theta_t)^2/2 \qquad\qquad b_W \equiv \beta_W + \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^2/2.$$

So we can view (b_V, b_W) as the data augmentation instead of θ and thus the state sampler is

$$[b_V, b_W|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+1)}, W^{(k+1)}|b_V, b_W].$$

Thus the dependence between (V, W) and (b_V, b_W) in the posterior will determine how much the state sampler moves in a given iteration and, in particular, it is possible that V and W have very different serial dependence from each other since we are drawing them jointly. When the dependence between V and b_V is high, the (V, W) step will hardly move V even if it drastically moves W since V and W are independent. However, the (b_V, b_W) step may move both elements a moderate amount since they both depend on (V, W).

In Figure K.1 we see that the posterior correlation between V and b_V is high in magnitude and positive when $R^* > 1$ while the posterior correlation between V and b_W is moderate to low and negative. When R^* is large enough though, the posterior correlation between V and b_W evaporates. Similarly when $R^* < 1$ the posterior correlation between W and b_W is high and positive and the posterior correlation between W and

b_V is high and negative. Again as R^* becomes large enough the correlation between W and b_V goes to zero. So when $R^* > 1$, the draw of (b_V, b_W) is unlikely to move b_V much since b_V is so highly correlated with V and essentially uncorrelated with b_W, but b_W is essentially uncorrelated with W and negatively correlated with V so b_W is likely to move a fair amount. Furthermore the draw of V is highly correlated with b_V while the draw of W is essentially independent of b_W (and the draws of V and W are independent conditional on b_V and b_W). Thus when $R^* > 1$ we should expect high serial dependence for V and low serial dependence for W, and so low ESP for V and high ESP for W, which is exactly what we see in Figure M.4. By similar reasoning when $R^* < 1$, we should expect low serial dependence for V and high serial dependence for W and thus high ESP for V and low ESP for W, which can also be seen in Figure M.4.

For the SD sampler, things are a bit more complicated. The draw of $V|W, \gamma$ still depends on b_V since it is the same inverse gamma draw as in the state sampler, but the draw of $W|V, \gamma$ now depends on a_γ and b_γ defined in Section D.2 as

$$a_\gamma \equiv \frac{1}{2V} \sum_{t=1}^{T} \left(\sum_{j=1}^{t} \gamma_j \right)^2 \qquad\qquad b_\gamma \equiv \frac{1}{V} \sum_{t=1}^{T} (y_t - \gamma_0) \left(\sum_{j=1}^{t} \gamma_j \right).$$

So the dependence between V and b_V determines how much the chain moves in the V step, and the dependence between W and (a_γ, b_γ) determines how much it moves in the W step. The dependence between (V, W) and γ determines how much the chain moves in the DA step, but we can view this step instead as a draw of b_V in which case the dependence between W and b_V determines how much the chain moves in that step. So if any one of these steps has high dependence, we should expect every element of the chain, and (V, W) in particular, to have high serial dependence in the chain. The SE sampler is analogous to the SD sampler except with b_W, a_ψ and b_ψ where

$$a_\psi = \frac{1}{2W} \sum_{t=1}^{T} (\mathcal{L}\psi_t)^2 \qquad\qquad b_\psi = \frac{1}{W} \sum_{t=1}^{T} (\mathcal{L}\psi_t \mathcal{L}y_t).$$

In order to analyze the SD sampler, first suppose $R^* > 1$. Then from Figure K.1 b_V has high correlation with V and low correlation with W, so the draw of b_V should not move the chain much. Next, the draw of V should again not move the chain much because of the high correlation between V and b_V. Finally the draw of W has a fair chance to move the chain because it has low correlation with both a_γ and b_γ. But this has

