Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

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

Matthew Simpson, Jarad Niemi & Vivekananda Roy

To cite this article: Matthew Simpson, Jarad Niemi & Vivekananda Roy (2017) Interweaving Markov Chain Monte Carlo Strategies for Efficient Estimation of Dynamic Linear Models, Journal of Computational and Graphical Statistics, 26:1, 152-159, DOI: 10.1080/10618600.2015.1105748

To link to this article: https://doi.org/10.1080/10618600.2015.1105748

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 26, NO. 1, 152–159
http://dx.doi.org/10.1080/10618600.2015.1105748

Taylor & Francis
Taylor & Francis Group

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

Matthew Simpson[a], Jarad Niemi[b], and Vivekananda Roy[b]

[a]Department of Statistics, University of Missouri–Columbia, Columbia, Missouri; [b]Department of Statistics, Iowa State University, Ames, Iowa

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 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 online for this article.

ARTICLE HISTORY
Received September 2014
Revised July 2015

KEY WORDS
Ancillary augmentation; Centered parameterization; Data augmentation; Noncentered parameterization; Reparameterization; State-space model; Sufficient augmentation; Time series

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, Sahu and Carlin 1995; Roberts and Sahu 1997; Meng and Van Dyk 1998; Van Dyk and Meng 2001; Bernardo et al. 2003; Papaspiliopoulos, Roberts, and Sköld 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), that is, 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, Papaspiliopoulos, and Dellaportas 2004; Bos and Shephard 2006; Strickland, Martin, and Forbes 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 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, 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 article 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.

CONTACT Matthew Simpson ✉ *themattsimpson@gmail.com*

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JCGS*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

2. Variations of Data Augmentation

Let $p(\phi|y)$ be a probability density, for example, 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, Roberts, and Sköld 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), that is, $\gamma = M(\theta; \phi, y)$ where $M(.; \phi, y)$ is one-to-one, and thus $[\gamma|\theta, \phi]$ is deterministic. The key difference between the eGIS 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 noncentered augmentation or noncentered 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. In fact, Yu and Meng (2011) showed 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.

Algorithm: [CIS]. Componentwise Interweaving Strategy

$$\begin{aligned}&\left[\theta_1|\phi_1^{(k)}, \phi_2^{(k)}\right] \rightarrow \left[\gamma_1|\phi_2^{(k)}, \theta_1\right] \rightarrow \left[\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_1\right]\\ &\rightarrow \left[\theta_2|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_1\right] \rightarrow \left[\gamma_2|\phi_1^{(k+1)}, \theta_2\right]\\ &\rightarrow \left[\phi_2^{(k+1)}|\phi_1^{(k+1)}, \gamma_2\right].\end{aligned}$$

Here 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 and third rows do 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, Campagnoli and Petrone 2009; Prado and West 2010) and is defined as

$$\begin{aligned}y_t &= F_t\theta_t + v_t \qquad v_t \overset{\text{ind}}{\sim} N_k(0, V_t) \qquad \text{(observation equation)}\\ \theta_t &= G_t\theta_{t-1} + w_t \qquad w_t \overset{\text{ind}}{\sim} N_p(0, W_t) \qquad \text{(system equation)},\end{aligned}$$

where $N_d(\mu, \Sigma)$ is a d-dimensional multivariate normal distribution with mean μ and covariance Σ. The observation errors, $v_{1:T} \equiv (v_1', v_2', \dots, v_T')'$, and the system disturbances, $w_{1:T} \equiv (w_1', w_2', \dots, w_T')'$ are independent. The observed data are $y \equiv y_{1:T} \equiv (y_1', y_2', \dots, y_T')'$, while the latent states are $\theta \equiv \theta_{0:T} \equiv$

$(\theta_0', \theta_1', \dots, \theta_T')'$. For each $t = 1, 2, \dots, T$, 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

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

for $t = 1, 2, \dots, T$. We assume the conditionally conjugate priors: θ_0, V, and W independent with $\theta_0 \sim N_p(m_0, C_0)$, $V \sim \text{IW}(\Lambda_V, \lambda_V)$ and $W \sim \text{IW}(\Lambda_W, \lambda_W)$ where m_0, C_0, Λ_V, λ_V, Λ_W, and λ_W are known hyperparameters and $\text{IW}(\Lambda, \lambda)$ 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 y:

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

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

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

$\tilde{m}_{Tp\times 1} = (m_0', m_0', \dots m_0')'$. 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

To construct an ASIS algorithm, we need to find an SA and an AA for the DLM. Papaspiliopoulos, Roberts, and Sköld (2007) noted 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 (7), 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

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

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 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, that is, the lower triangular matrix L_W such that $L_W L_W' = W$. Then, we will define the SDs, $\gamma \equiv \gamma_{0:T} \equiv (\gamma_0', \gamma_1', \dots, \gamma_T')'$, by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \dots, T$. There are actually $p!$ different versions of the SDs depending on how we order the elements of θ_t, but we use 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 $t = 1, 2, \dots, T$. Using the SDs, the model is

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

for $t = 1, 2, \dots, T$ 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), that is, $v_t = y_t - F_t\theta_t$ 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 $t = 1, 2, \dots, T$ 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 $t = 1, 2, \dots, T$ 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 $t = 2, 3, \dots, T$. Then, we can write the model as

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

for $t = 1, 2, \dots, T$, 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 $t = 1, 2, \dots, T$ 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

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

for $t = 1, 2, \ldots, T$ 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

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

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

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

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, Miller and Pelletier 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

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

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, for example, Frühwirth-Schnatter (2004) and Section 6.

