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'll focus on the local level model primarily for simplicity. Drawing from $p(W|V, \gamma_{0:T}, y_{1:T})$ and $p(V|W, \psi_{0:T}, y_{1:T})$ in particular is diffult since these turn out not to be of a known distributional form, but the simplicity of the local level model helps to clarify what the issues are. It is possible to implement a metropolis step for the difficult conditional or for (V, W) jointly, but first we would like to see what sort of gains are possible if we sample directly from the desired distributions. 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 $\theta_t = E[y_t|\theta_{0:T}]$ is the average value of y_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)**.**

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$ *using FFBS.*
2. *Draw* $(V^{(k+1)}, W^{(k+1)})$ *from* $p(V, W|\theta_{0:T}, y_{1:T})$*.*

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, 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'll add an extra Gibbs step and draw V and W separately. This gives us the scaled disturbance sampler:

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

1. *Draw* $\gamma_{0:T}$ *from* $p(\gamma_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$*, possibly using FFBS to sample* $\theta_{0:T}$ *then transforming.*
2. *Draw* $V^{(k+1)}$ *from* $p(V|W^{(k)}, \gamma_{0:T}, y_{1:T})$*.*
3. *Draw* $W^{(k+1)}$ *from* $p(W|V^{(k+1)}, \gamma_{0:T}, y_{1:T})$*.*

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 isn't any known form and is difficult to sample from. 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 $b^2 > \frac{32}{9\beta_w}(\alpha_w + 1)^3(1 - 2sgn(b)/3)$ implies that the density is log concave where

$$sgn(b) = \begin{cases} 1 & \text{if } b > 0 \\ 0 & \text{if } b = 0 \\ -1 & \text{if } b < 0. \end{cases}$$

This condition is equivalent to $\partial^2 \log p(W|.)/\partial W^2 < 0$ at the W^* that maximizes $\partial^2 \log p(W|.)/\partial W^2$ and hence guarantees the density is globally 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 isn't 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, but this is much more computationally expensive when necessary.

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)**.**

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$*, possibly using FFBS to sample* $\theta_{0:T}$ *then transforming.*
2. *Draw* $V^{(k+1)}$ *from* $p(V|W^{(k)}, \psi_{0:T}, y_{1:T})$*.*
3. *Draw* $W^{(k+1)}$ *from* $p(W|V^{(k+1)}, \psi_{0:T}, y_{1:T})$*.*

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 $b = \sum_{t=1}^{T}(L\psi_t Ly_t)/W$. So we can use the same methods to sample from this density – adaptive rejection sampling, as in Gilks and Wild [1992], will work as long as $b^2 > \frac{32}{9\beta_V}(\alpha_V + 1)^3(1 - 2sgn(b)/3)$, and otherwise a t proposal in a rejection sampler will work but will be substantially slower.

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 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, there is little to add. We'll 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 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. In Section **??** we saw that one version of the CIS algorithm is the same as the "error-dist" GIS algorithm (i.e. the dist-error algorithm but flipping the order in which the DAs are used). The CIS algorithm below is not the same as the CIS algorith which is equivalent to the error-dist GIS algorithm, but the difference is only the order in which the DAs are used within each Gibbs step. Thus we don't expect it to perform much differently from the dist-error GIS algorithm, but we include it for completeness.

1. state-dist GIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\gamma_{0:T}|V,W,\theta_{0:T}] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}]$
2. state-error GIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\psi_{0:T}|V,W,\theta_{0:T}] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$
3. dist-error GIS algorithm:
   $[\gamma_{0:T}|V,W] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}] \rightarrow [\psi_{0:T}|V,W,\gamma_{0:T}] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$
4. triple GIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\gamma_{0:T}|V,W,\theta_{0:T}] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}] \rightarrow [\psi_{0:T}|V,W,\gamma_{0:T}] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$
5. full CIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V|W,\theta_{0:T}] \rightarrow [\psi_{0:T}|V,W,\theta_{0:T}] \rightarrow [V|W,\psi_{0:T}] \rightarrow [\theta_{0:T}|V,W] \rightarrow [W|V,\theta_{0:T}] \rightarrow [\gamma_{0:T}|V,W] \rightarrow [W|V,\gamma_{0:T}]$

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 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.