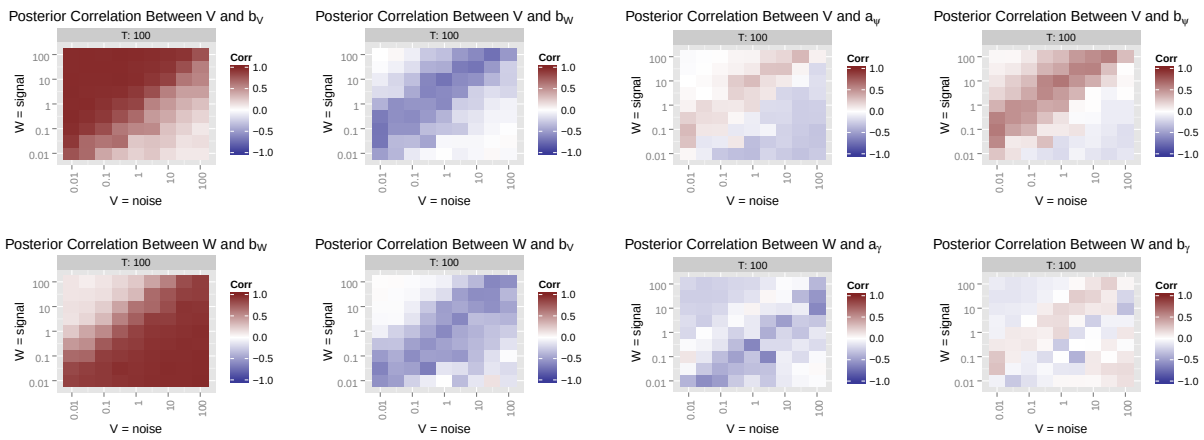
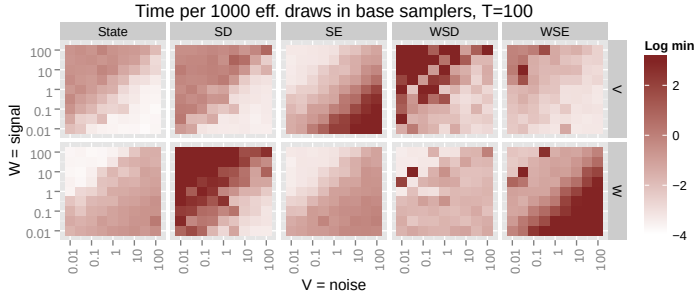


Figure K.1: Posterior correlation between V or W and b_V, b_W, a_γ, b_γ, a_ψ or b_ψ. X and Y axes indicate the true values of V and W respectively for the simulated data with $T = 100$.

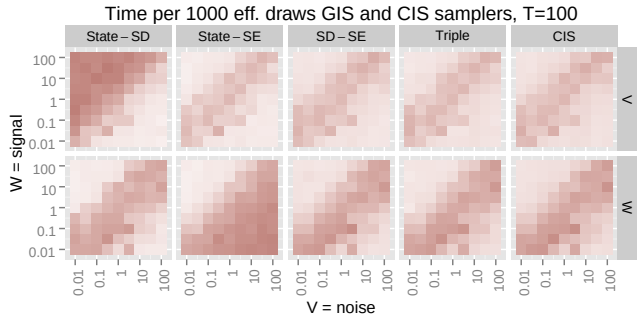
little impact on b_V and thus the entire chain since b_V is so highly correlated with V but hardly correlated with W. So when $R^* > 1$, we should expect high serial dependence and low ESP for V. We should also expect similar behavior for W since the entire chain is hardly moving so W's hyperparameters are hardly moving. This is roughly what we see in Figure M.4, though this reasoning does not allow us to predict which of V and W will have lower ESP. When $R^* < 1$ the posterior correlation in each of the steps is broken, though in the W step the correlation between W and both a_γ and b_γ becomes negative and somewhat high in magnitude. Here we should not expect less serial dependence in V or W, but we should perhaps expect higher ESP's since negatively correlated draws decrease Monte Carlo standard error. Indeed, we see ESP's near one for both variances in Figure M.4. The SE sampler is analogous to the SD sampler and a similar analysis applies — the posterior correlations between V or W and b_W, a_ψ or b_ψ in Figure K.1 roughly predict the ESP of the SE sampler in Figure M.4. When one or more of the correlations are high, ESPs for V and W are low while when all of the correlations are low, both ESPs are high. We omit a similar analysis of the wrongly-scaled samplers for brevity, but note that their behavior will allows us to predict the behavior of the CIS sampler.

