Ancillarity-Sufficiency or not; Interweaving to Improve MCMC Estimation of the Local Level DLM

Matt Simpson

April 16, 2014

Abstract

In dynamic linear models (DLMs), MCMC sampling can often be very slow for estimating the posterior density — especially for longer time series. In particular, in some regions of the parameter space the standard data augmentation algorithm can mix very slowly. Recently ancillarity-sufficiency interweaving has been introduced as a method to take advantage of alternate parameterizations in multilevel models in order to improve the mixing and convergence properties of the chain. Focusing on the local level DLM, we explore alternate parameterizations and various interweaving algorithms through simulation in order to improve mixing. We conclude by explaining what our results may mean for MCMC in a more general DLM.

1 Model

The general dynamic linear model (DLM) is a linear, gaussian, state space model and can be written as

$$y_t = F_t\theta_t + v_t \qquad v_t \overset{ind}{\sim} N_k(0, V_t) \tag{1}$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \overset{ind}{\sim} N_p(0, W_t) \tag{2}$$

for $t = 1, 2, \cdots, T$, and $v_{1:T}$, $w_{1:T}$ independent. Equation (1) is called the *observation equation* and equation (2) is called the *system equation*. Similarly, $v_{1:T}$ are called the observation errors, $V_{1:T}$ are called the observation variances, $w_{1:T}$ are called the system disturbances and $W_{1:T}$ are called the system variances. The observed data is $y_{1:T}$ while $\theta_{0:T}$ are called the latent states. For each $t = 1, 2, \cdots, T$, F_t is a $k \times p$ matrix and G_t is a $p \times p$ matrix. Let ϕ denote the vector of unknown parameters in the model. Then possibly $F_{1:T}$, $G_{1:T}$, $V_{1:T}$, and $W_{1:T}$ are all functions of ϕ.

We will focus our attention on a simple version of the DLM. Typically additional model structure is used to learn about $V_{1:T}$ and $W_{1:T}$ if time dependence is enforced – e.g. a stochastic volatility prior which would require a statespace model describing the $V_{1:T}$'s and $W_{1:T}$'s as data. Because of this additional complexity, we focus on the time-constant variances model, though many of our results may be useful in more complicated time-varying variance models. So we set $V_t = V$ and $W_t = W$ for $t = 1, 2, \cdots, T$. We will also suppose that F_t and G_t are known matrices for $t = 1, 2, \cdots, T$, though this constraint is immaterial since relaxing it will simply add one or more Gibbs steps to the algorithms we explore so long as no parameter that enters any F_t or G_t also enters V or W. Note, however, that in one of the data augmentations that we discuss, the scaled error data augmentation, there is a bit more housekeeping associated with $F_{1:T}$ depending on an unknown parameter (Section 2.2).

When $\phi = (V, W)$ is our unkown parameter vector and we can write the model as

$$y_t|\theta_{0:T} \overset{ind}{\sim} N(F_t\theta_t, V) \tag{3}$$

$$\theta_t|\theta_{0:t-1} \sim N(G_t\theta_{t-1}, W) \tag{4}$$

To complete the model specification in a Bayesian context, we need priors on θ_0, V, and W. We'll use the standard approach and assume that they're mutually independent a priori and that $\theta_0 \sim N(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ 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. This allows us to write the full joint distribution of $(V, W, \theta_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \theta_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\theta_0 - m_0)' C_0^{-1}(\theta_0 - m_0)\right] \\ &\times |V|^{-(\lambda_V + k + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(y_t - F_t\theta_t)' V^{-1}(y_t - F_t\theta_t)\right] \\ &\times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\theta_t - G_t\theta_{t-1})' W^{-1}(\theta_t - G_t\theta_{t-1})\right] \end{aligned} \tag{5}$$

where $p = dim(\theta_t)$, $k = dim(y_t)$, and tr(.) is the matrix trace operator.

2 Estimating the Model via Data Augmentation: Parameterization Issues

A well known method to estimate the DLM is via data augmentation (DA) often using forward filtering backward sampling (FFBS), as in **?** and **?**. The basic idea is to implement a Gibbs sampler with two blocks. The generic DA algorithm with parameter ϕ, augmented data θ, and data y obtains the $k+1$'st state of the Markov chain, $\phi^{(k+1)}$, from the k'th state, $\phi^{(k)}$ as follows:

Algorithm 1.

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

The first block runs a simulation smoother which draws the latent states from their conditional posterior distribution given the model parameters. This can be accomplished in a number of ways. FFBS is the original method proposed by **?** and **?**, but alternatives include **?**, **?** and **?**. The second block draws $\phi = (V, W)$ from their joint conditional posterior which in this model turns out to be independent inverse Wishart distributions. In particular

$$\begin{aligned} V|\theta_{0:T}, y_{1:T} &\sim IW\left(\Lambda_V + \sum_{t=1}^{T} v_t v_t', \lambda_V + T\right) \\ W|\theta_{0:T}, y_{1:T} &\sim IW\left(\Lambda_W + \sum_{t=1}^{T} w_t w_t', \lambda_W + T\right) \end{aligned}$$

where $v_t = y_t - F_t\theta_t$ and $w_t = \theta_t - G_t\theta_{t-1}$. We are calling this algorithm the *state sampler*.

The main problem with the state sampler is that in some regions of the parameter space the Markov chain mixes poorly for some of the parameters. For example, in the univariate local level model ($F_t = G_t = 1$ for $t = 1, 2, \cdots, T$) and similar models it is known that if the time constant variance of the latent states, W, is too small, mixing will be poor for W **?**.

One well known method of improving mixing and convergence in MCMC samplers is reparameterizing the model. **?** is a good summary. Most of the work in some way focuses on what are called centered and noncentered parameterizations. In our general notation where ϕ is the parameter, θ is the DA and y is the data, the parameterization (ϕ, θ) is a *centered parameterization* (CP) if $p(y|\theta, \phi) = p(y|\theta)$. The parameterization is a *noncentered parameterization* (NCP) if $p(\theta|\phi) = p(\theta)$. When (ϕ, θ) is a CP, θ is called a *centered augmentation* (CA) for ϕ and when (ϕ, θ) is a NCP, θ is called a *noncentered augmentation* (NCA)

for ϕ. A centered augmentation is sometimes called a *sufficient augmentation* (SA) and a noncentered augmentation is sometimes called an *ancillary augmentation* (AA), e.g. in **?**. Like **?**, we prefer the latter terminology because it immediately suggests the intuiton that a sufficient augmentation is like a sufficient statistic while an ancillary augmentation is like an ancillary statistic.

The key reasoning behind the emphasis on SAs and AAs is that typically when the DA algorithm based on the SA has nice mixing and convergence properties the DA algorithm based on the AA has poor mixing and convergence propeties and vice versa. In other words, the two algorithms form a "beauty and the beast" pair. This property suggests that there might be some way to combine the two DA algorithms or the two underlying parameterizations in order to construct a sampler which has "good enough" properties all the time. **?** for example suggest alternating between the two augmentations within a Gibbs sampler. Some work focuses on using partially noncentered parameterizations that are a sort of bridge between the CP and NCP, e.g. **?** for general hierarchical models and **?** in the context of a partiuclar DLM — a dynamic univarite regression with a stationary AR(1) coefficient.

Another method of combining the two DAs is through what **?** call interweaving. The idea is pretty simple: suppose that ϕ denotes the parameter vector, θ denotes one augmented data vector, γ denotes another augmented data vector, and y denotes the data. Then an MCMC algorithm that *interweaves* between θ and γ performs the following steps in a single iteration to obtain the $k+1$'st draw, $\phi^{(k+1)}$, from the k'th draw, $\phi^{(k)}$:

Algorithm 2.

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

Notice that an additional step is added to algorithm 1, and the final step now draws ϕ conditional on γ instead of θ. This is the intuition behind the name "interweaving"—the draw of the second augmented data vector is weaved in between the draws of θ and ϕ. This particular method of interweaving is called a *global* interweaving strategy (GIS) since interweaving occurs globally across the entire parameter vector. It's possible to define a *componentwise* interweaving strategy (CIS) that interweaves within specific steps of a Gibbs sampler as well. Step two of the GIS algorithm is typically accomplished by sampling $\phi|\theta, y$ and then $\gamma|\theta, \phi, y$. 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 a one-to-one function. In this case, the algorithm becomes:

Algorithm 3.

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

When γ is a one-to-one transformation of θ, step 4 is an update $\gamma = M(\theta; \phi, y)$. The GIS algorithm is directly comparable to the *alternating* algorithm suggested by **?**. Given the same two DAs, θ and γ, and parameter vector ϕ, the alternating algorithm for sampling from $p(\phi|y)$ is as follows:

Algorithm 4.

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

The key difference between this algorithm and algorithm 3 is in step 3: instead of drawing from $p(\gamma|\theta, \phi, y)$, the alternating algorithm draws from $p(\gamma|\phi, y)$. In other words it alternates between two data augmentation algorithms in a single iteration. The interweaving algorithm, on the other hand, connects or "weaves" the two separate iterations together in step 3 by drawing γ conditonal on θ in addition to ϕ and y.

? call a GIS approach where one of the DAs is a SA and the other is an AA an *ancillary sufficient interweaving strategy*, or an ASIS. They show that the GIS algorithm has a geometric rate of convergence no worse than the worst of the two underlying algorithms and in some cases better than the the corresponding alternating algorithm. In particular, their Theorem 1 suggests that the weaker the dependence between two data augumentations in the posterior, the more efficient the GIS algorithm. In the limit of a posteriori indepedent data augmentations, the GIS algorithm will even obtain iid draws from the posterior of the model

parameter. This motivates their focus on ASIS — conditional on the model parameter, a SA and an AA are independent, which suggests that the dependence between the two DAs will be limited in the posterior. In fact, when the prior on ϕ is nice in some sense, **?** show that the ASIS algorithm is the same as the optimal PX-DA algorithm of **?**, **?**, **?** and **?**. Their results suggest that ASIS and interweaving generally is a promising approach to improve the speed of MCMC in a variety of models no matter what region of the parameter space the posterior is concentrated.

