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

1 Estimating the Model via Data Augmentation: Parameterization Issues

The usual way to estimate the model is via data augmentation (DA) using forward filtering backward sampling (FFBS), as in Frühwirth-Schnatter [1994] and Carter and Kohn [1994]. 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.

1. *Draw* θ *from* $p(\theta|\phi^{(k)}, y)$
2. *Draw* $\phi^{(k+1)}$ *from* $p(\phi|\theta, y)$

The first block samples the states conditional on the data and model parameters while the second block samples the parameters conditonal on the states and the data. We're calling this algorithm the "state sampler." The FFBS step consists of running the Kalman filter to obtain a draw from $\theta_T|V_{1:T}, W_{1:T}, y_{1:T}$, then moving backward to obtain draws from $\theta_t|V_{1:T}, W_{1:T}, y_{1:T}, \theta_{t+1:T}$ for $t = T-1, T-2, \cdots, 0$. Frühwirth-Schnatter [1994], Carter and Kohn [1994], and Petris et al. [2009] contain the details of this process. For the subset of DLMs we are considering, the algorithm cashes out like this:

Algorithm 2.

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

   $V_1, V_2, \cdots, V_T$ *and* $W_1, W_2, \cdots, W_T$ *are conditionally independent given* $(\theta_{0:T}, y_{1:T})$*, with distributions*

   $$V_t|\theta_{0:T}, y_{1:T} \sim IW(\Psi_t + v_t v_t', \eta_t + 1) \qquad W_t|\theta_{0:T}, y_{1:T} \sim IW(\Omega_t + w_t w_t', \delta_t + 1)$$

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

We can immediately see why the "standard priors" are standard – they are conditionally conjugate for each parameter in question so that the full conditional distributions are all easy to sample from. The main problem with this algorithm is that computation time increases quickly with the length of the time series because the Kalman filter essentially requires drawing from $\theta_t|V_{1:T}, W_{1:T}, \theta_{0:t}, y_{0:T}$ for $t = 0, 1, \cdots, T$, so the FFBS step represents $2T$ multivariate draws. A second problem 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 and similar models it's known that if the time constant variance of the latent states, W, is too small, mixing will be poor for W Frühwirth-Schnatter [2004].

One well known method of improving mixing and convergence in MCMC samplers is reparameterizing the model. Papaspiliopoulos et al. [2007] 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 Yu and Meng [2011]. Like Yu and Meng, 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. Some work focuses on using partially noncentered parameterizations that are a sort of bridge between the CP and NCP, e.g. Papaspiliopoulos et al. for general hierarchical models and Frühwirth-Schnatter [2004] in the context of a partiuclar DLM — a dynamic univarite regression with a stationary AR(1) coefficient. But this doesn't quite accomplish what we want because it still picks a single parameterization to use that may depend on the region of the parameter space the posterior conentrates most of its mass. The interweaving concept of Yu and Meng [2011] does precisely what we want, however. 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 3.

1. *Draw* θ *from* $p(\theta|\phi^{(k)}, y)$
2. *Draw* $\gamma^{(k+1)}$ *from* $p(\gamma|\theta, y)$
3. *Draw* $\phi^{(k+1)}$ *from* $p(\phi|\gamma^{(k+1)}, y)$.

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 4.

1. *Draw* θ *from* $p(\theta|\phi^{(k)}, y)$
2. *Draw* ϕ *from* $p(\phi|\theta, y)$
3. *Draw* γ *from* $p(\gamma|\theta, \phi, y)$
4. *Draw* $\phi^{(k+1)}$ *from* $p(\phi|\gamma, y)$

When γ is not a one-to-one transformation of θ, step 4 is an update $\gamma = M(\theta; \phi, y)$. The GIS algorithm is directly comparable to an *alternating* algorithm. 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 5.

1. *Draw* θ *from* $p(\theta|\phi^{(k)}, y)$
2. *Draw* ϕ *from* $p(\phi|\theta, y)$
3. *Draw* γ *from* $p(\gamma|\phi, y)$
4. *Draw* $\phi^{(k+1)}$ *from* $p(\phi|\gamma, y)$

The key difference between this algorithm and algorithm 4 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.