L. COMPUTATIONAL TIME

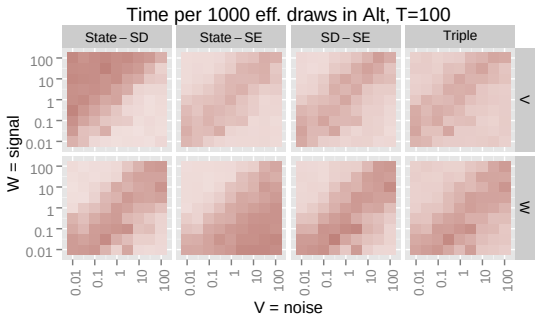
From a practical standpoint a more important question than how well the chain mixes is the full computational time required to adequately characterize the target posterior distribution. In order to investigate this, we compute the natural log of the average time in minutes required for each sampler to achieve an effective sample size of 1000 — in other words the log minutes per 1000 effective draws. All simulations were performed on a server with Intel Xeon X5675 3.07 GHz processors. While different systems will yield different absolute times, the relative times should be similar. Figure L.2 contains plots of the log minutes per 1000 effective draws for both V and W and for each of the samplers.



(a)



(b)

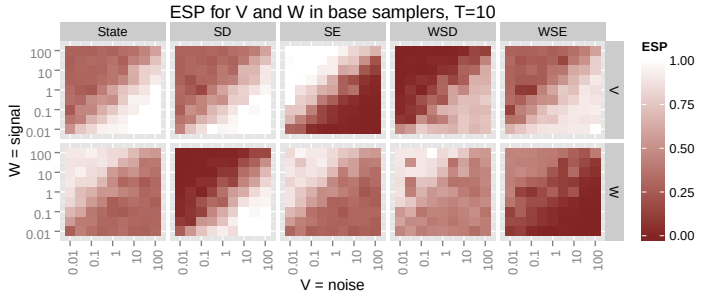


(c)

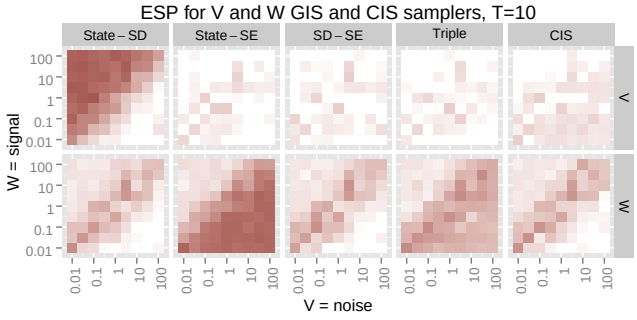
Figure L.2: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 100$ in each sampler. Figure L.2a contains the base samplers, Figure L.2b contains the GIS and CIS samplers, while Figure L.2c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.

For $T = 100$ the pattern we saw for ESP also appears for log minutes per 1000 effective draws. The State sampler becomes slow to reach 1000 effective draws for V when $R^* > 1$ and for W when $R^* < 1$. The SD and SE samplers behave as expected — the SD sampler is slow for both V and W when $R^* > 1$ while the SD sampler is slow for both V and W when $R^* < 1$. The SD-SE GIS, Triple GIS and CIS algorithms appear to be the big winners here and are almost indistinguishable. All three algorithms are slightly slower for both V and W when R^* is near one, though for larger T, when R^* is near or below one all three are slow for W (plots available in Appendix M). Compared to the state sampler, all three offer large gains over most of the parameter space. There appears to be no difference between a GIS algorithm and the corresponding alternating algorithm in terms of log time per 1000 effective draws, so the SD-SE Alt and Triple Alt algorithms are both just as efficient as the best interweaving algorithms. This may not always be the case though — the GIS version of an algorithm is computationally cheaper than the Alt version since it consists of three of the four same steps, and in the fourth step the Alt algorithm has to obtain a random draw while the GIS algorithm typically only has to make a transformation. The more expensive that draw is relative to the transformation, the faster GIS will be relative to Alt. For example, in a longer the time series the transformation will be significantly cheaper relative to the random draw. We find that With very long time series, e.g. $T > 100,000$, GIS is cheaper than Alt even in the local level model (Appendix N).