Next, we can use γ 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

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

It is easy to show that $V|W, \gamma, y \sim IW(\Lambda_V + \sum_{t=1}^{T} v_tv_t', \lambda_V + T)$ 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

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

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 p. 153. For example, the *State-SD alternating sampler* obtains the $k+1$'st iteration of (V, W) from the kth as follows:

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

The first line is an iteration of the state sampler while the second and third lines are 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 p. 153. For example, the *State-SD GIS sampler* is:

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

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

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

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 p. 153. 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:

$$\begin{aligned}&[\psi|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+0.5)}|W^{(k)}, \psi]\\ &\quad\rightarrow [\tilde{\psi}|V^{(k+0.5)}, W^{(k)}, \psi] \rightarrow [V^{(k+1)}|W^{(k)}, \tilde{\psi}]\\ &\quad\rightarrow [\tilde{\gamma}|V^{(k+1)}, W^{(k)}, \tilde{\psi}] \rightarrow [W^{(k+0.5)}|V^{(k+1)}, \tilde{\gamma}]\\ &\quad\rightarrow [\gamma|V^{(k+1)}, W^{(k+0.5)}, \tilde{\gamma}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].\end{aligned}$$

The first two lines are essentially a Gibbs step for drawing V that interweaves between ψ and $\tilde{\psi}$, while the last two lines are 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 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:

$$\begin{aligned}&[\psi|V^{(k)}, W^{(k)}] \rightarrow [V^{(k+0.5)}|W^{(k)}, \psi]\\ &\quad\rightarrow [\psi|V^{(k+0.5)}, W^{(k)}, \theta] \rightarrow [V^{(k+1)}|W^{(k)}, \theta]\\ &\quad\rightarrow [W^{(k+0.5)}|V^{(k+1)}, \theta] \rightarrow [\gamma|V^{(k+1)}, W^{(k+0.5)}, \theta]\\ &\quad\rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma].\end{aligned}$$

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

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

for $t = 1, 2, \ldots, T$. The priors on (θ_0, V, W) from Section 3 become $\theta_0 \sim N(m_0, C_0)$, $V \sim \mathrm{IG}(\alpha_V, \beta_V)$ and $W \sim \mathrm{IG}(\alpha_W, \beta_W)$ with θ_0, V, and W mutually independent, where $\mathrm{IG}(\alpha, \beta)$ 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 $t = 1, 2, \ldots, T$. Similarly, the SEs are $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \ldots, T$. 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 $t = 1, 2, \ldots, T$.

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, and 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: $\theta_0 \sim N(0, 10^7)$, $V \sim \text{IG}(5, 4V^*)$, and $W \sim \text{IG}(5, 4W^*)$, 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 ($\text{ESP} = \text{ESS}/n$). When $\text{ESP} = 1$ the chain is behaving as if it obtains iid draws from the posterior. Occasionally $\text{ESP} > 1$, if the draws are negatively correlated, but we round it down to one in our plots.

6.3 Simulation Results

Figure 1(a) 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.

We fit the model using several interweaving (GIS and CIS) samplers as well. Since the 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 1(b) has plots of ESP for each of the GIS and CIS algorithms while Figure 1(c) 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 narrowband 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

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 and 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) found that the signal-to-noise ratio with the AR(1) coefficient determine the convergence rate of a Gibbs sampler.