Yu and Meng 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 models with a "nice" prior on ϕ in some sense, they also show that the ASIS algorithm is the same as the optimal PX-DA algorithm of Meng and Van Dyk [1999], Liu and Wu [1999], Van Dyk and Meng [2001] and Hobert and Marchev [2008]. Their results suggest that ASIS 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 this is so, 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 ultimate goal here 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 $p(\phi|\theta,y)$ won't move the chain much, resulting in high autocorrelation in the chain. Interweaving helps break this autocorrelation in two ways. First, by inserting the extra step, e.g. steps 2 and 3 together in 4, the chain gets an additional chance to move in a single iteration thereby weaking the autocorrelation. Second, when one of θ and γ is a "beauty" and the other is a "beast", as is often the case when they form a SA-AA pair, one of steps 2 and 4 in algorithm 4 will significantly move the chain even if the other step will not. This intuition suggests that the key isn't so much that θ and γ form a SA-AA pair as that they form a beauty and the beast pair. It just so happens that SA-AA pairs are often great at accomplishing this.

1.1 The Scaled Disturbances

The next step is to apply the ideas of interweaving to sampling from the posterior of the dynamic linear model. Papaspiliopoulos et al. 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 (**??**) and (**??**) we see that $V_{1:T}$ is in the observation equation so that $\theta_{0:T}$ isn't a SA for $(V_{1:T},W_{1:T})$ while $W_{1:T}$ is in the system equation so that $\theta_{0:T}$ isn't an AA for $(V_{1:T},W_{1:T})$ either. In order to find a SA we need to somehow move $V_{1:T}$ from the observation equation (**??**) to the system equation (**??**) while leaving $W_{1:T}$ in the system equation. Alternatively, we could find an AA by somehow moving $W_{1:T}$ from the system equation to the observation equation while leaving $V_{1:T}$ 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 $(V_{1:T},W_{1:T})$. The disturbances $w_{0:T}$ are defined by $w_t=\theta_t-G_t\theta_{t-1}$ for $t=0,1,...,T$ and we define $\theta_{-1}=0$ so that $w_0=\theta_0$. However the DA algorithm based on the w_t's is identical to the algorithm based on the θ_t's. This is because $w_{0:T}$ is a one-to-one function of $\theta_{0:T}$ that doesn't depend on $V_{1:T}$ or $W_{1:T}$, the conditional distributions $p(V_{1:T},W_{1:T}|\theta_{0:T},y_{1:T})$ and $p(V_{1:T},W_{1:T}|w_{0:T},y_{1:T})$ are identical.

Papaspiliopoulos et al. 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 $V_{1:T}$ constant then $\theta_{0:T}$ is a SA for $W_{1:T}$ from the observation and system equations, (**??**) and (**??**), i.e. we say $\theta_{0:T}$ is a SA for $W_{1:T}$ given $V_{1:T}$, or for $W_{1:T}|V_{1:T}$. Similarly $\theta_{0:T}$ is an AA for $V_{1:T}|W_{1:T}$. This suggests that if we scale θ_t by W_t for all t appropriately we'll have an ancillary augmentation for $V_{1:T}$ and $W_{1:T}$ jointly. The same intuition suggests scaling $w_t=\theta_t-G_t\theta_{t-1}$ by W_t for all t appropriately in order to find an ancillary augmentation for $(V_{1:T},W_{1:T})$. We'll work with the latter case though, again these two ideas are ultimately equivalent since the resulting DAs are one-to-one transformations of each other.

To define the scaled disturbances in the general DLM, let L_t denote the Cholesky decomposition of W_t, i.e. $L_t'L_t=W_t$, for $t=1,2,\cdots,T$. Then we'll define the scaled disturbances $\gamma_{0:T}$ by $\gamma_0=\theta_0$ and $\gamma_t=L_t^{-1}(\theta_t-G_t\theta_{t-1})$ for $t=1,2,\cdots,T$. We can confirm our intuition that the scaled disturbances are an AA for $V_{1:T}$ and $W_{1:T}$ jointly. The reverse transformation is defined recursively by $\theta_0=\gamma_0$ and

$\theta_t = L_t\gamma_t + G_t\theta_{t-1}$ for $t = 1, 2, \cdots, T$. Then the Jacobian is block lower triangular with the identity matrix and the L_t's along the diagonal blocks, so $|J| = \prod_{t=1}^{T} |L_t| = \prod_{t=1}^{T} |W_t|^{1/2}$. Then from (**??**) we can write the full joint distribution of $(V_{1:T}, W_{1:T}, \gamma_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V_{1:T}, W_{1:T}, \gamma_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\gamma_0 - m_0)' C_0^{-1}(\gamma_0 - m_0)\right] \\ &\times \prod_{t=1}^{T} |W_t|^{-(\delta_t+p+2)/2} \exp\left[-\frac{1}{2} tr\left(\Omega_t W_t^{-1}\right)\right] \exp\left[-\frac{1}{2}\gamma_t'\gamma_t\right] |V_t|^{-(\eta_t+k+3)/2} \\ &\times \exp\left[-\frac{1}{2}\left(tr\left(\Psi_t V_t^{-1}\right) + \sum_{t=1}^{T} \left[y_t - F_t\theta_t(\gamma_{0:T}, W_{1:T})\right]' V^{-1} \left[y_t - F_t\theta_t(\gamma_{0:T}, W_{1:T})\right]\right)\right] \end{aligned} \tag{1}$$

where $\theta_t(\gamma_{0:T}, W_{1:T})$ denotes the recursive back transformation defined by the scaled disturbances.