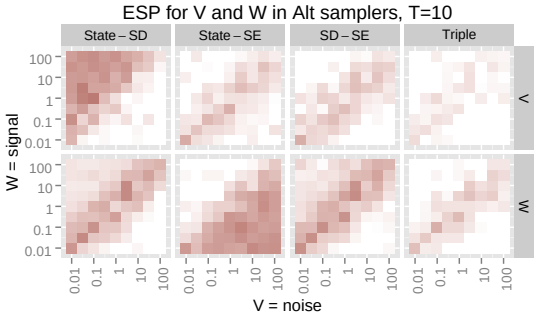
M. PLOTS FOR ALL VALUES OF T



(a)

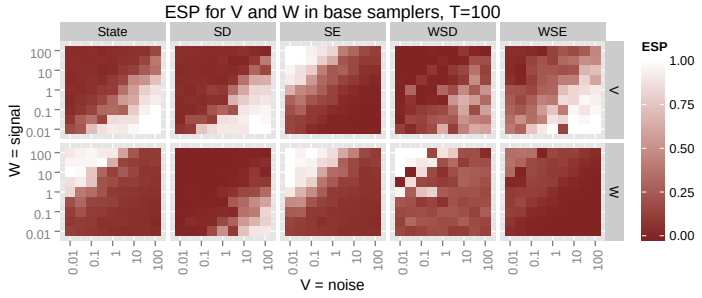


(b)

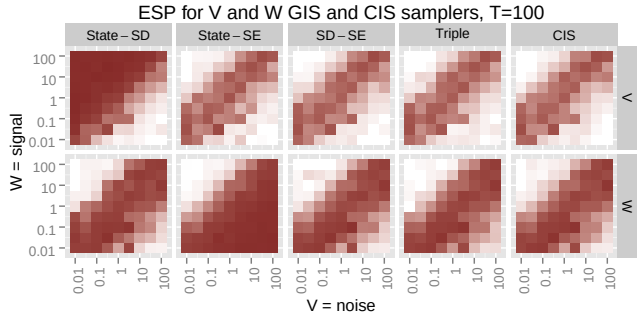


(c)

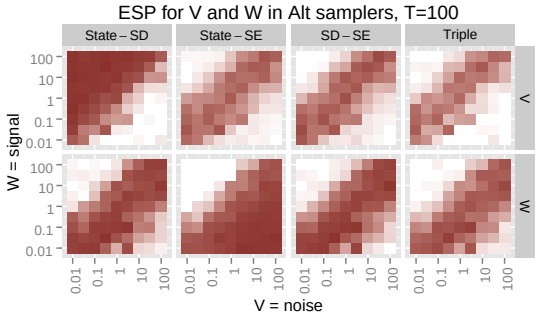
Figure M.3: Effective sample proportion in the posterior sampler for a time series of length $T = 10$, for V and W in the each sampler. Figure M.3a contains ESP for V and W for the base samplers, Figure M.3b contains ESP in the GIS and CIS samplers, and Figure M.3c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.



(a)

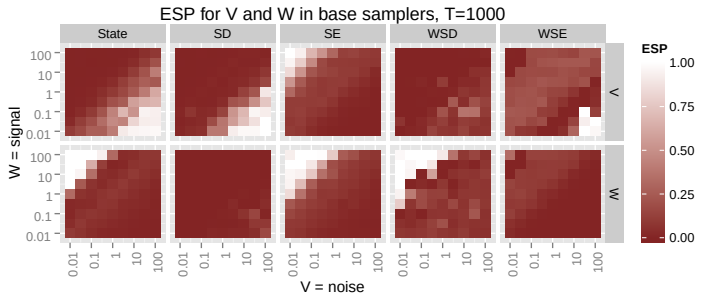


(b)

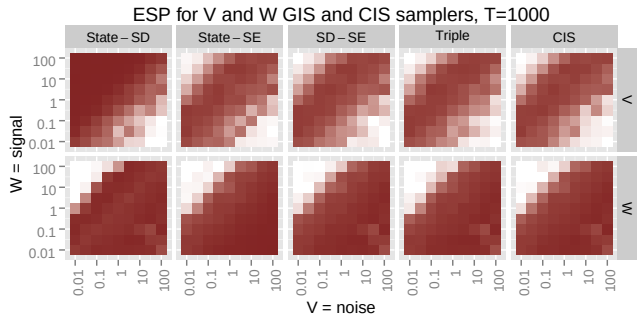


(c)

