WARNING: REFERENCES TO OTHER PARTS OF THE DOCUMENT WON'T WORK HERE

1 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

$$y_t|\theta_{0:T} \overset{ind}{\sim} N(\theta_t, V) \tag{1}$$
$$\theta_t|\theta_{0:t-1} \sim N(\theta_{t-1}, W) \tag{2}$$

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 **??** cash out to independent inverse gamma priors for the local level model, viz $V \sim IG(\alpha_V, \beta_V)$ and $W \sim IG(\alpha_W, \beta_W)$.

1.1 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 1 (State Sampler for LLM)**.**

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

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 (**??**), we can write the joint posterior of $(V, W, \gamma_{0:T})$ as

$$p(V, W, \gamma_{0:T}|y_{1:T}) \propto V^{-(\alpha_V+1+T/2)} \exp\left[-\frac{1}{V}\left(\beta_V + \frac{1}{2}\sum_{t=1}^{T}(y_t - \gamma_0 - \sqrt{W}\sum_{s=1}^{t}\gamma_s)^2\right)\right]$$
$$\times W^{-(\alpha_W+1)} \exp\left[-\frac{\beta_W}{W}\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\gamma_t^2\right] \exp\left[-\frac{1}{2C_0}(\gamma_0 - m_0)^2\right] \tag{3}$$

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 2 (Scaled Disturbance Sampler for LLM)**.**

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

In step 2, V is drawn from the same inverse gamma distribution as in step 2 of algorithm 1. 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 Gilks and Wild [1992]. 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 (**??**) we can write $p(V,W,\psi_{0:T}|y_{1:T})$ as

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

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 (3) with V and W switching places. The scaled error sampler obtained from drawing V and W separately is:

Algorithm 3 (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 1. Drawing V in step 2 is more complicated, but exactly analogous to drawing W in algorithm 2. 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

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

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

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

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

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

where now

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

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.

1.2 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 difference 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 1. Note the distributional forms for each of these steps (in some cases a transformation) are in Section 1.1. We omit the partial CIS algorithm, though note that in practice it is essentially the same as the State-Dist algorithm.

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 2 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 3 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

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

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

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.

1.3 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 simulated a dataset for all possible combinations of V and W with each of $T = 10, 100, 1000$. Then for each

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

dataset, we fit the local level model using each algorithm in Table 3. 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.

1.4 Base Results

Figure 1 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 will 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 typical case where the signal to noise ratio close to one, the state sampler has a modest to low ESP for both V and W. Note that the particular values of V and W do not 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 1 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 do not 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 are the preferred data augmentation for low signal-to-noise ratios and the scaled errors are the preferred data augmentation for high signal-to-noise ratios, while the states 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 1 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 $W/V < 1$

Parameter	State	Dist	Error	W-Dist	W-Error
V	$\frac{W}{V} < 1$	$\frac{W}{V} < 1$	$\frac{W}{V} > 1$	$\frac{W}{V} < 1$	$\frac{W}{V} < 1$
W	$\frac{W}{V} > 1$	$\frac{W}{V} < 1$	$\frac{W}{V} > 1$	$\frac{W}{V} > 1$	$\frac{W}{V} > 1$

Table 4: Rule of thumb for when each base algorithm has a high effective sample size for each variable as a function of the signal-to-noise ratio, W/V.

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 $T = 1000$.

The behavior of the wrongly scaled data augmentation algorithms is consistent with what we showed in section **??** — that the Full CIS algorithm based on the scaled errors and disturbances and the wrongly scaled errors and disturbances is equivalent to the full CIS algorithm that replaces the wrongly scaled DAs with the usual latent states. Since the behavior of the state sampler and the wrongly scaled disturbance sampler are the same for W, we might expect that when drawing W, it does not matter whether we use the states or the wrongly scaled disturbances. Similarly since the behavior of the state sampler and the wrongly scaled error sampler are the same for V, we might expect that it does not matter which one we use when drawing V. In fact this is what we found when constructing the Full CIS algorithm. Even though $(\gamma_{0:T}, \tilde{\gamma}_{0:T})$ forms an AA-SA pair for $W|V$ and $(\psi_{0:T}, \tilde{\psi}_{0:T})$ forms an AA-SA pair for $V|W$ while $\theta_{0:T}$ is not a SA for V, replacing $\tilde{\gamma}_{0:T}$ and $\tilde{\psi}_{0:T}$ with $\theta_{0:T}$ does not actually change the CIS algorithm.