So ultimately under the scaled disturbance parameterization we can write the model as

$$\begin{aligned} y_t|\gamma_{0:T}, V_{1:T}, W_{1:T} &\overset{ind}{\sim} N\left(F_t\theta_t(\gamma_{0:T}, W_{1:T}), V_t\right) \\ \gamma_t &\overset{iid}{\sim} N(0, I_p) \end{aligned} \tag{2}$$

for $t = 1, 2, \cdots, T$ where I_p is the $p \times p$ identity matrix. Neither $V_{1:T}$ nor $W_{1:T}$ are in the system equation, so the scaled disturbances are an AA for $(V_{1:T}, W_{1:T})$. This parameterization is well known, e.g. Frühwirth-Schnatter [2004] use it in a dynamic regression model with stationary regression coefficient.

The DA algorithm based on $\gamma_{0:T}$ is as follows:

Algorithm 6.

1. *Draw* $\gamma_{0:T}$ *from* $p(\gamma_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$
2. *Draw* $(V_{1:T}^{(k+1)}, W_{1:T}^{(k+1)})$ *from* $p(V_{1:T}, W_{1:T}|\gamma_{0:T}, y_{1:T})$.

Step 1 can be accomplished directly with the disturbance smoother of Koopman [1993] or indirectly by using FFBS to draw the states and then transform them to the scaled disturbances. Step 2 ends up being complicated because the joint conditional posterior of V and W isn't a known density. We'll go through an example of this when both y_t and θ_t are scalars later.

We previously mentioned that the intuition behind the scaled disturbances also suggests trying the scaled states, i.e. $\theta_{0:T}^s$ where $\theta_0^s = \theta_0$ and for $t = 1, 2, \cdots, T$, $\theta_t^s = L_t^{-1}\theta_t$u. Note that $\theta_{0:T}^s$ and $\gamma_{0:T}$ are completely determined by each other independently of $V_{1:T}$ and $W_{1:T}$, which suggests the conditional distribution of $(V_{1:T}, W_{1:T})$ is unchanged. This intuition is dangerously close to running right into the Borel-Kolmogorov paradox, but in this case there is no issue since the determinant of the Jacobian will be the same whether we are transforming to the scaled disturbances or the scaled states.

1.2 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_t in the general DLM. Now let K_t denote the Cholesky decomposition of V_t, that is $K_t'K_t = V_t$. Then we can define the scaled errors as $\psi_0 = \theta_0$ and $\psi_t = K_t^{-1}(y_t - F_t\theta_t)$ 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".

This isn't where the difficulties end either. With this definition of $\psi_{0:T}$, it isn't straightforward to determine $p(\psi_{0:T}|V_{1:T}, W_{1:T})$, i.e. to write down the model in terms of $\psi_{0:T}$ instead of $\theta_{0:T}$. 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 $\theta_t = F_t^{-1}(y_t - K_t\psi_t)$ 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 $F_t^{-1}K_t$'s along the diagonal, so $|J| = \prod_{t=1}^{T} |F_t|^{-1}|V_t|^{-1/2}$. Then from (**??**) we can write the joint distribution of $(V_{1:T}, W_{1:T}, \psi_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V_{1:T}, W_{1:T}, \psi_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\psi_0 - m_0)' C_0^{-1}(\psi_0 - m_0)\right] \\ &\times \prod_{t=1}^{T} |V_t|^{-(\eta_t+k+2)/2} \exp\left[-\frac{1}{2} tr\left(\Psi_t V_t^{-1}\right)\right] \exp\left[-\frac{1}{2}\psi_t'\psi_t\right] \\ &\times |W_t|^{-(\delta_t+k+3)/2} \exp\left[-\frac{1}{2}\left(tr\left(\Omega_t W_t^{-1}\right) + (y_t - \mu_t)'(F_t W_t F_t')^{-1}(y_t - \mu_t)\right)\right] \end{aligned} \tag{3}$$

where we define $\mu_t = K_t\psi_t + F_t G_t F_{t-1}(y_{t-1} - K_{t-1}\psi_{t-1})$, $y_0 = 0$, $K_0 = I_k$, and $F_0 = I_k$ where I_k is the $k \times k$ identity matrix. The $|F_t|^{-1}$'s have been absorbed into the normalizing constant, but note that if the F_t's depended on some unknown parameter then we couldn't do this. Now we can write the model in terms of the scaled error parameterization:

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

for $t = 1, 2, \cdots, T$. Now we see immediately that the scaled errors, $\psi_{0:T}$, are also an AA for $(V_{1:T}, W_{1:T})$ since neither V nor W are in the system equation of this model, though note that both $V_{1:T}$ and $W_{1:T}$ are in the observation equation, so $\psi_{0:T}$ is not a SA for $(V_{1:T}, W_{1:T})$ or for either one given the other.