1. State-Dist alternating algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\gamma_{0:T}|V,W] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}]$
2. State-Error alternating GIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\psi_{0:T}|V,W] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$
3. Dist-Error alternating GIS algorithm:
   $[\gamma_{0:T}|V,W] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}] \rightarrow [\psi_{0:T}|V,W] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$
4. Triple alternating GIS algorithm:
   $[\theta_{0:T}|V,W] \rightarrow [V,W|\theta_{0:T}] \rightarrow [\gamma_{0:T}|V,W] \rightarrow [V|W,\gamma_{0:T}] \rightarrow [W|V,\gamma_{0:T}] \rightarrow [\psi_{0:T}|V,W] \rightarrow [V|W,\psi_{0:T}] \rightarrow [W|V,\psi_{0:T}]$

Table 2: Alternating algorithms for the local level model

Finally, we also consider random kernel algorithms. In this context, a random kernel algorithm randomly chooses from the state sampler, scaled disturbance sampler, and scaled error sampler in each iteration where the selection probabilities are constant with respect to the iteration. We consider four random kernel algorithms based on any two or three of the base samplers with an equal probability of selecting each base sampler included in the algorithm. For example, the State-Dist random kernel algorithm selects either the state sampler or the scaled disturbance sampler with equal probability at every iteration, while the triple random kernel algorithm selects from the state sampler, scaled disturbance sampler, or the scaled error sampler with equal probability at every iteration.

Table 3 contains each algorithm we considered for the local level model. The basic idea here is that the alternating algorithms and the random kernel 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 shouldn't be large. So we would like the GIS algorithms to have at least as quick mixing as the corresponding alternating algorithms. The random kernel algorithms, however, have do do half as much computation to obtain a single draw (or a third as much computation in the case of the triple random kernel algorithm). Thus in some sense, we would like the GIS algorithms to have mixing which is twice as fast as the corresponding random kernel algorithm, or three times as fast in the case of the triple algorithms. We can make this notion precise by considering the effective sample size (ESS) of the Markov chain – we'd like the GIS algorithms to have an ESS about twice as large as the corresponding random kernel algorithms, or three times as large for the triple algorithms.

Base	State	(wrongly) Scaled Disturbance	(wrongly) Scaled Error	
GIS	State-Dist	State-Error	Dist-Error	Triple
Alt	State-Dist	State-Error	Dist-Error	Triple
RandKern	$\frac{1}{2}$State $+\frac{1}{2}$Dist	$\frac{1}{2}$State $+\frac{1}{2}$Error	$\frac{1}{2}$Dist $+\frac{1}{2}$Error	$\frac{1}{3}$State $+\frac{1}{3}$Dist $+\frac{1}{3}$Error
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

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 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 sample size (ESS) 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's possible to obtain $ESP > 1$ if the draws are negatively correlated and this happens occasionally with some of our samplers, but we round this down to $ESP = 1$ in order to simplify 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'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 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 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 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 $W/V < 1$ 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 $T = 1000$.

*INSERT PARAGRAPH EXPLAINING SOME INTUITION BEHIND *WHY* THESE ALGORITHMS BEHAVE THE WAY THEY DO - MAY NEED TO GO BACK AND RE-READ SOME OF THE PAPERS ON CENTRAL AND NONCENTRAL PARAMETERIZATIONS*

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 Gilks and Wild [1992] 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 **??** 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 2. 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 3 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