To gain some intuition about why interweaving works, recall that a typical problem with slow MCMC is that there is high autocorrelation in the Markov chain for ϕ, $\{\phi^{(k)}\}_{k=1}^{K}$, leading to imprecise estimates of $\mathrm{E}[f(\phi)]$ for some function f. Our goal is to reduce this dependence. In the usual DA algorithm, e.g. algorithm 1, when ϕ and θ are highly dependent in the joint posterior, the draws from $p(\theta|\phi, y)$ and then from $p(\phi|\theta, y)$ will hardly move the chain which results in high autocorrelation. Interweaving helps break this autocorrelation in two ways. First, by inserting the extra step, e.g. steps 2 and 3 together in 3, the chain gets an additional chance to move in a single iteration thereby weakening the autocorrelation. This is a feature of an alternating algorithm as well, but **?** show that the corresponding interweaving algorithm is often even more efficient. The key is the second point — when the posterior dependence between the two DAs is low, steps 2 and 3 in 3, i.e. step 2 in 2, is enough to almost completely break the dependence in the chain. For the alternating algorithm, it is typically not feasible to find a data augmentation such that step 2 or step 3 of 4 completely breaks the depedence in the chain — this would require finding a DA such that the model parameter and the DA are essentially independent which, in turn, would likely mean that drawing from the conditional posterior of the parameter given the DA is nearly as difficult as drawing from the marginal posterior of the model parameter.

Aside from the intuition of finding a posteriori (nearly) independent DAs, both alternating and interweaving strategies suggest looking for a "beauty and the beast" pair of DAs — specifically both algorithms will tend to do better, all else equal, when the two underlying DA algorithms are efficient in opposite regions of the parameter space. In other words, when on DA algorithm does poorly (is a "beast") the other does well (is a "beauty").

2.1 The scaled disturbances

The next step is to apply the ideas of interweaving to sampling from the posterior of the dynamic linear model. **?** note that typically the usual parameterization results in a SA for the parameter ϕ. All that's necessary for an ASIS algorithm, then, is to construct an AA for ϕ. We immediately run into a problem because the standard DA for a DLM is the latent states $\theta_{0:T}$. From equations (3) and (4) we see that V is in the observation equation so that $\theta_{0:T}$ is not a SA for (V, W) while W is in the system equation so that $\theta_{0:T}$ is not an AA for (V, W) either. In order to find a SA we need to somehow move V from the observation equation (3) to the system equation (4) while leaving W in the system equation. We also need to find an AA by somehow moving W from the system equation to the observation equation while leaving V in the observation equation. A naive thing to try is to condition on the disturbances instead of the states and see if the disturbances for a SA or an AA for (V, W). The disturbances $w_{0:T}$ are defined by $w_t = \theta_t - G_t\theta_{t-1}$ for $t = 1, , 2, \cdots, T$ and and $w_0 = \theta_0$. However the DA algorithm based on the w_t's is identical to the algorithm based on the θ_t because it turns out that the conditional distributions $p(V, W|\theta_{0:T}, y_{1:T})$ and $p(V, W|w_{0:T}, y_{1:T})$ are identical.

? suggest that in order to obtain an ancillary augmentation for a variance parameter, we must scale the sufficient agumentation by the square root of that parameter. Based on this intuition, note that if we hold V constant then $\theta_{0:T}$ is a SA for W from the observation and system equations, (3) and (4), i.e. we say $\theta_{0:T}$ is a SA for W given V, or for $W|V$. Similarly $\theta_{0:T}$ is an AA for $V|W$. This suggests that if we scale θ_t by W appropriately for all t we'll have an ancillary augmentation for V and W jointly. The same intuition suggests scaling $w_t = \theta_t - G_t\theta_{t-1}$ by W appropriately for all t in order to find an ancillary augmentation for (V, W). We will work with the latter case since it has already been used in the literature, but the two are not the same even in the simplest DLMs.

To define the scaled disturbances in the general DLM, 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 we'll define the scaled disturbances $\gamma_{0:T}$

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 scaled disturbances depending on how we order the elements of θ_t, as **?** note in a different class of models. We will sidestep the issue of the best ordering of the latent states. No matter which ordering is chosen, we can confirm our intuition that the scaled disturbances are an AA for V and W jointly. 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 $|J| = |L_W|^T = |W|^{T/2}$. Then from (5) we can write the full joint distribution of $(V, W, \gamma_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \gamma_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\gamma_0 - m_0)'C_0^{-1}(\gamma_0 - m_0)\right] \\ &\times |W|^{-(\lambda_W + p + T + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] |V|^{-(\eta_t + k + T + 2)/2} \\ &\times \exp\left[-\frac{1}{2}\left(tr\left(\Lambda_V V^{-1}\right) + \sum_{t=1}^{T}\left[y_t - F_t\theta_t(\gamma_{0:T}, W]'\, V^{-1}\left[y_t - F_t\theta_t(\gamma_{0:T}, W)\right]\right)\right)\right] \end{aligned} \tag{6}$$

where $\theta_t(\gamma_{0:T}, W)$ denotes the recursive back transformation defined by the scaled disturbances. So ultimately under the scaled disturbance parameterization we can write the model as

$$\begin{aligned} y_t|\gamma_{0:T}, V, W &\overset{ind}{\sim} N\left(F_t\theta_t(\gamma_{0:T}, W), V\right) \\ \gamma_t &\overset{iid}{\sim} N(0, I_p) \end{aligned} \tag{7}$$

for $t = 1, 2, \cdots, T$ where I_p is the $p \times p$ identity matrix. Neither V nor W are in the system equation so the scaled disturbances are an AA for (V, W). This parameterization is well known, e.g. **?** use it in a dynamic regression model with stationary regression coefficient.

The DA algorithm based on $\gamma_{0:T}$ draws $\gamma_{0:T}$ from its conditional posterior and then (V, W) from their joint conditional posterior given $\gamma_{0:T}$. There are a couple methods of performing this draw, including applying one of the simulation smoothers directly to drawing $\gamma_{0:T}$, if possible, or using one of them to draw the latent states $\theta_{0:T}$ before transformting the states to the sclaed disturbances. The draw from the joint conditional posterior of (V, W) is tricky because it is not a known density. We will illustrate how to accomplish it in a worked example in Section **??**.

2.2 The scaled errors

The scaled disturbances immediately suggest another potential AA that seems like it should be analogous — the scaled observation errors or more succinctly the scaled errors. What we are referring to is $v_t = y_t - F_t\theta_t$ appropriately scaled by V in the general DLM. Now let L_V denote the Cholesky decomposition of V, that is $L_V L_V' = V$. Then we can define a version of the scaled errors (this time depending on how we order the elements of y_t) as $\psi_0 = \theta_0$ and $\psi_t = L_V^{-1}(y_t - F_t\theta_t)$ for $t = 1, 2, \cdots, T$. This is a bit strange since in general $dim(\psi_0) \neq dim(\psi_t)$ for $t = 1, 2, \cdots, T$. Ideally we might like an "F_0" so that we can set $\psi_0 = F_0\theta_0$ in order for ψ_0 to have the same dimension as ψ_1. However, in general there is no F_0. In some DLMs F_t is constant with respect to t so that we could set $F_0 = F$, but in dynamic regression for example, there is no natural "F_0" assuming that we do not have the time-zero values of the covariates. To avoid this issue in practice, we simply leave $\psi_0 = \theta_0$ though transforming the initial value could in principle result in an algorithm with better properties.

There is a real difficulty, however. With this definition of $\psi_{0:T}$ it is not straightforward to to write down the model in terms of $\psi_{0:T}$ instead of $\theta_{0:T}$ and determine $p(\psi_{0:T}|V, W)$. When F_t is $k \times k$ (so that $dim(y_t) = k = p = dim(\theta_t)$) and is invertible for $t = 1, 2, \cdots, T$, $\psi_{0:T}$ is a one-to-one transformation of $\theta_{0:T}$ and the problem is easier. Then $\theta_t = F_t^{-1}(y_t - L_V\psi_t)$ for $t = 1, 2, \cdots, T$ while $\theta_0 = \psi_0$. The Jacobian of this transformation is block diagonal with a single copy of the identity matrix and the $F_t^{-1}L_V$'s along the diagonal, so $|J| = (\prod_{t=1}^{T}|F_t|^{-1})|V|^{T/2}$. Then from (5) we can write the joint distribution of $(V, W, \psi_{0:T}, y_{1:T})$

as

$$
\begin{aligned}
p(V, W, \psi_{0:T}, y_{1:T}) \propto & \exp\left[-\frac{1}{2}(\psi_0 - m_0)' C_0^{-1}(\psi_0 - m_0)\right] \\
& \times |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \psi_t' \psi_t\right] \\
& \times |W|^{-(\delta_t + k + T + 2)/2} \exp\left[-\frac{1}{2}\left(tr\left(\Lambda_W W^{-1}\right) + (y_t - \mu_t)'(F_t W F_t')^{-1}(y_t - \mu_t)\right)\right] \qquad (8)
\end{aligned}
$$

where we define $\mu_1 = L_V \psi_1 + F_1 G_1 \psi_0$ and for $t = 2, 3, \cdots, T$, $\mu_t = L_V \psi_t - F_t G_t F_{t-1}^{-1}(y_{t-1} - L_V \psi_{t-1})$. The $|F_t|^{-1}$'s have been absorbed into the normalizing constant, but if the F_t's depended on some unknown parameter then we could not do this and as a result would havffe to take them into account in a Gibbs step for F_t. Now we can write the model in terms of the scaled error parameterization:

$$
\begin{aligned}
y_t | V, W, \psi_{0:T}, y_{1:t-1} &\sim N(\mu_t, F_t' W F_t) \\
\psi_t &\overset{iid}{\sim} N(0, I_k)
\end{aligned}
$$

for $t = 1, 2, \cdots, T$ where I_k is the $k \times k$ identity matrix. Now we see immediately that the scaled errors, $\psi_{0:T}$, are also an AA for (V, W) since neither V nor W are in the system equation of this model. However, both V and W are in the observation equation so that $\psi_{0:T}$ is not a SA for (V, W) or for either one conditional on the other.