The DA algorithm based on $\psi_{0:T}$ is:

Algorithm 7.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T} | V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$
2. *Draw* $(V_{1:T}^{(k+1)}, W_{1:T}^{(k+1)})$ *from* $p(V_{1:T}, W_{1:T} | \psi_{0:T}, y_{1:T})$.

Once again step 1 can be accomplished directly with Koopman's disturbance smoother or indirectly using FFBS. Step 2 is also once again complicated since the joint conditional posterior of $V_{1:T}$ and $W_{1:T}$ isn't a known density.

1.3 Conditionally conjugate priors and the choice of DA

After choosing the priors for θ_0, $V_{1:T}$ and $W_{1:T}$ we motivated the choice by appealing to conditional conjugacy and thus computation. If this is our main concern for choosing a prior, it's worth asking what the conditional conjugate priors are under the scaled disturbances and the scaled errors. We'll look closely at the scaled disturbances, but the scaled errors are analogous. Based on (2) we can write the augmented data likelihood as

$$p(y_{1:T} | \gamma_{0:T}, V_{1:T}, W_{1:T}) \propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \gamma_t'\gamma_t\right] \prod_{t=1}^{T} |V_t|^{-1/2} \exp\left[-\frac{1}{2}\left(y_t - F_t\theta_t(\gamma_{0:T}, W_{1:T})\right)' V_t^{-1}\left(y_t - F_t\theta_t(\gamma_{0:T}, W_{1:T})\right)\right].$$

Immediately we see that the conjugate prior for V_t is inverse Wishart, so no change on that front. For W_t on the other hand, it's unclear until we unpack $\theta_t(\gamma_{0:T}, W_{1:T})$. Recall that in our definition of the scaled disturbances for $t = 1, 2, \cdots, T$, $\gamma_t = L_t^{-1} w_t = L_t^{-1}(\theta_t - G_t\theta_{t-1})$ where $L_t' L_t = W_t$ while $\gamma_0 = \theta_0$. The

reverse transformation is thus the recursion $\theta_t = L_t\gamma_t + G_t\theta_{t-1}$ for $t = 1, 2, \cdots, T$. This implies that for $t = 0, 1, \cdots, T$

$$\theta_t = \sum_{s=1}^{t} \left(\prod_{r=s+1}^{t} G_r \right) L_s\gamma_s + \prod_{r=1}^{t} G_r\gamma_0$$

where for $s + 1 > t$, $\prod_{r=s+1}^{t} G_r = I_k$ the $k \times k$ identity matrix. Now recall that $K_t'K_t = V_t$ and let $H_{st} = \prod_{r=s+1}^{t} G_r$. This allows us to write the full conditional distribution of $W_{1:T}$, ignoring the prior, as

$$p(W_{1:T}|\gamma_{0:T}, \cdots) \propto \exp\left[-\frac{1}{2} \sum_{t=1}^{T} \left(b_t \sum_{s=1}^{t} B_{st}L_s\gamma_s \right)' \left(b_t - \sum_{s=1}^{t} B_{st}L_s\gamma_s \right) \right]$$

where we define $b_t = K_t^{-1}(y_t - F_tH_{0t}\gamma_0)$ and $B_{st} = K_t^{-1}F_tH_{st}$. This doesn't look like it has any conjugate form for W_t, but it looks like a normal kernel for $L_t\gamma_t$ where L_t is the Cholesky decomposition of W_t. It is not immediately clear whether L_t will have a normal distribution, but this turns out to be true. If we follow the approach of Frühwirth-Schnatter and Tüchler [2008] and vectorize L_t by stacking the nonzero elements of each column on top of each other, we will see that the conditionally conjugate prior for the nonzero elements of each of the L_t's is a normal distribution.