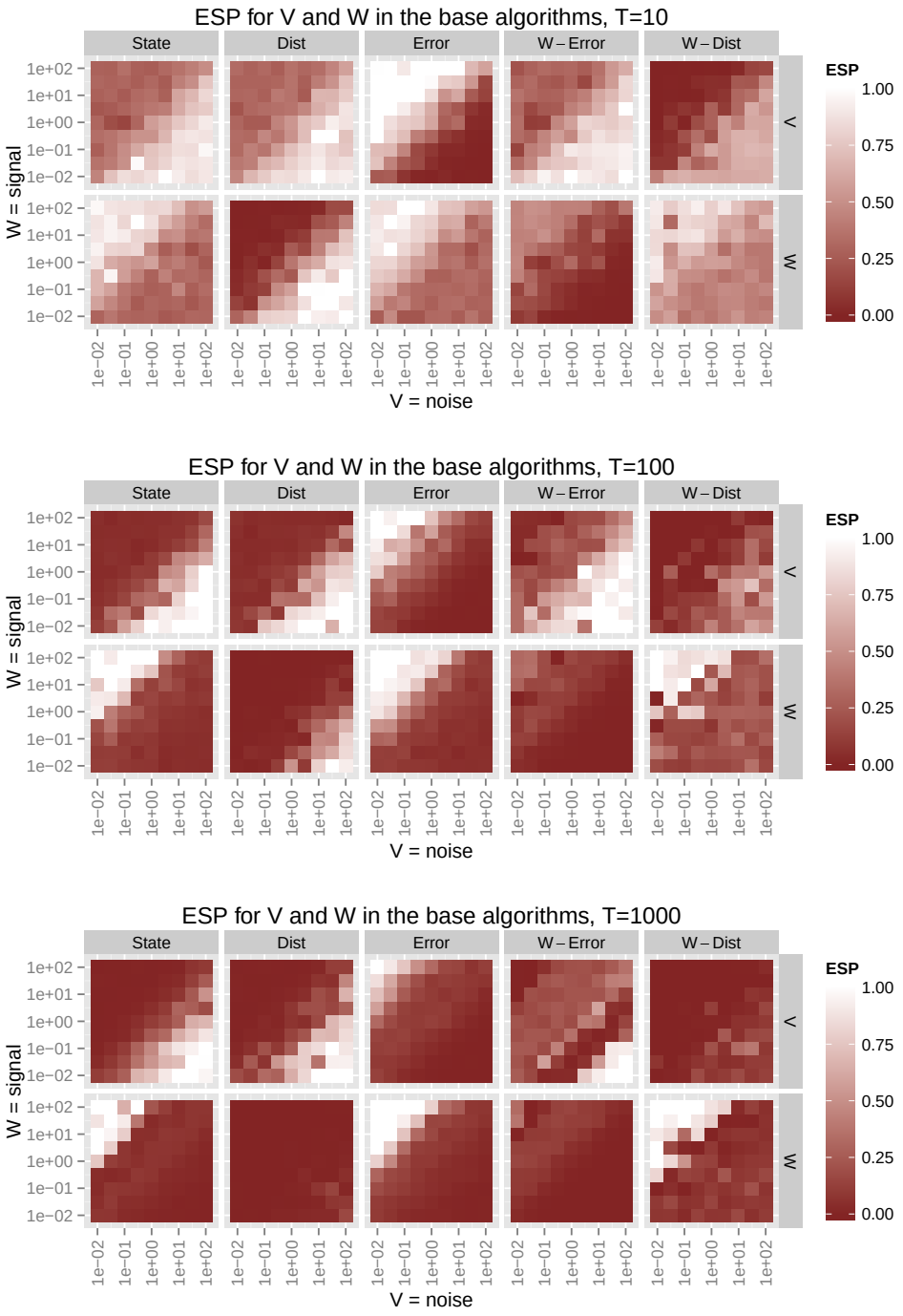

Figure 1: Effective sample proportion in the posterior sampler for a time series of lengths $T = 10$, $T = 100$, and $T = 1000$, for V and W, and for the state, scaled disturbance, scaled error, wrongly scaled disturbance, and wrongly scaleed error samplers. X and Y axes indicate the true values of V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