We summarize some of the above results for convenience in Table **??**. Most of the patterns of Figure 1 and Table **??** can be explained by Figure 2, which contains the estimated posterior correlations between various functions of parameters estimated using the simulations from the Triple-Alternating sampler. First we need to understand the correlations we are looking at. The state sampler consists of two steps — a draw of $\theta_{0:T}$ given V and W, and a draw of (V, W) given $\theta_{0:T}$. From Section 1.1 we have that conditional on $\theta_{0:T}$, V and W are independent in the posterior and each has an inverse gamma distribution that depends on the states only through the second parameter:

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

So can view (b_V, b_W) as the data augmentation instead of $\theta_{0:T}$ and thus the state sampler is

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

Thus the dependence between (V, W) and (b_V, b_W) in the posterior will determine how much the state sampler moves in a given iteration and, in particular, the dependence in the marginal chain for V is determined by the dependence between b_V and V while the dependence in the marginal chain for W is determined by the dependence between b_W and W.

For the scaled disturbance sampler, things are a bit more complicated. Now the data augmentation

becomes

$$b_V = \beta_V + \sum_{t=1}^{T}(y_t - \gamma_0 - \sqrt{W}\sum_{s=1}^{t}\gamma_t^2)^2/2 = \beta_V + \sum_{t=1}^{T}(y_t - \theta_t)^2/2$$
$$a_\gamma = \sum_{t=1}^{T}(\sum_{j=1}^{t}\gamma_j)^2/2V$$
$$b_\gamma = \sum_{t=1}^{T}(y_t - \gamma_0)(\sum_{j=1}^{t}\gamma_j)/V.$$

This again comes from Section 1.1. The draw of $V|W, \gamma_{0:T}$ is the same inverse gamma draw as in the state sampler. The draw of $W|V, \gamma_{0:T}$ depends on two random parameters, a_γ and b_γ defined above. So the dependence between V and b_V determines how much the marginal chain for V chain moves in a single iteration, while the dependence between W and (a_γ, b_γ) determines how much the marginal chain for W moves in a single iteration. The scaled error sampler is analogous to the scaled disturbance sampler except with

$$b_W = \beta_W + \sum_{t=2}^{T}(Ly_t - \sqrt{V}L\psi_t)^2/2 = \beta_W + \sum_{t=1}^{T}(\theta_t - \theta_{t-1})^2/2$$
$$a_\psi = \sum_{t=1}^{T}(L\psi_t)^2/2W$$
$$b_\psi = \sum_{t=1}^{T}(L\psi_t Ly_t)/W$$

and with the dependence in marginal chain for W depending on the dependence between b_W and W while the dependence in the marginal chain for V depending on the dependence betwee (a_ψ, b_ψ) and V.

In Figure 2 we see that the posterior correlation between V and b_V is high when $W/V > 1$ and is low when $W/V < 1$. This explains why both the state sampler and the scaled disturbance sampler have low ESP's for V when $W/V > 1$. Similarly the posterior correlation between W and b_W is high when $W/V < 1$ and is low when $W/V > 1$, which explains why both the state sampler and scaled error sampler have low ESP's for W when $W/V < 1$. *Insert explanation of marginal chain for W in the scaled disturbance sampler and for the marginal chain for V in the scaled error sampler. Right now it doesn't look right - the correlation between V and a_ψ or b_ψ is highest when W/V is high, but for the scaled error sampler, autocorrelation in the marginal chain for V is highest when W/V is low. However, here's a thought. Consider the marginal chain for W in the scaled disturbance sampler - this chain essentially consists of two steps, a draw of $(a_\gamma, b_\gamma|W)$ and a draw of $(W|a_\gamma, b_\gamma)$. When W is negatively correlated with both a_γ and b_γ, a large draw of W in the chain means that we expect small draws of a_γ and b_γ because of the negative correlation which, in turn, means that we expect a large draw of the next W again because of the negative correlation. This may explain what we are seeing.*