The DA algorithm based on $\psi_{0:T}$ is similar to that of $\gamma_{0:T}$ except we note that simiulation smoothing can be accomplished by directly applying the algorithm of **?** because the precision matrix of $\psi_{0:T}$ retains the necessary tridiagonal structure. Also we mention in passing that there is a bit of symmetry here — the joint conditional posterior of (V, W) given $\gamma_{0:T}$ is from the same family of densities as that of (W, V) given $\psi_{0:T}$ so that V and W essentially switch places. The upshot is that if we can draw from one we can draw from the other, so this part of our work has been essentially halved.

2.3 The elusive search for a sufficient augmentation

Having found two separate ancillary augmentations for the DLM, we would like to find a sufficient augmentation in order to implement take advantage of their likely weak posterior dependence and implement an ASIS. It turns out that this is no easy task. From equations (1) and (**??**) we can rewrite the by recursively substituting as

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

where $v_t \sim N(0, V)$ and $w_t \sim N(0, W)$ are independent. Here we see that θ_0 is given a special status relative to the other elements of the data augmentation which helps motivate not scaling it in the scaled disturbances or scaled errors. We are essentially treating it as a model parameter here and will continue to do so because it makes finding a sufficient augmentation easier (though still essentially impossible).

Now each y_t is a linear combination of normal distributions conditional on $\phi = (\theta_0, V, W)$, so $y_{1:T}$ has a marginal normal distribution such that

$$
\begin{aligned}
\mathrm{E}[y_t | \phi] &= F_t \prod_{s=t}^{1} G_s \theta_0 \\
\mathrm{Var}[y_t | \phi] &= V + F_t H_t W \\
\mathrm{Cov}[y_s, y_t | \phi] &= F_t H_t W
\end{aligned}
$$

where $\prod_{s=t}^{1} G_s = G_t G_{t-1} \cdots G_1$ and $H_t = I_p + G_t + G_t G_{t-1} + \cdots + G_t G_{t-1} \cdots G_2$. Next define

$$\mu = \begin{bmatrix} F_1 G_1 \theta_0 \\ F_2 G_2 G_1 \theta_0 \\ \vdots \\ F_T G_T G_{T-1} \cdots G_1 \theta_0 \end{bmatrix}, \quad \tilde{V}_{k\times k} = \begin{bmatrix} V & 0 & 0 & \ddots & 0 \\ 0 & V & 0 & \ddots & 0 \\ 0 & 0 & V & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \ddots & V \end{bmatrix}, \quad \tilde{W}_{k\times k} = \begin{bmatrix} F_1 H_1 \\ F_2 H_2 \\ \vdots \\ F_T H_T \end{bmatrix} W \begin{bmatrix} H_1' F_1' & H_2' F_2' & \dots & H_T' F_T' \end{bmatrix}.$$

Then we have the data model for $y_{1:T}$ without any data augmentation:

$$y_{1:T} \sim N_{p\times k}(\mu, \tilde{V} + \tilde{W}).$$

Now given a prior $p(\phi)$, this defines the posterior distribution of interest $p(\phi|y_{1:T})$.

Next we wish to find a sufficient augmentation θ (the lack of a subscript distinguishes this from the latent states $\theta_{0:T}$). Suppose we have such an augmentation and that conditional on ϕ, $(y_{1:T}, \theta)$ are normally distributed, in other words

$$\begin{bmatrix} \theta \\ y \end{bmatrix} \Bigg| \, \phi \sim N\left(\begin{bmatrix} \alpha_\theta \\ \mu \end{bmatrix}, \begin{bmatrix} \Omega_\theta & \Omega_{y,\theta}' \\ \Omega_{y,\theta} & \tilde{V} + \tilde{W} \end{bmatrix} \right)$$

which implies

$$\begin{aligned} y|\theta, \phi &\sim N(\mu + \Omega_{y,\theta}' \Omega_\theta^{-1}(\theta - \alpha_\theta), \tilde{V} + \tilde{W} - \Omega_{y,\theta}' \Omega_\theta^{-1} \Omega_{y,\theta}) \\ \theta|\phi &\sim N(\alpha_\theta, \Omega_\theta). \end{aligned}$$

Now for θ to be a sufficent augmentation we need $\mu + \Omega_{y,\theta}' \Omega_\theta^{-1}(\theta - \alpha_\theta)$ and $\tilde{V} + \tilde{W} - \Omega_{y,\theta}' \Omega_\theta^{-1} \Omega_{y,\theta}$ to be independent of ϕ. This requires that

$$\mu + \Omega_{y,\theta}' \Omega_\theta^{-1}(\theta - \alpha_\theta) = A\theta$$

where A a matrix which does not depend on ϕ. Rearranging, this gives $A = \Omega_{y,\theta}' \Omega_\theta^{-1}$ so that $\mu = A\alpha_\theta$ and $\Omega_{y,\theta} = \Omega_\theta A'$. Then using the second equation, we now require $\Sigma = \tilde{V} + \tilde{W} - A\Omega_\theta A'$ free of ϕ. This gives $A\Omega_\theta A' = \tilde{V} + \tilde{W} - \Sigma$. Consider $\tilde{\theta} = A\theta$, which is also a sufficient augmentation. Then we have

$$\tilde{\theta}|\phi \sim N(\mu, \tilde{V} + \tilde{W} - \Sigma)$$

and thus the posterior of ϕ given $\tilde{\theta}$ can be written as

$$p(\phi|\tilde{\theta}, y) \propto p(\phi)|\tilde{V} + \tilde{W} - \Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\tilde{\theta} - \mu)'(\tilde{V} + \tilde{W} - \Sigma)^{-1}(\tilde{\theta} - \mu) \right]$$

which for $\Sigma = 0$ is the posterior we wish to sample from. The transformation from $\tilde{\theta}$ to θ is unlikely to make this any easier.

The fundamental problem is that once we find a sufficient augmentation, in order to use it we must obtain draws from a density that appears just as hard to sample from as the posterior density we are already trying to approximate. We did treat θ_0 as a model parameter instead of an element of the data augmentation above, but changing this only makes the resulting conditional posterior of ϕ more complicated. The logic above does not rule out a useful sufficient augmentation, but it does suggest that it will be difficult to find one. Also, the problem we run into is unlikely to be unique to the time series setting but rather seems driven by trying to find a sufficient augmentation for a pair of variances, one on the data level and the other on the latent data level. For example, in a hierarchical model we expect there to be similar problems finding

a SA when both the observational and hierarchical variance are unknown. We run into a similar problem while trying to find two data augmentations that are independent in the posterior — which would guarantee an interweaving algorithm that yields iid draws of from the posterior distribution of the model parameters. After making sensible sounding assumptions about the nature of the DAs (i.e. joint with the data they are normally distributed along with some (in)dependence assumptions), the conditional posterior of ϕ ends of being identical to or just as complicated as the marginal posterior of ϕ.

2.4 The "wrongly scaled" DAs

The scaled disturbances are defined by $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ and the scaled errors are defined by $\psi_t = L_V^{-1}(y_t - \theta_t)$ for $t = 1, 2, \cdots, T$ where $L_W L_W' = W$ and $L_V L_V' = V$. Now 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 $\tilde{\psi}_0 = \tilde{\gamma}_0 = \theta_0$. In other words, the "tilde" versions of the scaled disturbances and the scaled errors are scaled by the "wrong" Cholesky decomposition, hence we call them the wrongly scaled disturbances and the wrongly scaled errors respectively. It is hard to motivate these DAs without looking forward to componentwise interweaving in the DLM (section **??**), but you can at least view them as the result of having thrown spaghetti against the walls to see what sticks. Once again both of these DAs have many variations depending on how the elements of θ_t or y_t are ordered, but we will ignore that issue.

First consider $\tilde{\gamma}_{0:T}$. Notice that for $t = 1, 2, \cdots, T$, $\tilde{\gamma}_t = L_V^{-1} L_W \gamma_t$ while $\tilde{\gamma_0} = \gamma_0$. The reverse transformation is then $\gamma_t = L_W^{-1} L_V \tilde{\gamma}_t$. The Jacobian is then block diagonal with $L_W^{-1} L_V$ along the diagonal. Thus $|J| = |L_W|^{-T}|L_V|^T = |W|^{-T/2}|V|^{T/2}$. Then from (6) we can write the joint distribution of $(V, W, \tilde{\gamma}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \tilde{\gamma}_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1}(\tilde{\gamma}_0 - m_0)\right] |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \\ &\times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)' V^{-1} \left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)\right] \\ &\times |W|^{-(\lambda_W + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \end{aligned} \tag{9}$$

Then under $\tilde{\gamma}_{0:T}$ we can write the model as

$$\begin{aligned} y_t | \tilde{\gamma}_{0:T}, V, W &\overset{ind}{\sim} N\left(F_t\theta_t(\tilde{\gamma}_{0:T}), V\right) \\ \tilde{\gamma}_t &\overset{ind}{\sim} N(0, L_V^{-1} W (L_V^{-1})') \end{aligned}$$

for $t = 1, 2, \cdots, T$. Since L_V is the Cholesky decomposition of V, the observation equation does not contain W. So $\tilde{\gamma}_{0:T}$ is a SA for $W|V$. Note also that since W and L_V are both in the system equation, $\tilde{\gamma}_{0:T}$ is not an AA for V nor for W.