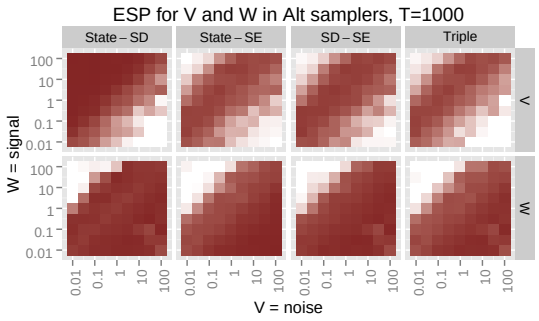
Figure M.4: Effective sample proportion in the posterior sampler for a time series of length $T = 100$, for V and W in the each sampler. Figure M.4a contains ESP for V and W for the base samplers, Figure M.4b contains ESP in the GIS and CIS samplers, and Figure M.4c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.



(a)

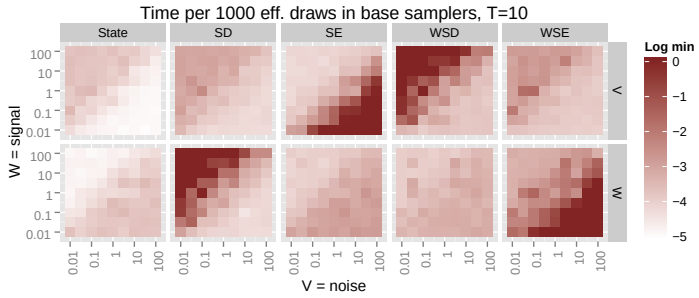


(b)

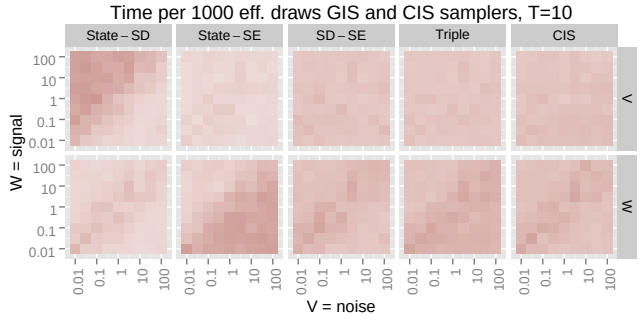


(c)

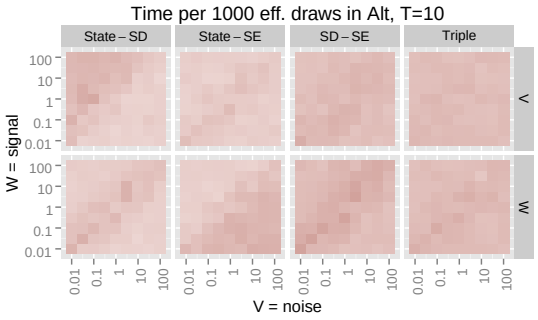
Figure M.5: Effective sample proportion in the posterior sampler for a time series of length $T = 1000$, for V and W in the each sampler. Figure M.5a contains ESP for V and W for the base samplers, Figure M.5b contains ESP in the GIS and CIS samplers, and Figure M.5c contains ESP in the Alt samplers. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.



(a)

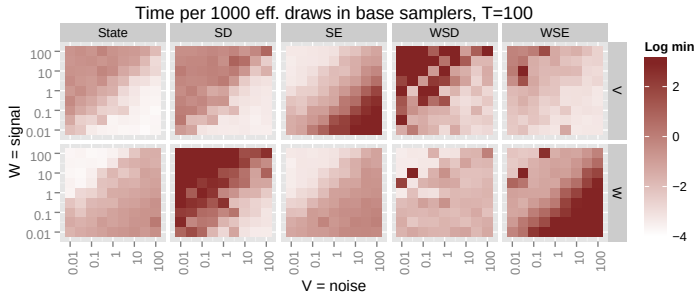


(b)