1.5 GIS and CIS Results

Based on the intuition in Section **??** above, both the GIS and alternating 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 autocorrelation in the GIS algrothim using multiple DAs. This suggests that the Dist-Error GIS and alternating algorithms will have the best performance of the GIS and alternating 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 algorithms should have trouble with V when W/V is high since both the state sampler and the scaled

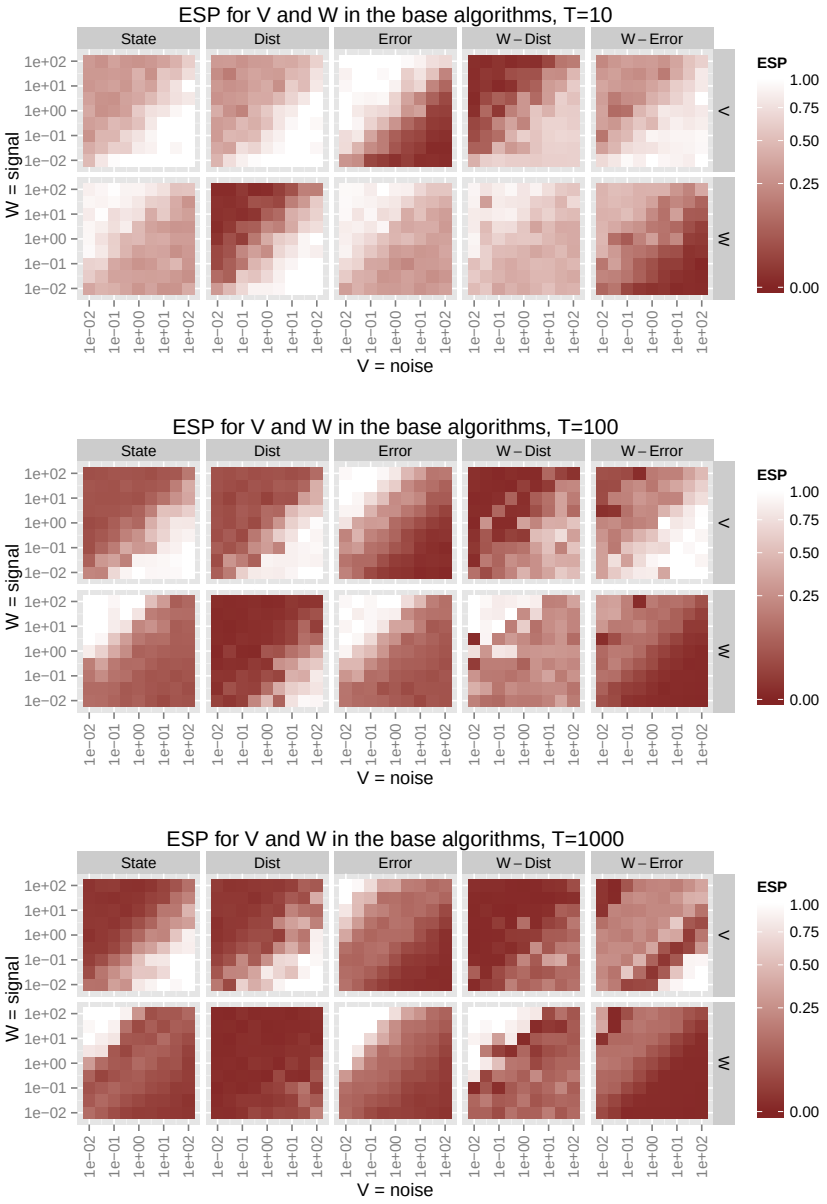

Figure 1: Effective sample proportion in the posterior sampler for a time series of lengths $T = 10$, $T = 100$, and $T = 1000$, for V and W, and for the state, scaled disturbance, scaled error, wrongly scaled disturbance, and wrongly scaleed error samplers. X and Y axes indicate the true values of V 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