Now consider $\tilde{\psi}_t = L_W^{-1} L_V \psi_t$ for $t = 1, 2, \cdots, T$ where again $\tilde{\psi}_0 = \psi_0 = \theta_0$ Then $\psi_t = L_V^{-1} L_W \tilde{\psi}_t$ and the Jacobian is block diagonal with $L_V^{-1} L_W$ along the diagonal. So $|J| = |V|^{-T/2}|W|^{T/2}$ and from (8) we can write the joint distribution of $(V, W, \tilde{\psi}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \tilde{\psi}_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)' C_0^{-1}(\tilde{\psi}_0 - m_0)\right] \\ &\times |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\psi}_t' (L_W^{-1} V (L_W^{-1})')^{-1} \tilde{\psi}_t\right] \\ &\times |W|^{-(\lambda_W + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_W W^{-1}\right)\right] |V|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} (y_t - \tilde{\mu}_t)' (F_t W F_t')^{-1} (y_t - \tilde{\mu}_t)\right] \end{aligned} \tag{10}$$

where we define $\tilde{\mu}_1 = L_W\tilde{\psi}_1 - F_1G_1\tilde{\psi}_0$ and for $t = 2, 3, \cdots, T$ $\tilde{\mu}_t = L_W\tilde{\psi}_t - F_tG_tF_{t-1}^{-1}(y_{t-1} - L_W\tilde{\psi}_{t-1})$ In terms of $\tilde{\psi}_{0:T}$, the model is then:

$$y_t|V, W, \tilde{\psi}_{0:T}, y_{1:t-1} \sim N(\tilde{\mu}_t, F_t'WF_t)$$
$$\tilde{\psi}_t \overset{iid}{\sim} N(0, L_W^{-1}V(L_W^{-1})')$$

for $t = 1, 2, \cdots, T$. Since $\tilde{\mu}_t$ only depends on W (through L_W) and not on V, V is absent from the observation equation. Thus $\tilde{\psi}_{0:T}$ is a SA for $V|W$. Again that both W and V are in the system equation so $\tilde{\psi}_{0:T}$ is not an AA for either V or W.

In the case of both wrongly scaled DA algorithms, the smoothing step can be accomplished in a manner analogous to the "correctly scaled" DA algorithms, i.e. the scaled disturbance and scaled error algorithms. The draw from the joint conditional posterior of (V, W) is from a nonstandard density that, like for the correctly scaled DA algorithms, has a certain symmetry property. Specifically $V, W|\tilde{\gamma}_{0:T}, y_{1:T}$ and $W, V|\tilde{psi}_{0:T}, y_{1:T}$ have densities from the same family so that by changing which of $\tilde{\gamma}_{0:T}$ or $\tilde{\psi}_{0:T}$ is conditioned on, V and W essentially switch places. This class of densities is different from the correctly scaled DA case, however. We will demonstrate this through an example in Section .

3 Interweaving in the DLM: Global and Componentwise

We now have five DAs for the generic DLM with known F_t's and G_t's. For simplicity we'll assume that $dim(y_t) = dim(\theta_t)$ and F_t invertible for $t = 1, 2, \cdots, T$ so that the scaled errors are easy to work with. The five DAs are the states, $\theta_{0:T}$, the scaled disturbances $\gamma_{0:T}$, the scaled errors $\psi_{0:T}$, the wrongly scaled disturbances $\tilde{\gamma}_{0:T}$, and the wrongly scaled errors $\tilde{\psi}_{0:T}$. This allows us to construct several GIS algorithms based on algorithm 3. The main algorithms we consider are the state-dist, state-error, dist-error, and triple interweaving algorithms. The names should be intuitive, but for example the state-dist algorithm interweaves between the states and the scaled disturbances, while the triple interweaving algorithm interweaves between the states, the scaled diturbances, and the scaled errors. Strictly speaking the order in which we sample the DAs in the algorithm does matter, but **?** note that this tends not to make much difference. We always construct our algorithms so that the DAs are used in the order they were presented earlier in this paragraph.

To illustrate the GIS algorithms, **??** is the state-dist GIS algorithm:

Algorithm 5.

$$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \to [V, W|\theta_{0:T}, y_{1:T}] \to [\gamma_{0:T}|V, W, \theta_{0:T}, y_{1:T}] \to [V^{(k+1)}, W^{(k+1)}|\gamma_{0:T}, y_{1:T}]$$

the third step is actually a one-to-one transformation from $\theta_{0:T}$ to $\gamma_{0:T}$. In practice we may want to break up step 4 into two steps if it is easier to draw from the full conditionals of V and W rather than drawing them jointly, though this will cost us both in terms of MCMC efficiency and theoretical tractibility of analyzing the algorithm.

None of the GIS algorithms we can construct are ASIS algorithms — none of the DAs are a SA for (V, W). The states, $\theta_{0:T}$, are a SA for $W|V$ though, so this motivates a CIS algorithm. A partial CIS algorithm is immediate:

Algorithm 6.

$$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \to [V^{(k+1)}|W^{(k)}, \theta_{0:T}$$
$$[W|V^{(k+1)}, \theta_{0:T}, y_{1:T}] \to [\gamma_{0:T}^{(k+1)}|V^{(k+1)}, W, \theta_{0:T}, y_{1:T}] \to [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$$

Steps 1-2 of this algorithm correspond to a Gibbs step for V while steps 3-5 correspond to a Gibs step for W. Step 4 is a transformation step since conditional on V and W, $\gamma_{0:T}$ is a one-to-one transformation of $\theta_{0:T}$. This algorithm is actually the same as a version of the state-dist interweaving algorithm with some of the steps rearranged, specifically algorithm **??**. So it should be similar in performance to a GIS algorithm.

With a little more work, we can also construct a full CIS algorithm that also turns out to be essentially the same as another GIS algorithm. Here we employ the wrongly scaled disturbances $\tilde{\gamma}_{0:T}$ and wrongly scaled errors $\tilde{\psi}_{0:T}$. Now we already now that $\gamma_{0:T}$ is an AA for $W|V$ and $\tilde{\gamma}_{0:T}$ is a SA for $W|V$, so the two form an AA-SA pair for $W|V$. Similarly, $\psi_{0:T}$ is an AA for $V|W$ while $\tilde{psi}_{0:T}$ is a SA for $V|W$ so together they form an AA-SA pair for $V|W$. Now we can construct a full CIS algorithm:

Algorithm 7.

$[\tilde{\psi}_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \to [V|W^{(k)}, \tilde{\psi}_{0:T}, y_{1:T}] \to [\psi_{0:T}|V, W^{(k)}, \tilde{\psi}_{0:T}, y_{1:T}] \to [V^{(k+1)}|W^{(k)}, \psi_{0:T}, y_{1:T}] \to$
$[\tilde{\gamma}_{0:T}|V^{(k+1)}, W^{(k)}, \psi_{0:T}, y_{1:T}] \to [W|V^{(k+1)}, \tilde{\gamma}_{0:T}, y_{1:T}] \to [\gamma_{0:T}|V^{(k+1)}, W, \tilde{\gamma}_{0:T}, y_{1:T}] \to [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$

Steps 1-4 constitute a Gibbs step for V and steps 5-8 constitute a Gibbs step for W. Steps 3, 5 and 7 are transformation steps — the parameter we are drawing is a one to one function of the parameters we are conditioning on. It turns out that $p(W|V, \tilde{\gamma}_{0:T}, y_{1:T})$ and $p(W|V, \theta_{0:T}, y_{1:T})$ are the same density, and also that $p(V|W, \tilde{\psi}_{0:T}, y_{1:T})$ and $p(V|W, \theta_{0:T}, y_{1:T})$ are the same density. The upshot is that step 1 of algorithm **??** can be replaced with a draw from $p(\theta_{0:T}|V, W, y_{1:T})$, and any time we condition on one of the "wrongly scaled" variables, we can condition on $\theta_{0:T}$ instead, yielding the following version of the same CIS algorithm:

Algorithm 8.

$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \to [V|W^{(k)}, \theta_{0:T}, y_{1:T}] \to [\psi_{0:T}|V, W^{(k)}, \theta_{0:T}, y_{1:T}] \to [V^{(k+1)}|W^{(k)}, \psi_{0:T}, y_{1:T}] \to$
$[\theta_{0:T}|V^{(k+1)}, W^{(k)}, \psi_{0:T}, y_{1:T}] \to [W|V^{(k+1)}, \theta_{0:T}, y_{1:T}] \to [\gamma_{0:T}|V^{(k+1)}, W, \theta_{0:T}, y_{1:T}] \to [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$

It can also be shown that this algorithm and the dist-error algorithm employ the same steps, just in a different order. This suggests that we should expect the two algorithms to perform similarly.

4 Application: The Local Level Model

In order to illustrate how these algorithms work, we will focus on the local level model for simplicity though there are still some difficulties. The local level model (LLM) is a DLM with univariate data y_t for $t = 1, 2, \cdots, T$ and a univariate latent state θ_t for $t = 0, 2, \cdots, T$ that satisfies

$$y_t|\theta_{0:T} \overset{ind}{\sim} N(\theta_t, V) \tag{11}$$
$$\theta_t|\theta_{0:t-1} \sim N(\theta_{t-1}, W) \tag{12}$$

with $\theta_0 \sim N(m_0, C_0)$. Here $\theta_t = E[y_t|\theta_{0:T}]$. The states are $\theta_{0:T}$, the scaled disturbances are $\gamma_{0:T}$ with $\gamma_0 = \theta_0$ and $\gamma_t = (\theta_t - \theta_{t-1})/\sqrt{W}$ for $t = 1, 2, \cdots, T$, and the scaled errors are $\psi_{0:T}$ with $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \cdots, T$. The independent inverse Wishart priors on V and W in Section 1 cash out to independent inverse gamma priors for the local level model, viz $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$.

4.1 Base Samplers

The joint density of $(V, W, \theta_{0:T}, y_{1:T})$ is:

$$p(V, W, \theta_{0:T}, y_{1:T}) \propto V^{-(\alpha_V+1+T/2)} \exp\left[-\frac{1}{V}\left(\beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \theta_t)^2\right)\right]$$
$$W^{-(\alpha_W+1+T/2)} \exp\left[-\frac{1}{W}\left(\beta_W + \frac{1}{2}\sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2\right)\right] \exp\left[-\frac{1}{2C_0}(\theta_0 - m_0)^2\right]$$

This immediately gives the state sampler:

Algorithm 9 (State Sampler for LLM)**.**

$$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V^{(k+1)}, W^{(k+1)}|\theta_{0:T}, y_{1:T}]$$

In step 2, V and W are independent with $V \sim IG(a_V, b_V)$ and $W \sim IG(a_W, b_W)$ where $a_V = \alpha_V + T/2$, $b_V = \beta_V + \sum_{t=1}^{T}(y_t - \theta_t)^2/2$, $a_W = \alpha_W + T/2$, and $b_W = \beta_W + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2/2$.