In order to show this, we first have to introduce a bit of notation. Let $A = (a_{ij})$ be a $p \times q$ matrix. First we define the vectorization of A, $\text{vec}(A)$, as the $pq \times 1$ column vector obtained by stacking the columns of A on top of each other. If $p = q$ then we define the half vectorization of A, $\text{vech}(A)$, as the $p(p+1)/2 \times 1$ column vector obtained by first deleting the elements of A above the main diagonal, then stacking the remaining elements on top of each other column by column. Formally,

$$\text{vec}(A) = (a_{11}, a_{21}, \cdots, a_{p1}, a_{12}, a_{22}, \cdots a_{p2}, \cdots, a_{pq})'$$
$$\text{vech}(A) = (a_{11}, a_{21}, \cdots, a_{p1}, a_{22}, a_{32}, \cdots a_{p2}, \cdots, a_{pp})'.$$

In particular if a is a column vector, $\text{vec}(a) = a$ and $\text{vec}(a') = \text{vec}(a)'$. Suppose $p = q$ so that A is square. Then there exists a unique $p(p+1)/2 \times p^2$ matrix S_p such that $\text{vech}(A) = S_p \text{vec}(A)$, called the elimination matrix. We list a few of the properties of S_p here, but see Magnus and Neudecker [1980] or Magnus [1988] for more details.

1. $S_pS_p' = I_{p(p+1)/2}$
2. If A is lower triangular $\text{vec}(A) = S_p' \text{vech}(A) = S_p'S_p \text{vec}(A)$.
3. $S_p = \sum_{i \geq j}^{p} u_{ij} \text{vec}(E_{ij})'$ where E_{ij} is a $p \times p$ matrix of zeroes except for a one in the ij'th spot and u_{ij} is a $p(p+1)/2 \times 1$ column vector of zeroes except for a one in the $[(j-1)p + i - j(j-1)/2]$'th spot.

Now since we can assume the Jacobian from $W_t \to L_t$ will be absorbed into the prior, which we are

ignoring, we have

$$
\begin{aligned}
p(L_{1:T}|\gamma_{0:T},\cdots) &\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t-\sum_{s=1}^{t}B_{st}L_s\gamma_s\right)'\left(b_t-\sum_{s=1}^{t}B_{st}L_s\gamma_s\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t-\sum_{s=1}^{t}B_{st}\operatorname{vec}(L_s\gamma_s)\right)'\left(b_t-\sum_{s=1}^{t}B_{st}\operatorname{vec}(L_s\gamma_s)\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t-\sum_{s=1}^{t}B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L_s))\right)'\left(b_t-\sum_{s=1}^{t}B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L_s)\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(\sum_{s=1}^{t}\sum_{r=1}^{t}\operatorname{vec}(L_s)(\gamma_s'\otimes I_p)'B_{st}'B_{rt}(\gamma_r'\otimes I_p)\operatorname{vec}(L_r)-2\sum_{s=1}^{t}b_t'B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L_s)\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{s=1}^{T}\sum_{r=1}^{T}\operatorname{vec}(L_s)'(\gamma_s\gamma_r'\otimes C_{sr})\operatorname{vec}(L_r)-2\sum_{s=1}^{T}c_s'(\gamma_s'\otimes I_p)\operatorname{vec}(L_s)\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\left(\sum_{s=1}^{T}\sum_{r=1}^{T}\operatorname{vech}(L_s)'S_p(\gamma_s\gamma_r'\otimes C_{sr})S_p'\operatorname{vech}(L_r)-2\sum_{s=1}^{T}c_s'(\gamma_s'\otimes I_p)S_p'\operatorname{vech}(L_s)\right)\right]
\end{aligned}
$$

where we define $C_{sr}=\sum_{t=r\vee s}^{T}B_{st}'B_{rt}$ and $c_s=\sum_{t=s}^{T}B_{st}'b_t$ and where $a\vee b=\max(a,b)$. Here we use two properties of the Kronecker product $\otimes$. First for any matrices $A:p\times q$ and $B:q\times r$, $\operatorname{vec}(AB)=(B'\otimes I_p)\operatorname{vec}(A)$. Second, for any $p\times 1$ vectors a and b and any $p\times p$ matrix A, $(a'\otimes I_p)'A(b'\otimes I_p)=ab'\otimes A$.