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 algorithms add the state sampler into the Dist-Error algorithms, it seems plausible that it might improve mixing for one of V or W since for W/V different from one, the state sampler has good

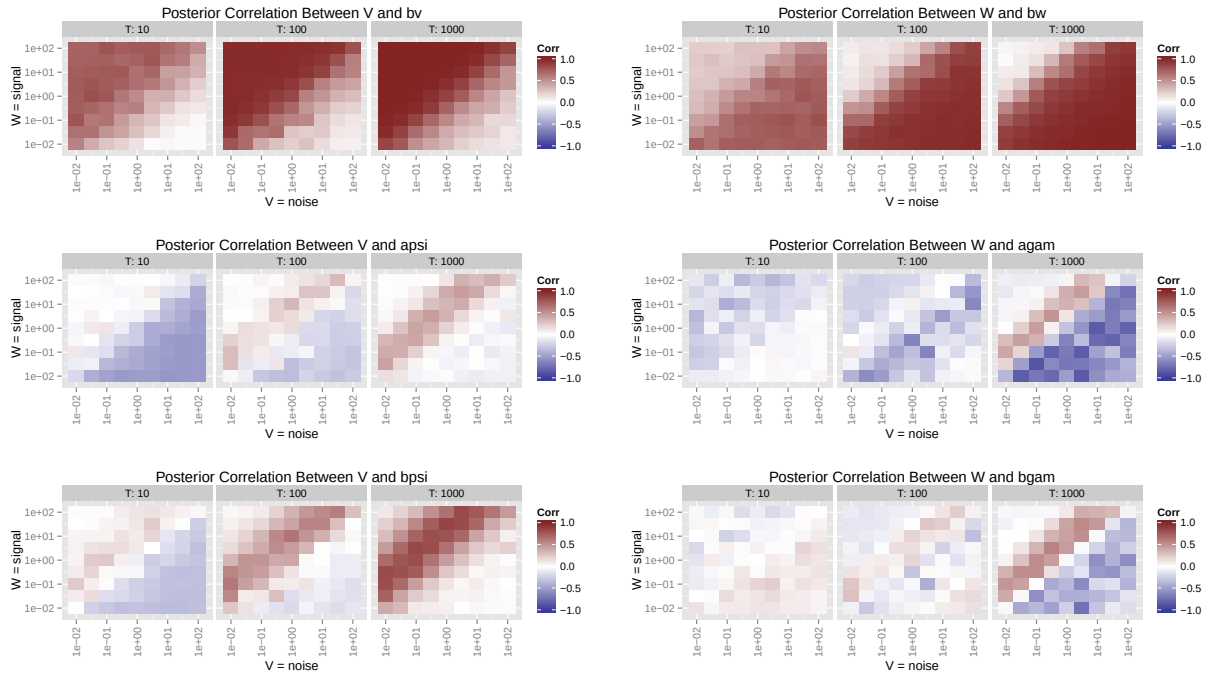

Figure 2: Posterior correlation between V or W and b_V or b_W. X and Y axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

mixing for at least one of V of W.

There are a couple of ways to gain some intuition about what we expect the Full CIS algorithm to do before seeing the results. First, we saw in section **??** that the Full CIS and the Dist-Error GIS algorithm consist of the same steps, just rearranged. This suggests that they should perform similarly so that we expect the Full CIS algorithm to have good mixing for both V and W when W/V is sufficiently different from one. We can draw the same conclusion in a different way by noticing that in the Gibbs step for V, the CIS algorithm interweaves between the states and the scaled errors and in the Gibbs step for W it interweaves between the states and the scaled disturbances. Since the state sampler has a high ESP for V when $W/V < 1$ and the scaled disturbance sampler has a high ESP for V when $W/V > 1$ we should expect the Full CIS sampler to have a high ESP for V when W/V is different from one. Similarly, since the state sampler has a high ESP for W when $W/V > 1$ and the scaled eror sampler has a high ESP for W when $W/V < 1$, we should expect the Full CIS sampler to have a high ESP for W when W/V is different from one.