The scaled disturbance sampler, i.e. the DA algorithm based on the scaled disturbances, is a bit more complicated. In this context $\gamma_0 = \theta_0$ and $\gamma_t = (\theta_t - \theta_{t-1})/\sqrt{W}$ for $t = 1, 2, \cdots, T$, and thus $\theta_t = \sqrt{W}\sum_{t=1}^{T}\gamma_t + \gamma_0$ for $t = 1, 2, \cdots, T$. Following (6), we can write the joint posterior of $(V, W, \gamma_{0:T})$ as

$$\begin{aligned} p(V, W, \gamma_{0:T}|y_{1:T}) \propto V^{-(\alpha_V+1+T/2)} \exp\left[-\frac{1}{V}\left(\beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \gamma_0 - \sqrt{W}\sum_{s=1}^{t}\gamma_s)^2\right)\right] \\ \times W^{-(\alpha_W+1)} \exp\left[-\frac{\beta_W}{W}\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\right] \exp\left[-\frac{1}{2C_0}(\gamma_0 - m_0)^2\right] \end{aligned} \tag{13}$$

Now V and W are no longer conditionally independent given $\gamma_{0:T}$ and $y_{1:T}$. Instead of attempting the usual DA algorithm, we will add an extra Gibbs step and draw V and W separately primarily for ease of computation. This gives us the scaled disturbance sampler:

Algorithm 10 (Scaled Disturbance Sampler for LLM)**.**

$$[\gamma_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V^{(k+1)}|W^{(k)}, \gamma_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$$

In step 2, V is drawn from the same inverse gamma distribution as in step 2 of algorithm **??**. In step 3, the draw of W is more complicated. The density can be written as

$$p(W|V, \gamma_{0:T}, y_{1:T}) \propto W^{-\alpha_W - 1} \exp\left[-\frac{1}{2V}\sum_{t=1}^{T}\left(y_t - \gamma_0 - \sqrt{W}\sum_{s=1}^{t}\gamma_s\right)^2\right] \exp\left[-\frac{\beta_W}{W}\right].$$

This density is not any known form and is difficult to sample from, though its functional form is similar to the generalized inverse gaussian distribution. The log density can be written as

$$\log p(W|V, \gamma_{0:T}, y_{1:T}) = -aW + b\sqrt{W} - (\alpha_W + 1)\log W - \beta_W/W + C$$

where C is some constant, $a = \sum_{t=1}^{T}(\sum_{j=1}^{t}\gamma_j)^2/2V$ and $b = \sum_{t=1}^{T}(y_t - \gamma_0)(\sum_{j=1}^{t}\gamma_j)/V$. It can be shown that $b > \left(\frac{(\alpha+1)^3}{\beta}\right)^{1/2}\frac{4\sqrt{2}}{3\sqrt{3}}$ implies that the density is log concave. It turns out that this tends to hold over a wide region of the parameter space — so long as V is smaller or is not much larger than W. This allows for the use of adaptive rejection sampling in order to sample from this distribution in many cases, e.g. using **?**. An alternative is to use a t approximation to the conditional density as a proposal in a rejection sampler. This is much more computationally expensive when necessary, but it works ok on the log scale.

The scaled error sampler is similar to the scaled disturbance sampler and this is easy to see in the local level model. Here $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \cdots, T$ so that $\theta_t = y_t - \sqrt{V}\psi_t$ for $t = 1, 2, \cdots, T$. From (8) we can write $p(V, W, \psi_{0:T}|y_{1:T})$ as

$$\begin{aligned} p(V, W, \psi_{0:T}, y_{1:T}) \propto W^{-(\alpha_W+1+T/2)} \exp\left[-\frac{1}{W}\left(\beta_W + \frac{1}{2}\sum_{t=1}^{T}(Ly_t - \sqrt{V}L\psi_t)^2\right)\right] \\ V^{-(\alpha_V+1)} \exp\left[-\frac{\beta_V}{V}\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\psi_t^2\right] \exp\left[-\frac{1}{2C_0}(\psi_0 - m_0)^2\right] \end{aligned}$$

where we define $Ly_t = y_t - y_{t-1}$ for $t = 2, 3, \cdots, T$ & $Ly_1 = y_1 - \psi_0$ and $L\psi_t = \psi_t - \psi_{t-1}$ for $t = 2, 3, ..., T$ & $L\psi_1 = \psi_1 - 0$. Once again, V and W are no longer conditionally independent given $\psi_{0:T}$ and $y_{1:T}$. In fact, the density is analgous to (**??**) with V and W switching places. The scaled error sampler obtained from drawing V and W separately is:

Algorithm 11 (Scaled Error Sampler for LLM)**.**

$$[\psi_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \to [V^{(k+1)}|W^{(k)}, \psi_{0:T}, y_{1:T}] \to [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$$

In step 3, W is drawn from the same inverse gamma distribution as in step 2 of algorithm **??**. Drawing V in step 2 is more complicated, but exactly analogous to drawing W in algorithm **??**. The log density of $V|W, \psi_{0:T}, y_{1:T}$ can be written as

$$\log p(V|W, \psi_{0:T}, y_{1:T}) = -aV + b\sqrt{V} - (\alpha_V + 1)\log V - \beta_V/V + C$$

where again C is some constant, but now $a = \sum_{t=1}^{T}(L\psi_t)^2/2W$ and $b = \sum_{t=1}^{T}(L\psi_t Ly_t)/W$ but otherwise the form of the density is the same as that of $W|V, \gamma_{0:T}, y_{1:T}$.

We can also contruct the DA algorithms based on the "wrongly scaled" disturbances or errors. The wrongly scaled disturbances are defined by $\tilde{\gamma}_t = \gamma_t \frac{\sqrt{W}}{\sqrt{V}}$ for $t = 1, 2, \cdots, T$ and $\tilde{\gamma}_0 = \gamma_0$ while the wrongly scaled errors are defined by $\tilde{\psi}_t = \psi_t \frac{\sqrt{V}}{\sqrt{W}}$ for $t = 1, 2, \cdots, T$ and $\tilde{\psi}_0 = \psi_0$. For $\tilde{\gamma}_{0:T}$ we have

$$\begin{aligned} p(V, W|\tilde{\gamma}_{0:T}, y_{1:T}) \propto & W^{-\alpha_W - T/2 - 1} \exp\left[-\frac{1}{2W/V}\sum_{t=1}^{T}\tilde{\gamma}_t^2\right] \exp\left[-\frac{\beta_W}{W}\right] \\ & \times V^{-\alpha_V - 1} \exp\left[-\frac{\beta_V}{V}\right] \exp\left[-\frac{1}{2V}\sum_{t=1}^{T}\left(y_t - \tilde{\gamma}_0 - \sqrt{V}\sum_{s=1}^{t}\tilde{\gamma}_s\right)^2\right]. \end{aligned}$$

Thus the conditional posterior of W given V and $\tilde{\gamma}_{0:T}$ is the same as if we had conditioned on $\theta_{0:T}$ instead of $\tilde{\gamma}_{0:T}$. In other words

$$p(W|V, \tilde{\gamma}_{0:T}, y_{1:T}) \propto W^{-(\alpha_W + T/2) - 1} \exp\left[-\frac{1}{W}\left(\beta_W + \frac{1}{2}V\sum_{t=1}^{T}\tilde{\gamma}_t^2\right)\right]$$

so that $V|W, \tilde{\gamma}_{0:T}, y_{1:T} \sim IG(a_W, b_W)$ where $a_W = \alpha_W + T/2$ and

$$b_W = \beta_W + \frac{1}{2}V\sum_{t=1}^{T}\tilde{\gamma}_t^2 = \beta_W + \frac{1}{2}\sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2.$$

The conditional posterior of V is more complicated. We have

$$\begin{aligned} p(V|W, \tilde{\gamma}_{0:T}, y_{1:T}) & \propto \exp\left[-\frac{1}{2W/V}\sum_{t=1}^{T}\tilde{\gamma}_t^2\right] V^{-\alpha_V - 1} \exp\left[-\frac{\beta_V}{V}\right] \exp\left[-\frac{1}{2V}\sum_{t=1}^{T}\left(y_t - \tilde{\gamma}_0 - \sqrt{V}\sum_{s=1}^{t}\tilde{\gamma}_s\right)^2\right] \\ & \propto V^{-\alpha_V - 1} \exp\left[-\frac{a}{V} + \frac{b}{\sqrt{V}} - cV\right] \end{aligned}$$

where

$$\begin{aligned} a &= \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)^2 > 0 \\ b &= \sum_{t=1}^{T}(y_t - \tilde{\gamma}_0)\sum_{s=1}^{t}\tilde{\gamma}_s \\ c &= \frac{1}{2W}\sum_{t=1}^{T}\tilde{\gamma}_t^2 > 0. \end{aligned}$$

We will return to this density momentarily.

For the wrongly scaled errors, we have

$$p(V, W|\tilde{\psi}_{0:T}, y_{1:T}) \propto V^{-\alpha_V - T/2 - 1} \exp\left[-\frac{1}{2V/W}\sum_{t=1}^{T}\tilde{\psi}_t^2\right] \exp\left[-\frac{\beta_V}{V}\right]$$
$$\times W^{-\alpha_W - 1} \exp\left[-\frac{1}{2W}\sum_{t=1}^{T}\left(\tilde{Ly}_t - \sqrt{W}(\tilde{L\psi}_t)\right)\right]$$

where we define $\tilde{Ly}_t = y_t - y_{t-1}$ for $t = 1, 2, \cdots, T$ and $\tilde{Ly}_1 = y_1 - \tilde{\psi}_0$, and $\tilde{L\psi}_t = \tilde{\psi}_t - \tilde{\psi}_{t-1}$ for $t = 1, 2, \cdots, T$ with $\tilde{L\psi}_1 = \tilde{\psi}_1$. Then the conditional posterior of V is the same as if we had conditioned on $\theta_{0:T}$ instead of $\tilde{\psi}_{0:T}$, i.e.

$$p(V|W, \tilde{\psi}_{0:T}, y_{1:T}) \propto V^{-(\alpha_V - T/2) - 1} \exp\left[-\frac{1}{V}\left(\beta_V + \frac{1}{2}W\sum_{t=1}^{T}\tilde{\psi}_t^2\right)\right]$$

so that $V|W, \tilde{\psi}_{0:T}, y_{1:T} \sim IG(a_V, b_V)$ where $a_V = \alpha_V + T/2$ and

$$b_V = \beta_V + \frac{1}{2}W\sum_{t=1}^{T}\tilde{\psi}_t^2 = \beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \theta_t)^2.$$