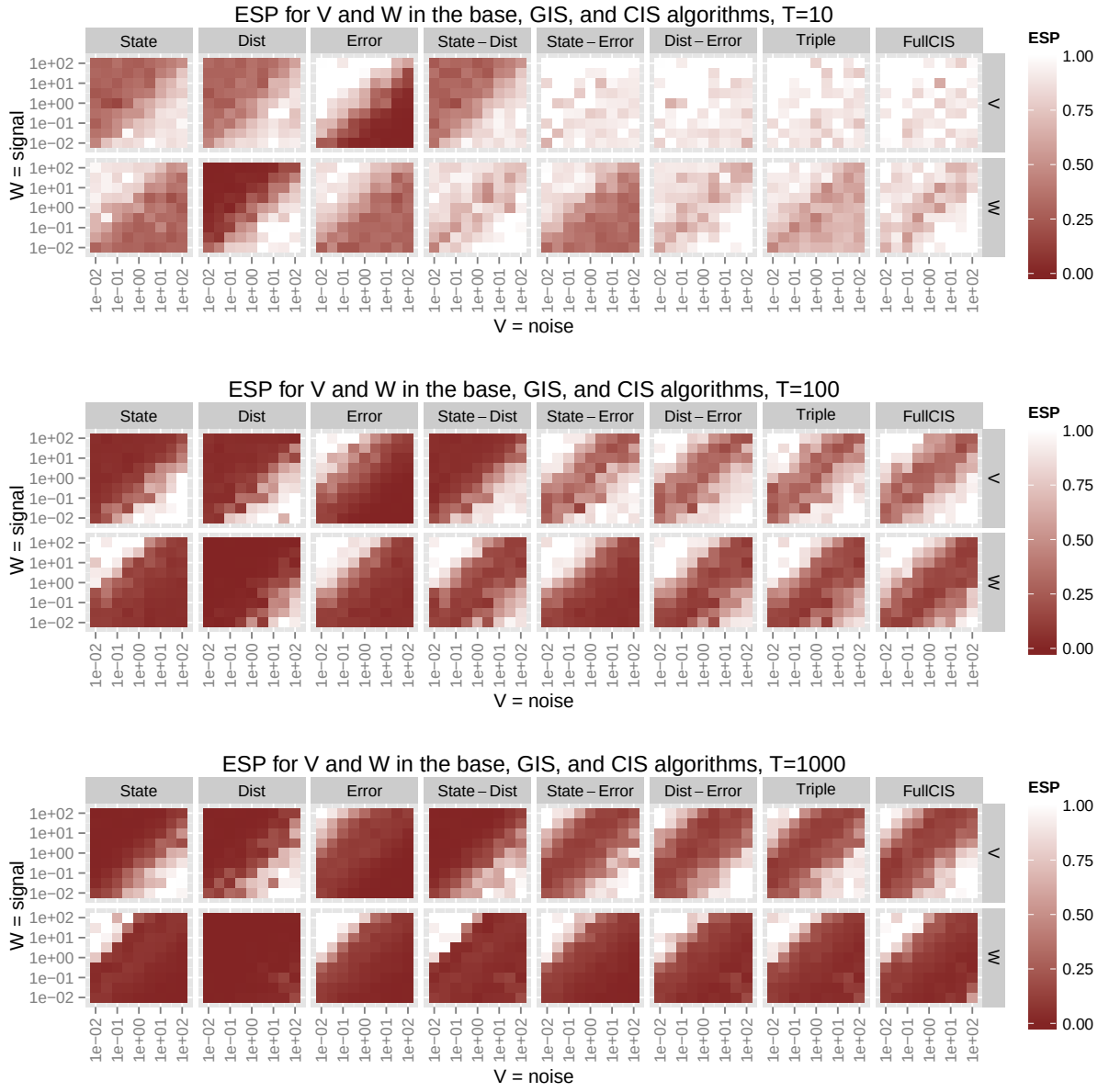

Figure 2: Effective sample proportion in the posterior sampler for V and W in for $T = 10$, $T = 100$, and $T = 1000$, in the state, scaled disturbance and scaled error samplers and for all three GIS samplers based on any two of these. Horizontal and vertical axes indicate the true values of V and W 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.