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. For V, the State-Error algorithm has a high ESP when W/V is far enough away from 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

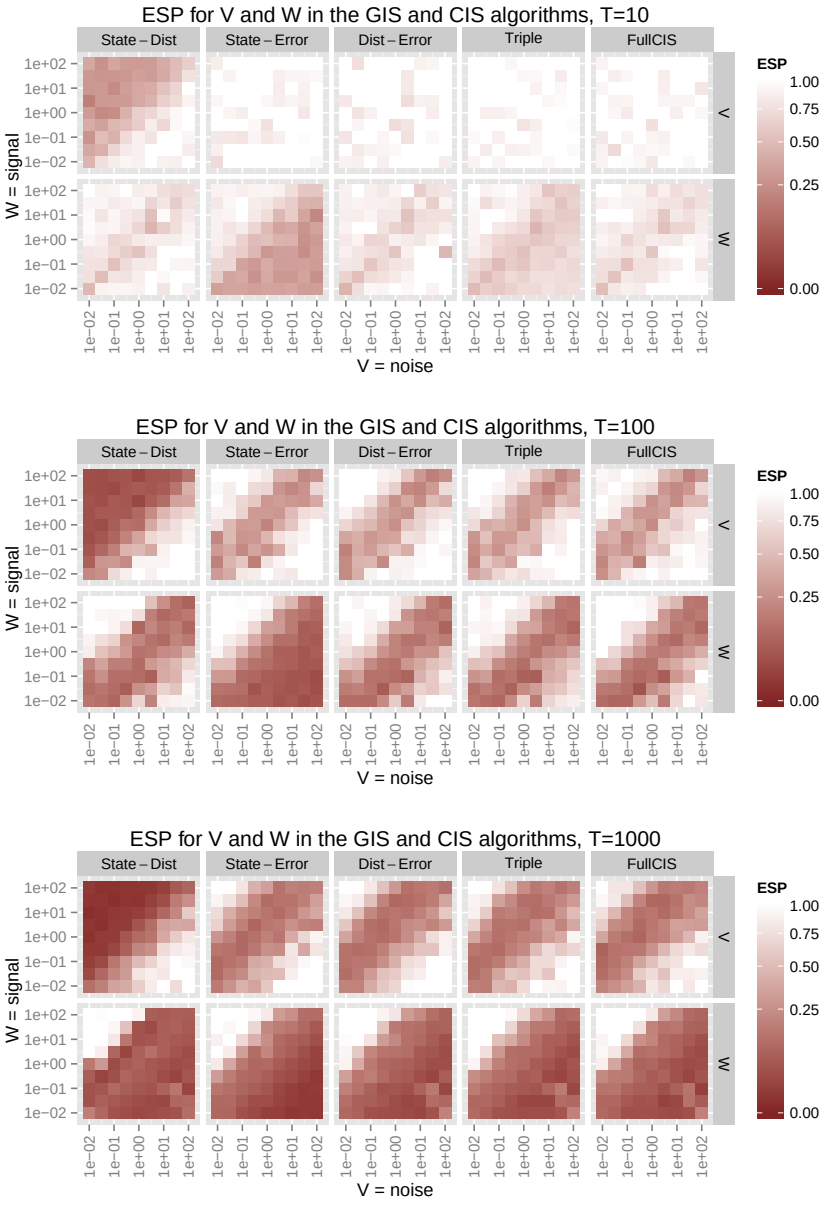

Figure 3: Effective sample proportion in the posterior sampler for V and W in for $T = 10$, $T = 100$, and $T = 1000$, in all three GIS samplers based on any two of the base samplers and the full CIS sampler. Horizontal and vertical axes indicate the true values of V and W respectively for the simulated 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.