The conditional posterior of W is more complicated but similar to that of V when we conditioned on $\tilde{\gamma}_{0:T}$. We have

$$p(W|V, \tilde{\psi}_{0:T}, y_{1:T}) \propto \exp\left[-\frac{1}{2V/W}\sum_{t=1}^{T}\tilde{\psi}_t^2\right] W^{-\alpha_W - 1} \exp\left[-\frac{1}{2W}\sum_{t=1}^{T}\left(\tilde{Ly} - \sqrt{W}\tilde{L\psi}\right)\right]$$
$$\propto W^{-\alpha_W - 1} \exp\left[-\frac{a}{W} + \frac{b}{\sqrt{W}} - cW\right]$$

where now

$$a = \beta_W + \frac{1}{2}\sum_{t=1}^{T}\tilde{Ly}_t^2 > 0$$
$$b = \sum_{t=1}^{T}\tilde{Ly}_t\tilde{L\psi}_t$$
$$c = \frac{1}{2V}\sum_{t=1}^{T}\tilde{\psi}_t^2 > 0.$$

So in the case of both wrongly scaled DAs we need to sample from a density of the form

$$p(X) \propto X^{-\alpha - 1} \exp\left[-\frac{a}{X} + \frac{b}{\sqrt{X}} - cX\right].$$

The density of $Y = \log(X)$ is

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

This density is easy to sample from fairly efficiently with rejection sampler using a t or normal approximation as a proposal. It is also typically log concave, so adaptive rejection sampling will work as well. In particular when $b \leq 0$ or $a > \frac{3b}{16}\left(\frac{b}{16c}\right)^{1/3}$ the density of Y is log concave.

4.2 Hybrid Samplers: Interweaving, Alternating and Random Kernel

Section **??** contains the details for the interweaving algorithms in the general DLM. In the local level model the only caveat is that we only sample V and W jointly when we condition on the states. We will consider all four GIS samplers based on any two or three of the base samplers and one CIS sampler. In the GIS samplers, the order of the parameterizations will always be the states ($\theta_{0:T}$), then the scaled disturbances ($\gamma_{0:T}$), then the scaled errors ($\psi_{0:T}$). All of the GIS algorithms and the full CIS algorithm are below in Table **??**. Note the distributional forms for each of these steps (in some cases a transformation) are in Section **??**. We ignore the partial CIS algorithm

1. state-dist GIS algorithm:
   $[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\gamma_{0:T}|V, W, \theta_{0:T}, y_{1:T}] \rightarrow [V^{(k+1)}|W, \gamma_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$

2. state-error GIS algorithm:
   $[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, \theta_{0:T}, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

3. dist-error GIS algorithm:
   $[\gamma_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V|W^{(k)}, \gamma_{0:T}, y_{1:T}] \rightarrow [W|V, \gamma_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, \gamma_{0:T}, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

4. triple GIS algorithm:
   $[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\gamma_{0:T}|V, W, \theta_{0:T}, y_{1:T}] \rightarrow [V|W, \gamma_{0:T}, y_{1:T}] \rightarrow [W|V, \gamma_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, \gamma_{0:T}, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

5. full CIS algorithm:
   $[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V|W^{(k)}, \theta_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, \theta_{0:T}, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [\theta_{0:T}|V^{(k+1)}, W, y_{1:T}] \rightarrow [W|V^{(k+1)}, \theta_{0:T}, y_{1:T}] \rightarrow [\gamma_{0:T}|V^{(k+1)}, W, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$

Table 1: GIS and CIS algorithms for the local level model

Interweaving algorithms are conceptually very similar to alternating algorithms. For every GIS algorithm, there's a corresponding alternating algorithm where each $[DA_2|V, W, DA_1]$ step is replaced by a $[DA_2|V, W]$ step (here DA_i is a data augmentation for $i = 1, 2.$). Table **??** contains each alternating algorithm. Note that there are two possible "hybrid triple" algorithms that we don't consider here where the move from $\theta_{0:T}$ to $\gamma_{0:T}$ interweaves and while the move from $\gamma_{0:T}$ to $\psi_{0:T}$ alternates and vice versa.

Table **??** contains each algorithm we considered for the local level model. The basic idea here is that the alternating algorithms should serve as a sort of baseline to compare the corresponding interweaving algorithms against. The GIS algorithm should be slightly faster than the alternating algorithm since the only difference is one step becoming a transformation instead of a random draw, but the difference should not be large since there is no reason to expect the scaled disturbances and the scaled errors to have low to zero dependence in the posterior. So we would like the GIS algorithms to have at least as quick mixing as the corresponding alternating algorithms. We can make this notion precise by considering the effective sample size (ESS) of the Markov chain – we would like the GIS algorithms to have an ESS that is larger than their corresponding alternating algorithms for the same actual sample size. We omit the partial CIS algorithm from our results because, as expected, it does not perform materially different from the state-dist algorithm.

4.3 Simulation Setup

In order to test these algorithms, we simulated a fake dataset from the local level model for various choices of V, W, and T. We created a grid over V–W space with (V, W) ranging from $(10^{-2}, 10^{-2})$ to $(10^2, 10^2)$ and we

1. State-Dist alternating algorithm:
$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\gamma_{0:T}|V, W, y_{1:T}] \rightarrow [V^{(k+1)}|W, \gamma_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}, y_{1:T}]$

2. State-Error alternating GIS algorithm:
$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

3. Dist-Error alternating GIS algorithm:
$[\gamma_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V|W, \gamma_{0:T}, y_{1:T}] \rightarrow [W|V, \gamma_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

4. Triple alternating GIS algorithm:
$[\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T}] \rightarrow [V, W|\theta_{0:T}, y_{1:T}] \rightarrow [\gamma_{0:T}|V, W, y_{1:T}] \rightarrow [V|W, \gamma_{0:T}, y_{1:T}] \rightarrow [W|V, \gamma_{0:T}, y_{1:T}] \rightarrow [\psi_{0:T}|V, W, y_{1:T}] \rightarrow [V^{(k+1)}|W, \psi_{0:T}, y_{1:T}] \rightarrow [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}, y_{1:T}]$

Table 2: Alternating algorithms for the local level model

Base	State	(wrongly) Scaled Disturbance	(wrongly) Scaled Error	
GIS	State-Dist	State-Error	Dist-Error	Triple
Alt	State-Dist	State-Error	Dist-Error	Triple
CIS	State-Error/WError-Error for $V\|W$; State-Dist/WDist-Dist for $W\|V$			

Table 3: Each algorithm considered for the local level model

simulated a dataset for all possible combinations of V and W with each of $T = 10, 100, 1000$. Then for each dataset, we fit the local level model using each algorithm in Table **??**. We used the same rule for constructing priors for each model: $\theta_0 \sim N(0, 10^7)$, $V \sim IG(5, 4\tilde{V})$, and $W \sim IG(5, 4\tilde{W})$, mutually independent where $(\tilde{V}, \tilde{W})$ are the true values of V and W used to simulate the time series. Thus both the prior and likelihood roughly agree about the likely values of V and W.

For each dataset and each sampler, we obtained $n = 3000$ draws and threw away the first 500 as burn in. The chains were started at the true values used to simulated the time series, so we can examine the behavior of the chains to determine how well they mix but not how quickly they converge. Define the effective sample proportion (ESP) for a scalar component of the chain as the effective number of independent draws, or effective sample size (ESS) (see e.g. **?**) of the component divided by the actual sample size, i.e. $ESP = ESS/n$. An $ESP = 1$ indicates that the Markov chain is behaving as if it obtains iid draws from the posterior. It is possible to obtain $ESP > 1$ if the draws are negatively correlated and occasionally for some of our samplers our estimates of ESS are negative, but we round this up to 0 so that the maximum ESP possible is 1 in our plots.

4.4 Base Results

Figure **??** contains plots of ESP for V and W in each chain of each base sampler for each of $T = 10$, $T = 100$, and $T = 1000$. We'll focus on $T = 10$ first. The state sampler has a low ESP for V and a high ESP for W when the signal-to-noise ratio, W/V, is larger than one. When the signal-to-noise ratio is smaller than one, on the other hand, the state sampler has a low ESP for W and a high ESP for V. In the usual case where the signal to noise ratio isn't too different from one, the state sampler has a modest to low ESP for both V and W. Note that the particular values of V and W don't seem to matter at all — just their relative values, i.e. the signal-to-noise ratio W/V. Moving up any diagonal on the plots for V and W in the state sampler, W/V is constant and the ESS appears roughly constant. The basic lesson here is that the state sampler has mixing issues for whichever of V or W is smaller.

Figure **??** tells a different story for the scaled disturbance sampler. When the signal-to-noise ratio is less

than one, ESPs for both V and W are nearly 1, i.e. the effective sample size is nearly the actual sample size of the chain. When the signal-to-noise ratio is greater than one, however, ESP for both V and W becomes small, especially for V. Once again the absolute values of V and W don't matter for this behavior — just the relative values. The scaled error sampler has essentially the opposite properties. When W/V is large, it has a near 1 ESP for both V and W. On the other hand, when W/V is small is has a low ESP for both V and W, especially for V. The lesson here seems to be that the scaled disturbances ($\gamma_{0:T}$) are the preferred data augmentation for low signal-to-noise ratios and the scaled errors ($\psi_{0:T}$) are the preferred data augmentation for high signal-to-noise ratios, while the states ($\theta_{0:T}$) are preferred for signal-to-noise ratios near 1. The wrongly scaled disturbances ($\tilde{\gamma}_{0:T}$) and wrongly scaled errors ($\tilde{\psi}_{0:T}$), on the other hand, look like worse versions of the state sampler. The pattern of mixing for V and W over the range of the parameter space is essentially the same as the state sampler, except the wrongly scaled disturbance sampler has worse mixing for V than the state sampler everywhere and similarly the wrongly scaled error sampler has worse mixing for W than the state sampler everywhere.