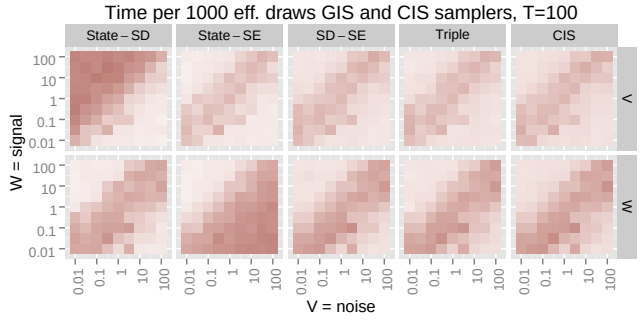


(c)

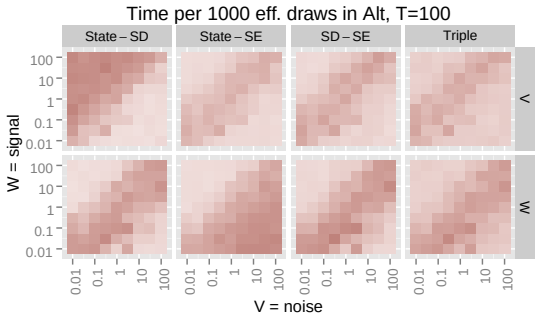
Figure M.6: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 10$ in each sampler. Figure M.6a contains the base samplers, Figure M.6b contains the GIS and CIS samplers, while Figure M.6c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.



(a)

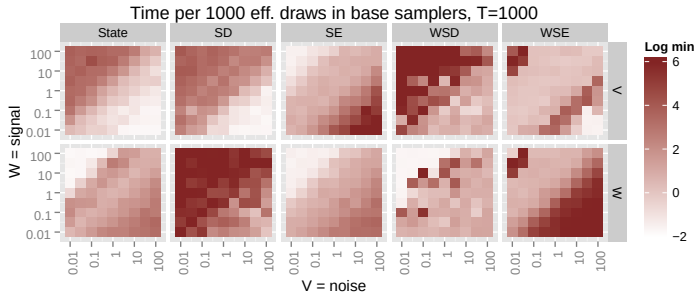


(b)

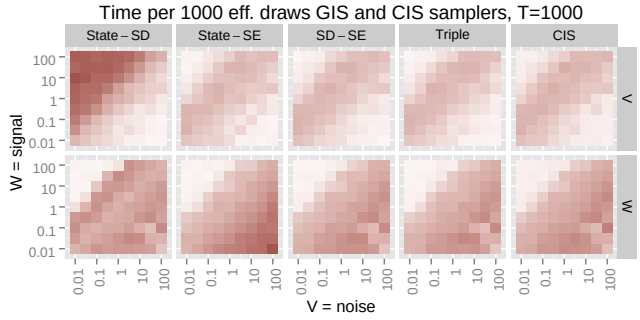


(c)

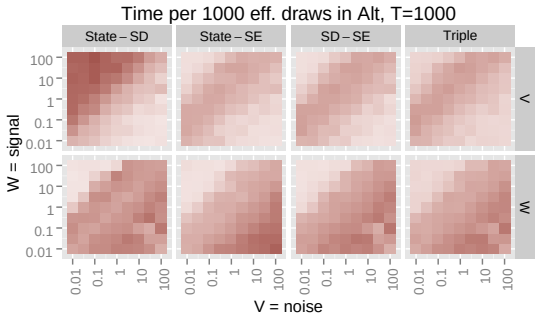
Figure M.7: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 100$ in each sampler. Figure M.7a contains the base samplers, Figure M.7b contains the GIS and CIS samplers, while Figure M.7c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.



(a)



(b)



(c)

Figure M.8: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 1000$ in each sampler. Figure M.8a contains the base samplers, Figure M.8b contains the GIS and CIS samplers, while Figure M.8c contains the Alt samplers. Log times larger than three log min are rounded down to three for plotting purposes.

N. GIS VS. ALT IN VERY LONG TIME SERIES

While our results indicate that GIS and the corresponding Alternating algorithms seem to perform equally well in terms of both mixing and computational time, this is not always the case. GIS is computationally cheaper per iteration and in long enough time series this different is significant. To illustrate this we again simulated data from the local level model with $V = 1$, $W = 1$, for various lengths of the time series starting at $T = 1000$ and increasing to $T = 500,000$. Then we fit each model using the same priors as before using the SD-SE GIS and SD-SE Alt algorithms. Figure N.9 contains plots of the time in minutes per 1000 effective draws for each of V and W in each sampler.