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 expected little if any

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 did not happen and furthermore the State-Dist and State-Error samplers did not improve the ESP for V or W respectively. When the underlying DA algorithms form a "beauty and the beast" pair, the interweaving algorithm appears to mix just as well as the best mixing single DA algorithm. Figure 4 allows us to see the ESP of the alternating algorithms in order to compare them to the GIS algorithms. There appears to be little practical difference between the alternating and interweaving versions of a given algorithm based on any two or three of the base DAs.

1.6 Computational time

A more important question than how well the chain mixes from a practical standpoint is the full computational time required to adequately characterize the posterior distribution. In order to investigate this, we compute the natural log of the average time in minutes required for each sampler to achieve an effective sample size of 1000 — in other words the log time per 1000 effective draws. All simulations were performed on *INSERT DETAILS ABOUT IMPACT3*. While different systems will yeld different absolute times, the relative times should still be valid. Figure 5 contains plots of the log time per 1000 effective draws for both V and W and for each of the base and interweaving samplers — note that the scales change as T changes. For $T = 10$ the standard state sampler is competitive with the other samplers over most of the parameter space, though the Triple interweaving and Full CIS algorithms appear to offer some improvement. In any case, for small T computation usually is not a problem, so the difference is practically insignificant.

For $T = 100$ and $T = 1000$ the pattern we saw for ESP begins to emerge for time per 1000 effective draws. The state sampler becomes very slow to reach 1000 effective draws for V when $W/V > 1$ and for W when $W/V < 1$. The scaled disturbance and scaled error samplers behave as expected — the scaled disturbance sampler is slow for both V and W when $W/V > 1$ while the scaled disturbance sapler is slow for both V and W when $W/V < 1$. The Dist-Error GIS, Triple GIS and Full CIS algorithms appear to be the big winners here and are almost indistinguishable. All three algorithms are slightly slower for V when W/V is near one, and when W/V is near or below one all three are slow for W. Compared to the state sampler though, all three offer large gains over most of the parameter space. When we compare the GIS algorithms to their alternating counterparts in terms of log time per 1000 effective draws, again there is little difference. Figure 6 shows the log time per 1000 effective draws for the alternating algorithms and we see essentially the same pattern as we saw for the GIS algorithms.