The plots for $T = 100$ and $T = 1000$ in Figure **??** tell basically the same story, with a twist. Increasing the length of the time series seems to exacerbate all problems without changing the basic conclusions. As T increases, W/V has to be smaller and smaller for the scaled disturbance sampler to have decent mixing, and similarly W/V has to be larger and larger for the scaled error sampler to have decent mixing. Interestingly, the scaled error sampler appears to mix well for both V and W over a larger region of the space $W/V < 1$ than the scaled disturbance sampler does over $W/V > 1$. The state sampler is stuck between a rock and a hard place, so to speak, since as T increases, good mixing for V requires W/V to be smaller and smaller, but good mixing for W requires W/V to be larger and larger. The wrongly scaled samplers are again pretty similar to the state sampler for larger T except the wrongly scaled sampler tends to be worse everywhere for the variance that was used to scale — i.e. once again the wrongly scaled disturbance sampler has worse mixing for V than the state sampler while the wrongly scaled error smapler has worse mixing for W than the state sampler. However, the wrongly scaled samplers do appear to have slightly better mixing than the state sampler for the variance that was *not* used to scale. In particular, the wrongly scaled error sampler appears to have slightly better mixing for V than the state sampler over part of the parameter space when $T = 100$ or $T = 1000$.

It's also worth noting that both the scaled error and scaled disturbance samplers run into trouble with their adaptive rejection sampling step in precisely the same region of the parameter space where they have good mixing for both V and W, though as T increases, this only happens in the increasingly extreme ends of the parameter space. More precisely, when $W/V > 1$, $p(W|V, \psi_{0:T}, y_{1:T})$ will often fail to be log concave, and when $W/V < 1$, $p(V|W, \gamma_{0:T}, y_{1:T})$ will often fail to be log concave, but as T increases the degree to which W/V must differ from one (in the appropriate direction) in order for log concavity to often or even occasionally fail increases. Outside of these respective regions, log-concavity of the relevant density failing is an extremely unlikely occurence. As a result, the adaptive rejection sampling algorithm of **?** won't work in general. Another option is to give up directly sampling from either conditonal density and use a metropolis step, perhaps for (V, W) jointly. In general, the sampling algorithm should be prepared to use something other than adaptive rejection sampling if necessary because it's possible that the chain enters a region of the parameter space where the relevant density is not log concave, no matter what the likely values of V and W are. *NOTE: ADD DETAILS ABOUT PRECISELY HOW LARGE OR SMALL W/V HAS TO BE TO THIS PARAGRAPH*

Based on the intuition in Section 2 above, the GIS algorithms should work best when at least one of the underlying base algorithms has a high ESP — the basic idea is that when least one of the underlying algorithms has low autocorrelation, we should have low autocorrleation in the GIS algrothim using multiple DAs. This suggests that the dist-error GIS algorithm will have the best performance of the GIS algorithms using two DAs for both V and W, especially for W/V far away from one. When W/V is near one it may offer no improvement, especially for large T. The state-dist GIS algorithm should have trouble with V when W/V is high since both the state sampler and the scaled disturbance sampler have trouble with V when W/V is high. Similarly, the state-error GIS algorithm should have trouble with W when W/V is low since both underlying samplers have trouble with W when W/V is low. Since the triple GIS algorithm adds the

state sampler into the dist-error GIS algorithm, it seems plausible that it might improve mixing for one of V or W since for V/W different from one, the state sampler has good mixing for at least one of V of W. The full CIS algorithm, on the other hand, is unlikely to be better than the dist-error GIS algorithm since in a certain sense one algorithm is the same as the other, just with the steps reordered.

We can verify most of these intuitions in Figure **??**. First, the state-dist GIS algorithm has high ESP for W except for a narrow band where W/V is near one, though this band becomes much wider as T increases. The state-dist GIS algorithm's mixing behavor for V appears identical to the original state sampler — high ESP when $W/V < 1$ and poor ESP when $W/V > 1$, and again the good region shrinks as T increases. So this algorithm behaves as expected — it takes advantage of the fact that the state and scaled disturbance DA algorithms make up a "beauty and the beast" pair for W and thus improves mixing for W. However, the two underlying DA algorithms behave essentially indentically for V so there is no improvement. Similarly the state-error GIS algorithm's ESP for W is essentially identical to the state and scaled error algorithms' ESP for W — high when W/V large and low when W/V small — but for V, the state-error algorithm has a high ESP when W/V isn't too close to one, especially when T is small. The dist-error GIS algorithm also behaves as predicted — when W/V is not too close to one it has high ESP for both V and W, though as T increases W/V has too be farther away from one in order for the ESPs to be high. The dist-error GIS algorithm behaves apparently identically to the full CIS and triple GIS algorithms, with some differences when T is small. The first of these is not surprising — based on the intuition that the dist-error GIS and full CIS algorithms are the same up to a reordering of each of their steps, we didn't expect much of a difference. However, we had some hope that the triple GIS algorithm would improve upon the dist-error GIS algorithm somewhat by further breaking the correlation between iterations in the Markov chain. This didn't happen, and furthermore the state-dist and state-error samplers didn't improve the ESP for V or W respectively. When the two underlying DA algorithms form a "beast and the beast" pair, the interweaving algorithm appears to mix just as well as the best mixing single DA algorithm.

Finally Figure **??** allows us to compare the GIS algorithms to the alternating and random kernel algorithms. Note that for the purposes of making a direct comparison, these plots show ESP/2 for the three two-DA random kernel algorithms and ESP/3 for the triple random kernel algorithm. We do this because the alternating and interweaving algorithms each have to two roughly twice as much computation as the random kernel algorithm in order to complete one full iteration of the sampler, or in the case of the triple algorithms three times as much. The main takeaway is that there doesn't appear to be any difference between itnerweaving and alternating, and the differences between the random kernel and the former two algorithms are small. For large T, the random kernel algorithm tends to be a bit worse than the GIS and alternating algorithms in the "good" region of the parameter space, but in the "bad" region the differences aren't meaningful.

INSERT SECTION ON TIMINGS – POINT OUT THE BAD TIMINGS IN CERTAIN REGIONS OF THE PARAMETER SPACE FOR ALL ALGORITHMS THAT USE THE SCALED ERRORS OR SCALED DISTURBANCES, THEN USE THIS TO SEGUE INTO THE NORMAL PRIOR ON THE STANDARD DEVIATION. IF THE MIXING RESULTS ARE THE SAME, DON'T SHOW GRAPHS JUST MENTION THIS, THEN SHOW GRAPHS OF TIMINGS WHICH, HOPEFULLY, ARE MUCH FASTER

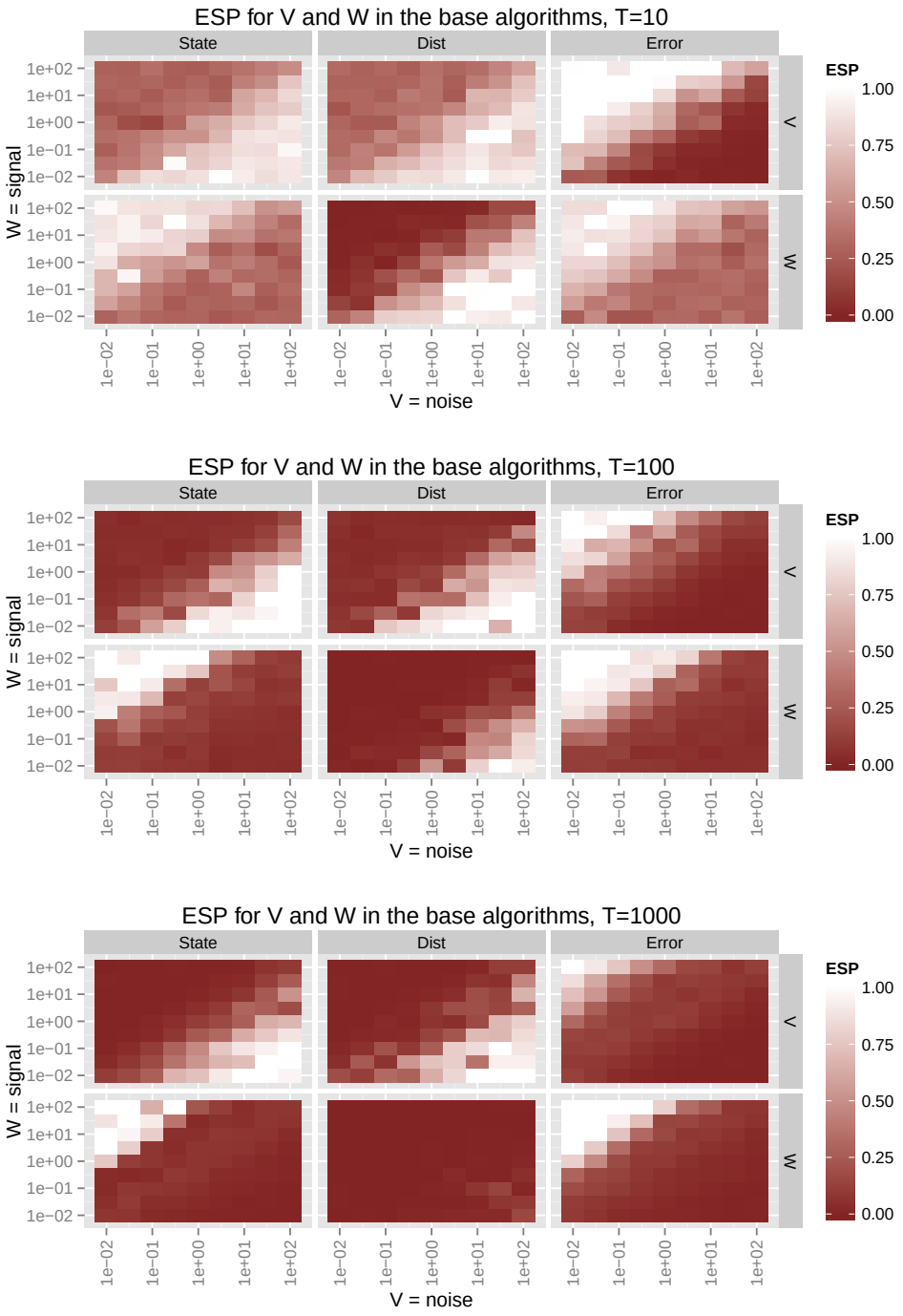

Figure 1: Effective sample proportion in the posterior sampler for a time series of lengths $T = 10$, $T = 100$, and $T = 1000$, for V and W, and for the state, scaled disturbance, scaled error, wrongly scaled disturbance, and wrongly scaleed error samplers. X and Y axes indicate the true values of V 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. Note that for plotting purposes, effective sample proportions larger than 1 were rounded down to 1