Now let $L^*=(\operatorname{vech}(L_1)',\cdots,\operatorname{vech}(L_T))$. Then

$$p(L^*|\gamma_{0:T},\cdots)\propto\exp\left[-\frac{1}{2}\left(L^*-\Omega^{-1}\omega\right)'\Omega\left(L^*-\Omega^{-1}\omega\right)\right]$$

where Ω is a $Tp(p+1)/2\times Tp(p+1)/2$ positive definite matrix with $p(p+1)/2\times p(p+1)/2$ blocks

$$\Omega_{sr}=S_p(\gamma_s\gamma_r'\otimes C_{sr})S_p'$$

and ω is a $Tp(p+1)/2\times 1$ vector with $p(p+1)/2\times 1$ blocks

$$\omega_s=S_p(\gamma_s\otimes I_p)\sum_{t=s}^{T}B_{st}'b_t.$$

So the conjugate prior on $L_{1:T}$ is a multivariate normal distribution. This seems a strange since we expect the diagonal elements of L_t to be positive since they are standard deviations, but this is no problem as long as we view this prior as a clever trick for defining a prior on $W_t=L_t'L_t$ so that the sign doesn't matter. Strictly speaking here, we have subtly changed the definition of L_t to a *signed* Cholesky decomposition of W_t, and thus subtly changed the defintion of the γ_t's to take into account the signs of the diagonal elements of L_t. Frühwirth-Schnatter and Tüchler [2008], Frühwirth-Schnatter and Wagner [2011] and *CITE THE DYNAMIC PAPER SYLVIA IS WORKING ON WITH ANGELA AND THE STOCAHSTIC VOLATILITY PAPER BY SYLVIA AND GREGOR KASTNER* use this approach to choosing priors for the system (or hierarchical) variance when working with the scaled disturbances in dynamic and non-dynamic models, but only the first considers the covariance matrix case, rather than the scalar variance case, and that paper does not give the exlicit construction of the covariance matrix of $\operatorname{vech}(L_t)$ using the elimination matrix.

We have two sets of covariance matrices, $W_{1:T}$ and $V_{1:T}$, and we want to put a different class of priors on each set. We can put the same sort of normal prior on K^*, the vectorized Cholesky decompositions of the V_t's. In the univariate case the conditional posterior of V_t will come out to be a generalized inverse gaussian

distribution which is a bit complicated but not awful to draw from. There's mild tension here though – depending on how we choose to write down the model we end up with a different class of prior distributions for at least $W_{1:T}$. Now the reason for this difference is ultimately computation — it is known that sometimes using the scaled disturbances improves mixing in the Markov chain — but ideally computational concerns should not have an effect on inference. It would be nice to unite these priors two priors under a common class without sacrificing their respective computaitonal advantages under the relevant data augmentations.

The above discussion assumes that the covariance matrices are time dependent. Typically this will not be the case, and when it is often $W_{1:T}$ and $V_{1:T}$ will each have a hierachical prior that depends on some hyperparameters, e.g. matrices Ψ_W and Ψ_V, each with their own hyperprior. The typical case, however, is W and V constant over time. The same basic reasoning applies, but the result is much simpler. First, redefine $B_{st} = K^{-1}F_tH_{st}$ and $b_t = K^{-1}(y_t - F_tH_{0t}\gamma_0)$. Then we get

$$
\begin{aligned}
p(L|\gamma_{0:T}, \cdots) &\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(K^{-1}(y_t - F_tH_{0t}\gamma_0) - K^{-1}F_t\sum_{s=1}^{t}H_{st}L\gamma_s\right)'\left(.\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t - \sum_{s=1}^{t}B_{st}L\gamma_s\right)'\left(b_t - \sum_{s=1}^{t}B_{st}L\gamma_s\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t - \sum_{s=1}^{t}B_{st}\operatorname{vec}(L\gamma_s)\right)'\left(b_t - \sum_{s=1}^{t}B_{st}\operatorname{vec}(L\gamma_s)\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(b_t - \sum_{s=1}^{t}B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L)\right)'\left(b_t - \sum_{s=1}^{t}B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L)\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\operatorname{vec}(L)'\sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{r=1}^{t}(\gamma_s\gamma_r'\otimes B_{st}'B_{rt})\operatorname{vec}(L) - 2\sum_{t=1}^{T}b_t'\sum_{s=1}^{t}B_{st}(\gamma_s'\otimes I_p)\operatorname{vec}(L)\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\operatorname{vech}(L)'\Omega\operatorname{vech}(L) - 2\omega'\operatorname{vech}(L)\right)\right]
\end{aligned}
$$

where we define

$$\Omega = S_p\sum_{t=1}^{T}\sum_{s=1}^{t}\sum_{r=1}^{t}(\gamma_s\gamma_r'\otimes B_{st}'B_{rt})S_p' = \sum_{s=1}^{T}\sum_{r=1}^{T}S_p(\gamma_s\gamma_r'\otimes C_{sr})S_p'$$

and

$$\omega = S_p\sum_{t=1}^{T}\sum_{s=1}^{t}(\gamma_s\otimes I_p)B_{st}'b_t = \sum_{s=1}^{T}S_p(\gamma_s\otimes I_p)c_s.$$

So again the condiitonally conjugate prior for the nonzero elements of L is a multivariate normal distribution.