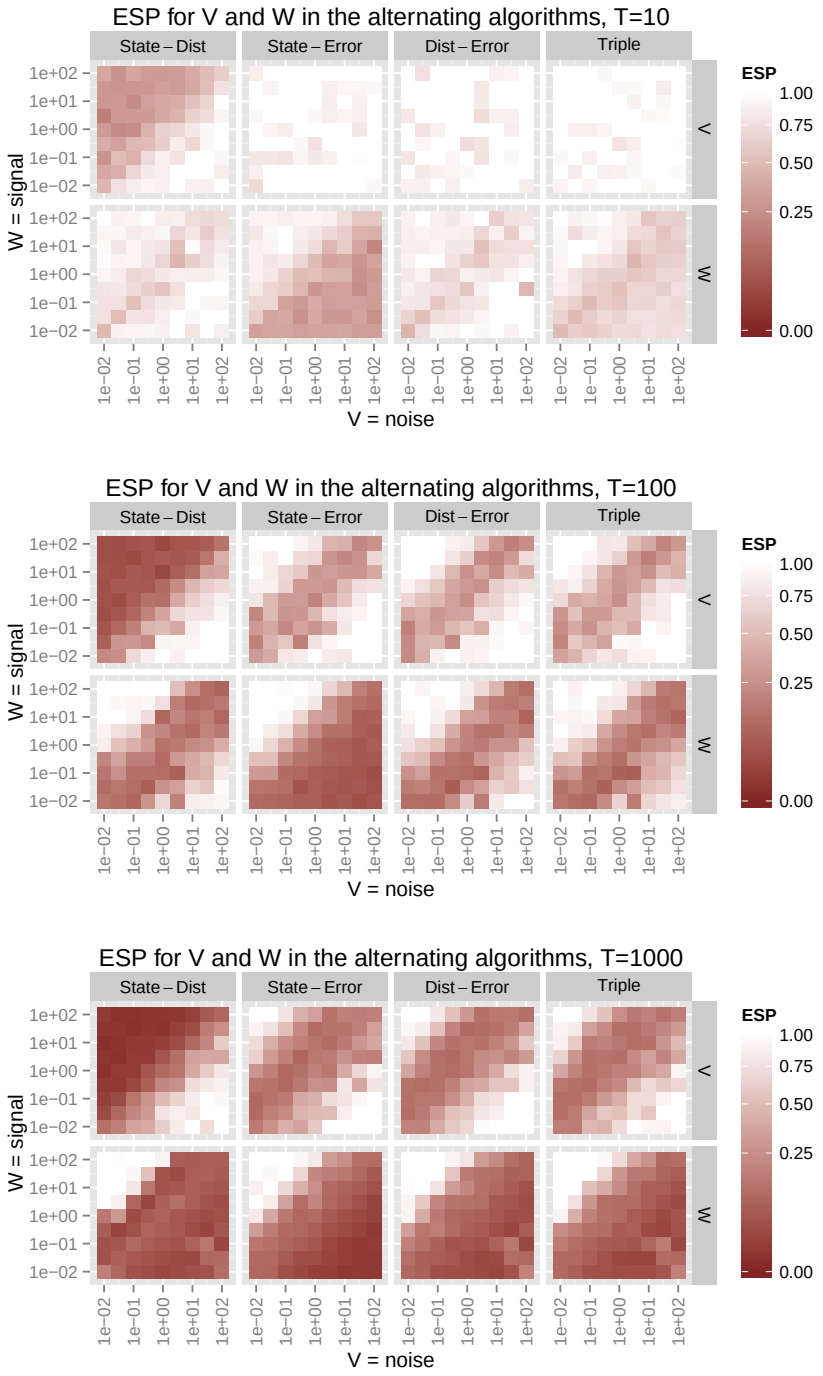

Figure 4: Effective sample proportion in the posterior sampler for V and W in for $T = 10$, $T = 100$, and $T = 1000$, in all four alternating samplers. Horizontal and vertical axes indicate the true values of V and W respectively for the simulated data. Note that the signal-to-noise ratio is constant 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.