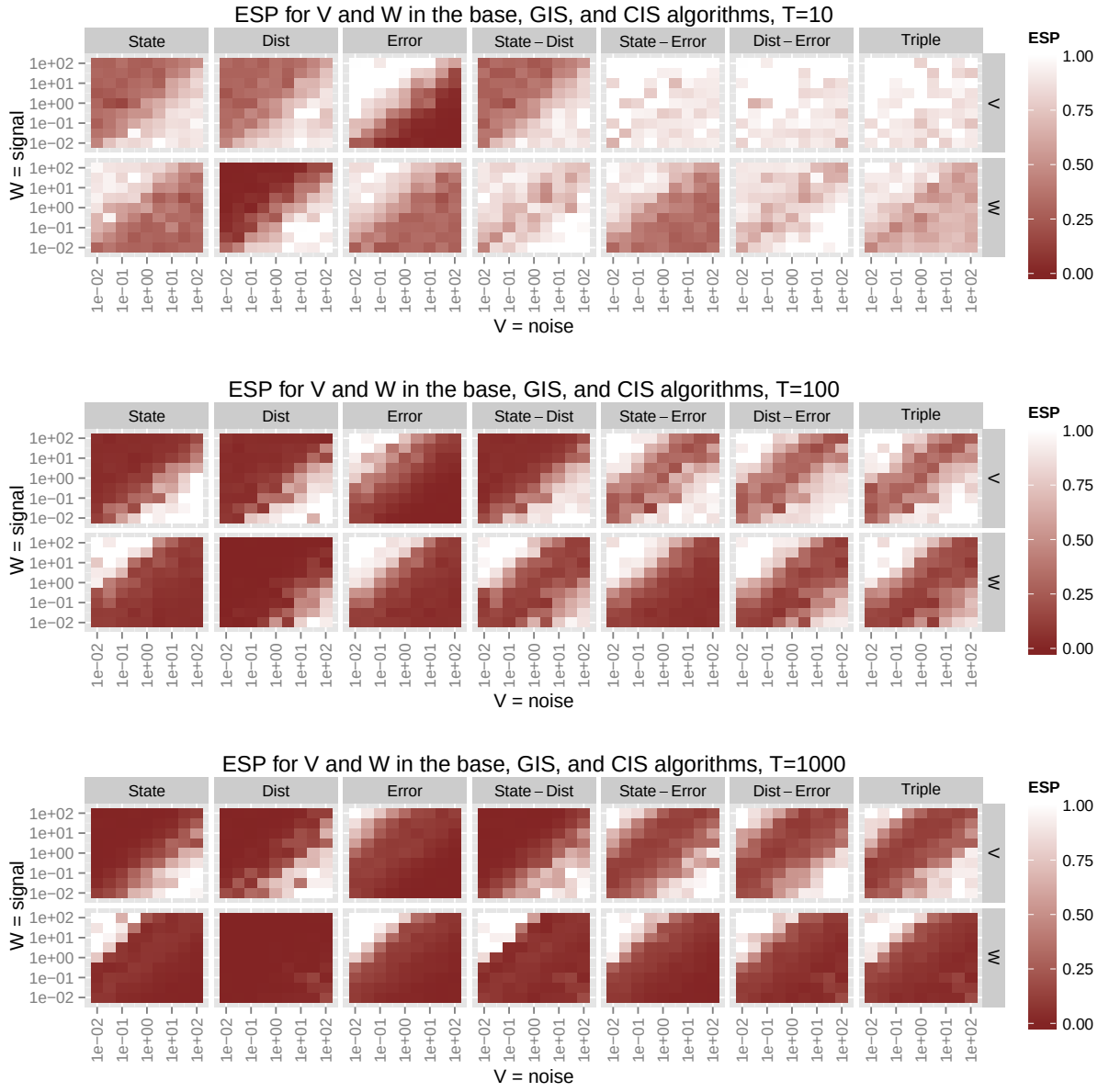

Figure 2: Effective sample proportion in the posterior sampler for V and W in for $T = 10$, $T = 100$, and $T = 1000$, in the state, scaled disturbance and scaled error samplers and for all three GIS samplers based on any two of these. Horizontal and vertical axes indicate the true values of V and W 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. Note that for plotting purposes, effective sample proportions larger than one were rounded down to one.

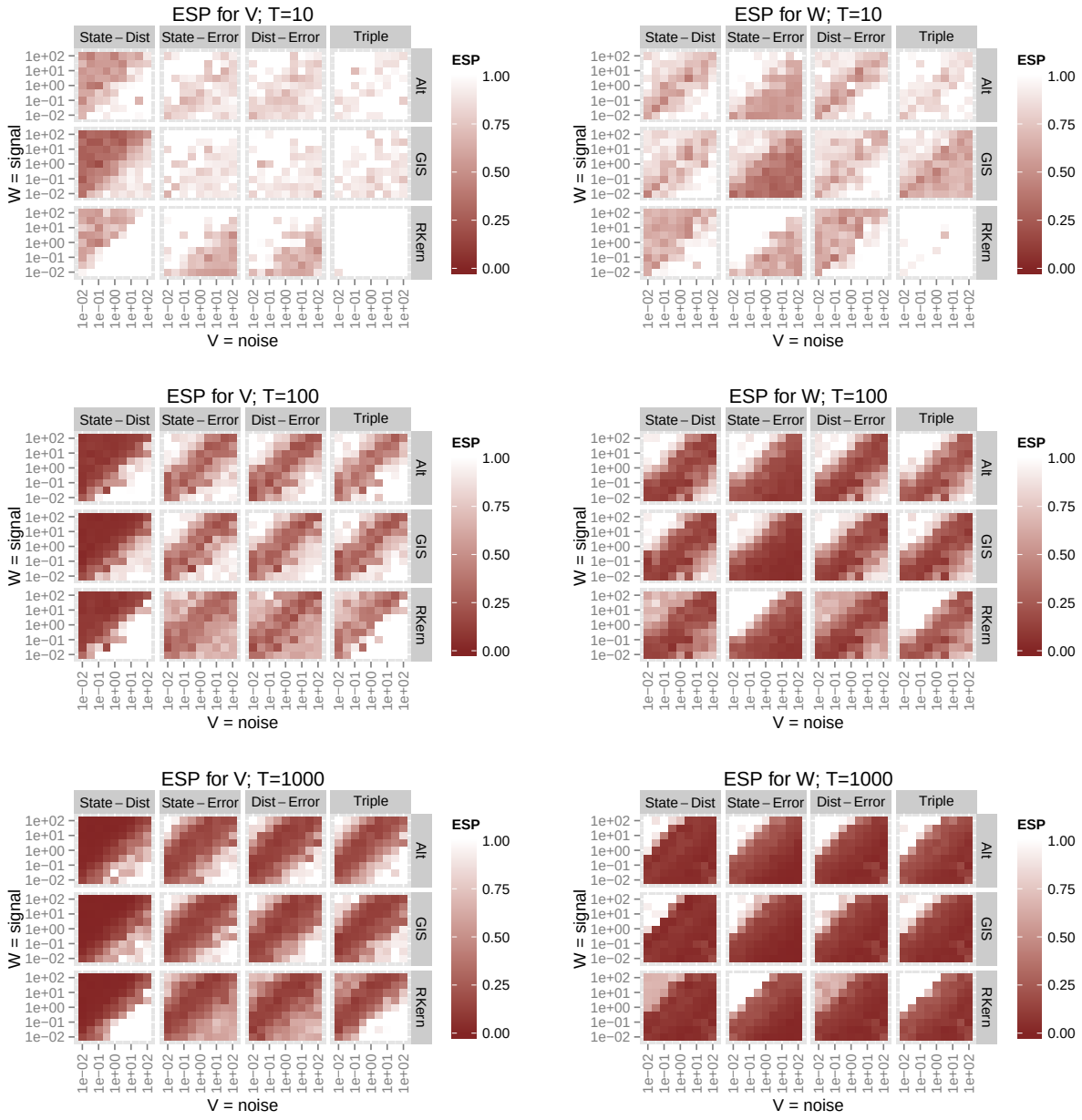

Figure 3: Effective sample proportion in the posterior sampler for a time series of length $T = 10$, $T = 100$, and $T = 1000$, for V and W, and for the GIS and alternating samplers based on the state, scaled disturbance, and scaled error samplers. X and Y axes indicate the true values of V and W respectively 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. Note that for plotting purposes, effective sample proportions larger than 1 were rounded down to 1. Also note that the ESP for the random kernel samplers has been multiplied by 2 or, in the case of the triple kern sampler, by 3, in order to make them comparable to the GIS and alternating samplers.

References

Chris K Carter and Robert Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994.

Piet De Jong and Neil Shephard. The simulation smoother for time series models. *Biometrika*, 82(2):339–350, 1995.

Sylvia Frühwirth-Schnatter. Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2):183–202, 1994.

Sylvia Frühwirth-Schnatter. Efficient Bayesian parameter estimation for state space models based on reparameterizations. *State Space and Unobserved Component Models: Theory and Applications*, pages 123–151, 2004.

Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

James P Hobert and Dobrin Marchev. A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms. *The Annals of Statistics*, 36(2):532–554, 2008.

Siem Jan Koopman. Disturbance smoother for state space models. *Biometrika*, 80(1):117–126, 1993.

Jun S Liu and Ying Nian Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, 1999.

William J McCausland, Shirley Miller, and Denis Pelletier. Simulation smoothing for state–space models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, 55(1):199–212, 2011.

X-L Meng and David Van Dyk. Fast em-type implementations for mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):559–578, 1998.

Xiao-Li Meng and David A Van Dyk. Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86(2):301–320, 1999.

Omiros Papaspiliopoulos, Gareth O Roberts, and Martin Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001.

Yaming Yu and Xiao-Li Meng. 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, 2011.