For completeness, we show that the conditionally conjugate prior for $K_{1:T}$ and, in the case of time-constant variances K, is also a multivariate normal distribution. From (3) we have

$$p(K_{1:T}|\psi_{0:T}, \cdots) \propto \exp\left[-\frac{1}{2}g(K_{1:T})\right]$$

where

$$g(K_{1:T}) = A + \sum_{t=1}^{T} \left[y_t - K_t\psi_t - F_tG_tG_{t-1}\left(y_{t-1} - K_{t-1}\psi_{t-1}\right)\right]' \left(F_t'W_tF_t\right)^{-1} \left[y_t - K_t\psi_t - F_tG_tG_{t-1}\left(y_{t-1} - K_{t-1}\psi_{t-1}\right)\right]$$
$$= A + \sum_{t=1}^{T} \left[y_t - K_t\psi_t - H_t\left(y_{t-1} - K_{t-1}\psi_{t-1}\right)\right]' D_t \left[y_t - K_t\psi_t - H_t\left(y_{t-1} - K_{t-1}\psi_{t-1}\right)\right]$$

where A is some constant with respect to $K_{1:T}$, $D_t = (F_t'W_tF_t)^{-1}$, $H_t = F_tG_tF_{t-1}$, $K_0 = I_k$, $F_0 = I_k$, and $y_0 = 0$. From this point forward we will ignore the constant A since it will be absorbed into the proportionality constant for $p(K_{1:T}|\psi_{0:T}, \cdots)$. Now up to an additive constant we have

$$\begin{aligned}
g(K_{1:T}) = &\sum_{t=1}^{T} \left(y_t - (\psi_t' \otimes I_k)S_p' \operatorname{vech}(K_t)\right)' D_t \left(y_t - (\psi_t' \otimes I_k)S_p' \operatorname{vech}(K_t)\right) \\
&+ \sum_{t=2}^{T} \left(y_{t-1} - (\psi_{t-1}' \otimes I_k)S_p' \operatorname{vech}(K_{t-1})\right)' H_t'D_tH_t \left(y_{t-1} - (\psi_{t-1}' \otimes I_k)S_p' \operatorname{vech}(K_{t-1})\right) \\
&+ \sum_{t=2}^{T} \left(y_t - (\psi_t' \otimes I_k)S_p' \operatorname{vech}(K_t)\right)' D_tH_t \left(y_{t-1} - (\psi_{t-1}' \otimes I_k)S_p' \operatorname{vech}(K_{t-1})\right) \\
&+ \sum_{t=2}^{T} \left(y_{t-1} - (\psi_{t-1}' \otimes I_k)S_p' \operatorname{vech}(K_{t-1})\right)' H_t'D_t \left(y_t - (\psi_t' \otimes I_k)S_p' \operatorname{vech}(K_t)\right) \\
&+ 2\operatorname{vech}(K_1)'S_p(\psi_1' \otimes I_k)'D_1H_1\psi_0.
\end{aligned}$$

But if we write

$$g(K_{1:T}) = (K^*)'\Omega K^* - 2\omega' K^*$$

where $K^* = (\operatorname{vech}(K_1)', \cdots, \operatorname{vech}(K_T)')'$ then we can identify blocks of Ω with the cross product terms in g and blocks of ω with the single product terms in g. Specifically, Ω and ω are defined by blocks

$$\begin{aligned}
\Omega_{TT} &= S_p(\psi_T\psi_T' \otimes D_T)S_p' \\
\Omega_{tt} &= S_p\left[\psi_t\psi_t' \otimes (D_t + H_{t+1}'D_{t+1}H_{t+1})\right]S_p' && \text{for } t = 1, 2, \cdots, T-1 \\
\Omega_{t,t-1} &= S_p(\psi_t\psi_{t-1}' \otimes D_tH_t)S_p' && \text{for } t = 1, 2, \cdots, T-1 \\
\Omega_{t-1,t} &= S_p(\psi_{t-1}\psi_t' \otimes H_t'D_t)S_p' = \Omega_{t,t-1}' && \text{for } t = 1, 2, \cdots, T-1 \\
\omega_1 &= S_p(\psi_1 \otimes I_k)(D_1y_1 + H_2'D_2H_2y_1 + H_2'D_2y_2 - H_1'D_1\psi_0) \\
\omega_T &= S_p(\psi_T \otimes I_k)(D_TY_T + D_TH_Ty_{T-1}) \\
\omega_t &= S_p(\psi_t \otimes I_k)(D_ty_t + H_{t+1}'D_{t+1}H_{t+1}y_t + D_tH_ty_{t-1} + H_{t+1}'D_{t+1}y_{t+1}) && \text{for } t = 2, 3, \cdots, T-1
\end{aligned}$$

where $\Omega_{t,t-i}$ is a matrix of zeroes for $i > 1$. Then given a prior proportional to 1, $K^*|\psi_{0:T}, \cdots \sim N(\Omega^{-1}\omega, \Omega^{-1})$.