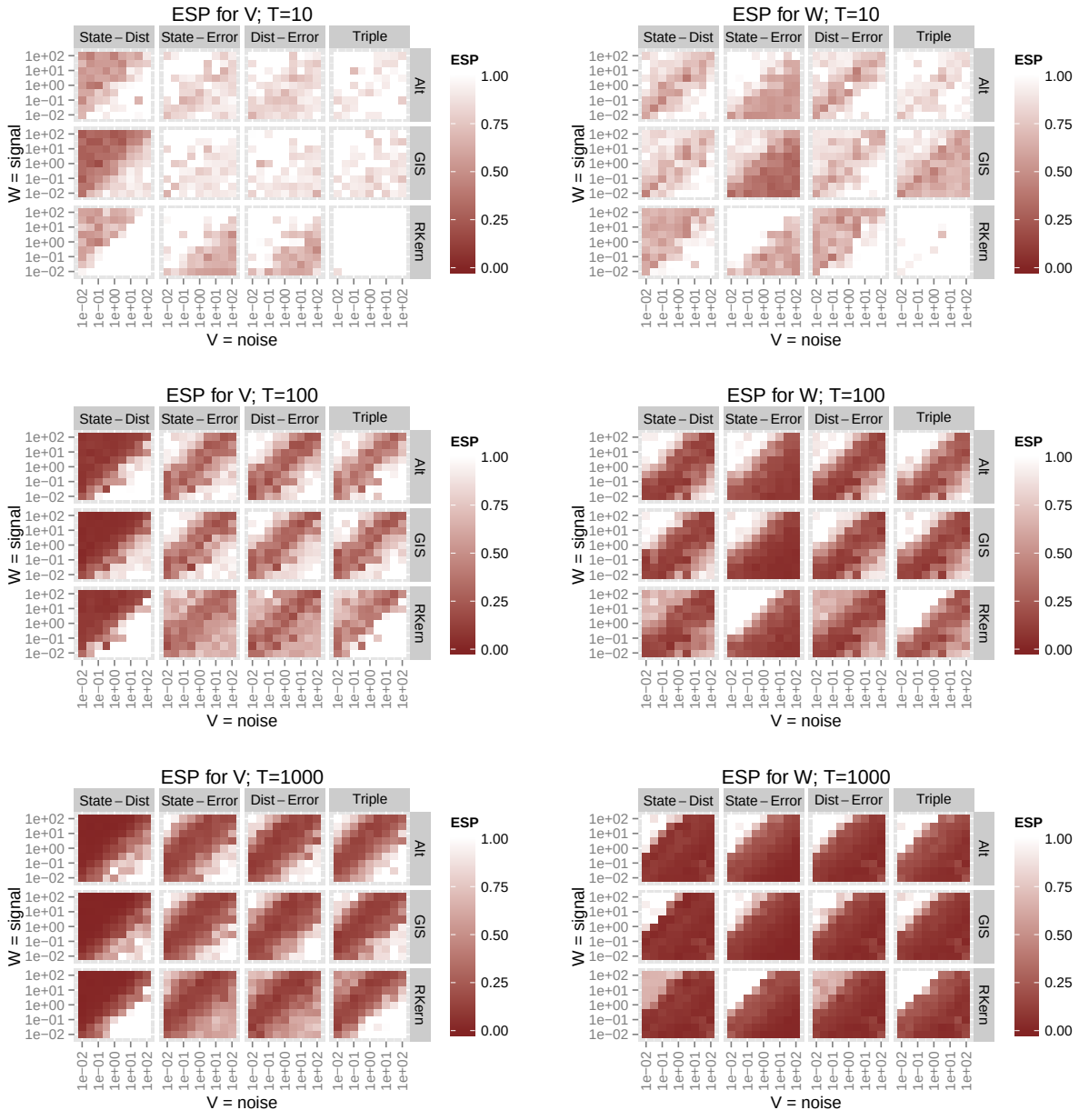

Figure 3: Effective sample proportion in the posterior sampler for a time series of length $T = 10$, $T = 100$, and $T = 1000$, for V and W, and for the GIS and alternating samplers based on the state, scaled disturbance, and scaled error samplers. X and Y axes indicate the true values of V and W respectively 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

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