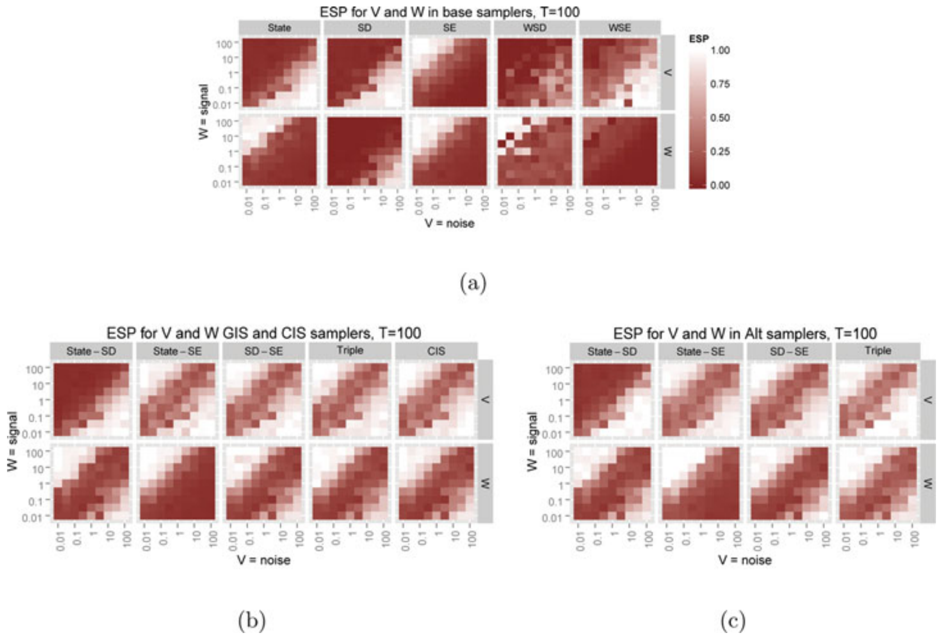

Figure 1. Effective sample proportion in the posterior sampler for a time series of length $T = 100$ for V and W in the base sampler (a), GIS and CIS samplers (b), and Alt samplers (c). The axes indicate the true values of V (horizontal) and W (vertical) for the simulated data. The signal-to-noise ratio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

When Frühwirth-Schnatter (2004) studied 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, for example, 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

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, $R^* = W^*/V^*$.

	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	$R^*>1$	$R^*<1$	$R^*>1$	$R^*>1$	$R^*>1$	$R^* \not\approx 1$	$R^*>1$	$R^* \not\approx 1$	$R^* \not\approx 1$	$R^* \not\approx 1$

NOTE: The right side of the table applies to both the interweaving and alternating algorithms.

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.

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)

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 article.

References

Basu, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhyā: The Indian Journal of Statistics*, 15, 377–380. [153]

Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., 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. [152]

Bos, C. S., and Shephard, N. (2006), "Inference for Adaptive Time Series Models: Stochastic Volatility and Conditionally Gaussian State Space Form," *Econometric Reviews*, 25, 219–244. [152]

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541–553. [152,155]

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, 1–38. [152]

Frühwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202. [152,155]

——— (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. [152,154,155,158]

Frühwirth-Schnatter, S., 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, New York: Springer, pp. 480–485. [152]

——— (2008), "Bayesian Estimation of the Multi-Factor Heston Stochastic Volatility Model," *Communications in Dependability and Quality Management*, 11, 5–25. [152]

Frühwirth-Schnatter, S., and Tüchler, R. (2008), "Bayesian Parsimonious Covariance Estimation for Hierarchical Linear Mixed Models," *Statistics and Computing*, 18, 1–13. [158]

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, 827–841. [152]

——— (2010), "Stochastic Model Specification Search for Gaussian and Partial Non-Gaussian State Space models," *Journal of Econometrics*, 154, 85–100. [158]

——— (2011), "Bayesian Variable Selection for Random Intercept Modeling of Gaussian and Non-Gaussian Data," in *Bayesian Statistics Vol. 9*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, Oxford: Oxford University Press, pp. 165–200. [158]

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parametrisations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488. [152]

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)," *Bayesian Analysis*, 1, 515–534. [158]

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. [157]

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, 532–554. [152]

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. [152,155]

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [152,153]

McCausland, W. J., Miller, S., and Pelletier, D. (2011), "Simulation Smoothing for State–Space models: A Computational Efficiency Analysis," *Computational Statistics & Data Analysis*, 55, 199–212. [155]

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, 59, 511–567. [152]

——— (1998), "Fast EM-type Implementations for Mixed Effects Models, *Journal of the Royal Statistical Society*, Series B, 60, 559–578. [152]

——— (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320. [152]

Papaspiliopoulos, O., and Roberts, G. O. (2008), "Stability of the Gibbs Sampler for Bayesian Hierarchical Models," *The Annals of Statistics*, 36, 95–117. [152]

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007), "A General Framework for the Parametrization of Hierarchical Models," *Statistical Science*, 22, 59–73. [152,153,154]

Petris, G., Campagnoli, P., and Petrone, S. (2009), *Dynamic Linear Models with R*, New York: Springer. [153]

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, 63–85. [152,157]

Prado, R., and West, M. (2010), *Time Series: Modeling, Computation, and Inference*, London: CRC Press. [153]

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, 66, 369–393. [152]

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*, 59, 291–317. [152]

Shephard, N. (1996), *Statistical Aspects of ARCH and Stochastic Volatility*, London: Springer. [152]

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, New York: Springer. [154]

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, 2911–2930. [152]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [152]

Van Dyk, D., and Meng, X.-L. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50. [152,158]

Van Dyk, D. A., and Tang, R. (2003), "The One-Step-Late PXEM Algorithm," *Statistics and Computing*, 13, 137–152. [152]

West, M., and Harrison, J. (1999), *Bayesian Forecasting & Dynamic Models* (2nd ed.), New York: Springer. [153]

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, 531–570. [152,153,156,157]