For both samplers the relationship between the length of the time series and the time per 1000 effective draws appears linear. However, for the alternating sampler the time required increases at a fast rate as the T increases. For example with a time series of 300,000 observations, the SD-SE GIS sampler requires about 1000 minutes (16.67 hours) to achieve an effective sample size of 1000 for W and about about 500 minutes (8.33 hours) for V. On the other hand the SD-SE Alt sampler requires 2000 minutes (33.33 hours) to achieve an effective sample size of 1000 for W and 1000 minutes for V. Theses differences do not add up to much when the time series is short enough – e.g. $T = 1000$ and below, but when T is on the order of $100,000$ the benefit of GIS starts to become significant.

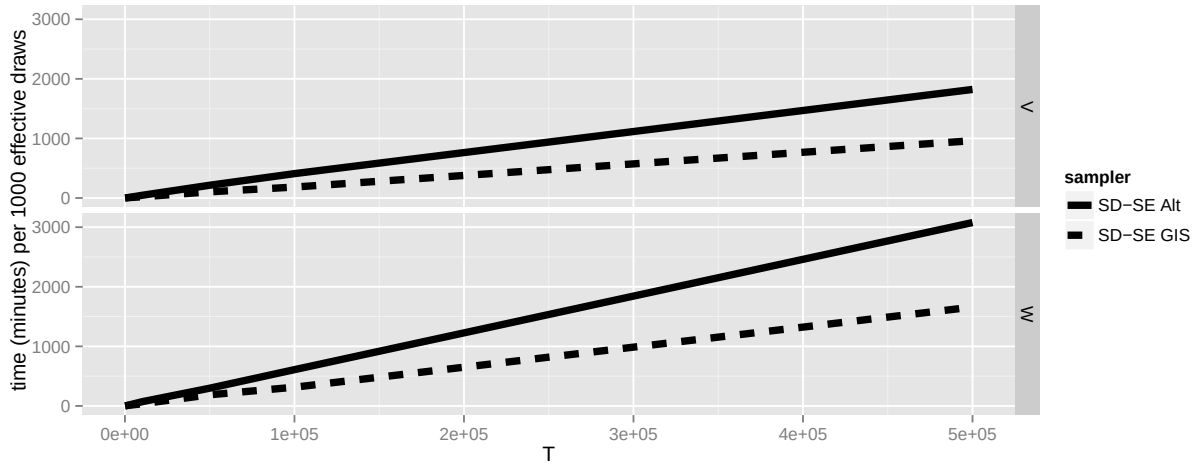


Figure N.9: Time in minutes per 1000 effective draws in the posterior sampler for V and W, for the SD-SE GIS and SD-SE Alt samplers in very long time series.

REFERENCES

Carter, C. K., & Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81(3), 541–553.

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15(2), 183–202.

Gilks, W. R., & Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41(2), 337–348.

Kastner, G., & Frühwirth-Schnatter, S. (2014), "Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of Stochastic Volatility Models," *Computational Statistics & Data Analysis*, 76, 408–423.

McCausland, W. J., Miller, S., & Pelletier, D. (2011), "Simulation Smoothing for State–Space models: A Computational Efficiency Analysis," *Computational Statistics & Data Analysis*, 55(1), 199–212.

Rodriguez, P. P. (2009), *ars: Adaptive Rejection Sampling*. R package version 0.4, original C++ code from Arnost Komarek based on ars.f written by P. Wild and W. R. Gilks.
URL: *http://CRAN.R-project.org/package=ars*

Rue, H. (2001), "Fast Sampling of Gaussian Markov Random Fields," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 325–338.

Simpson, M. (2015), "Application of Interweaving in DLMs to an Exchange and Specialization Experiment," in *Bayesian Statistics from Methods to Models and Applications: Research from BAYSM 2014*, eds. S. Frühwirth-Schnatter, A. Bitto, G. Kastner, & A. Posekany, Springer.

Yu, Y., & Meng, X.-L. (2011), "To Center or not to Center: That is not the Question - An Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency," *Journal of Computational and Graphical Statistics*, 20(3), 531–570.