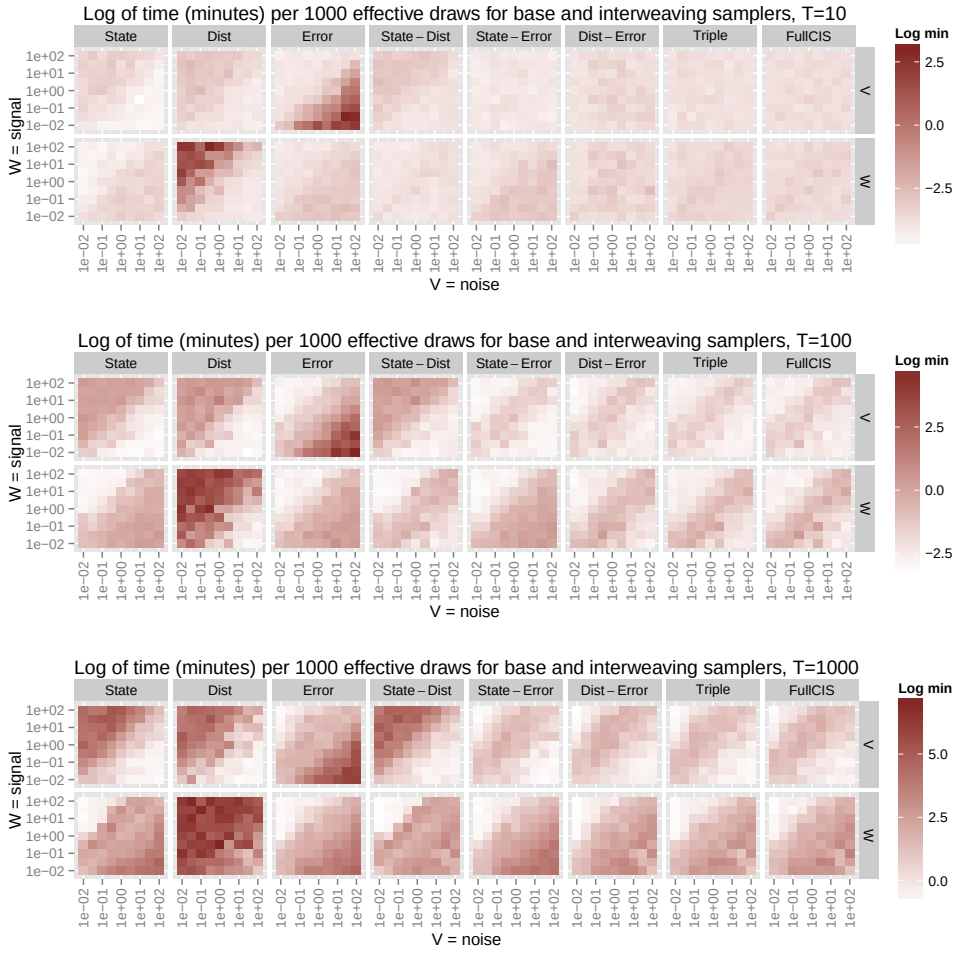

Figure 5: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 10$, $T = 100$, and $T = 1000$, in the state, scaled disturbance and scaled error samplers and for all five interweaving samplers. Horizontal and vertical axes indicate the true values of V and 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. For plotting purposes, times larger than the top of the scale are displayed in bright red.

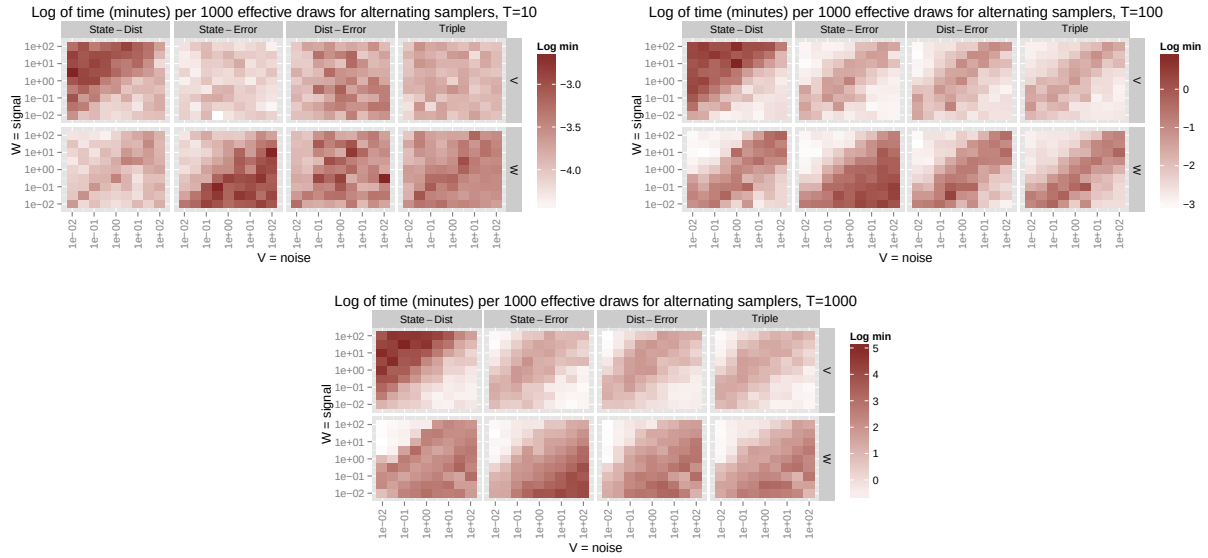

Figure 6: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, for $T = 10$, $T = 100$, and $T = 1000$, in the alternating, GIS, and random kernel samplers. Horizontal and vertical axes indicate the true values of V and W respectively for the simulated data. The signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high. For plotting purposes, times larger than the top of the scale are displayed in bright red.

References

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