For the time constant K case, redefine $D_t = (F_t' W F_t)^{-1}$. Then up to an additive constant we have

$$\begin{aligned}
g(K) = &\sum_{t=1}^{T} \left(y_t - (\psi_t' \otimes I_k) S_p' \operatorname{vech}(K)\right)' D_t \left(y_t - (\psi_t' \otimes I_k) S_p' \operatorname{vech}(K)\right) \\
&+ \sum_{t=2}^{T} \left(y_{t-1} - (\psi_{t-1}' \otimes I_k) S_p' \operatorname{vech}(K)\right)' H_t' D_t H_t \left(y_{t-1} - (\psi_{t-1}' \otimes I_k) S_p' \operatorname{vech}(K)\right) \\
&+ \sum_{t=2}^{T} \left(y_t - (\psi_t' \otimes I_k) S_p' \operatorname{vech}(K)\right)' D_t H_t \left(y_{t-1} - (\psi_{t-1}' \otimes I_k) S_p' \operatorname{vech}(K)\right) \\
&+ \sum_{t=2}^{T} \left(y_{t-1} - (\psi_{t-1}' \otimes I_k) S_p' \operatorname{vech}(K)\right)' H_t' D_t \left(y_t - (\psi_t' \otimes I_k) S_p' \operatorname{vech}(K)\right) \\
&+ 2 \operatorname{vech}(K)' S_p (\psi_1' \otimes I_k)' D_1 H_1 \psi_0.
\end{aligned}$$

This gives

$$\Omega = \sum_{t=1}^{T} S_p(\psi_t \psi_t' \otimes D_t) S_p' + \sum_{t=1}^{T-1} S_p'(\psi_t \psi_t' \otimes H_{t+1}' D_{t+1} H_{t+1}) S_p' + 2 \sum_{t=1}^{T-1} S_p(\psi_t \psi_{t+1}' \otimes H_{t+1}' D_{t+1}) S_p'$$

and

$$\begin{aligned}
\omega = &\sum_{t=1}^{T-1} S_p(\psi_t \otimes I_k)(D_t + H_{t+1}' D_{t+1} H_{t+1}) y_t + \sum_{t=1}^{T-1} S_p(\psi_{t+1} \otimes I_k) D_{t+1} H_{t+1} y_t \\
&+ \sum_{t=1}^{T-1} S_p(\psi_t \otimes I_k) H_{t+1}' D_{t+1} y_{t+1} + S_p \left[(\psi_T \otimes I_k) D_T y_T - (\psi_1 \otimes I_k) D_1 H_1 \psi_0\right]
\end{aligned}$$

so that if K has a prior proportional to 1, then $\operatorname{vech}(K)|\psi_{0:T}, \cdots \sim N(\Omega^{-1}\omega, \Omega^{-1})$. There are efficient routines for drawing from a $N(\Omega^{-1}\omega, \Omega^{-1})$ distribution when Ω is block tribanded as for K and $K_{1:T}$ above. For example, McCausland et al. [2011] describes these routines in the context of drawing the latent states $(\theta_{0:T})$ in a DLM.

2 Interweaving in the DLM: Global and Componentwise

We now have three 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 three DAs are the states, $\theta_{0:T}$, the scaled disturbances $\gamma_{0:T}$, and the scaled errors $\psi_{0:T}$. This allows us to construct four separate GIS algorithms based on algorithm 4: three algorithms that interweave between any two of $\theta_{0:T}$, $\gamma_{0:T}$, and $\psi_{0:T}$, the state-dist, state-error, and dist-error interweaving algorithms, and one algorithm that interweaves between all three, the triple interweaving algorithm. Strictly speaking, the order in which we sample the DAs in the algorithm does matter, but Yu and Meng note that this tends not to make much difference. So while we actually have twelve separate GIS samplers (two of each GIS sampler depending on two DAs, and six GIS samplers depending on all three), effectively we only have four. For example, algorithm 8 is the state-dist GIS algorithm:

Algorithm 8.

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$
2. *Draw* $(V_{1:T}^{(k+0.5)}, W_{1:T}^{(k+0.5)})$ *from* $p(V_{1:T}, W_{1:T}|\theta_{0:T}, y_{1:T})$

3. *Update* $\gamma_{0:T}^{(k+1)}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = L_t^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$

4. *Draw* $(V_{1:T}^{(k+1)}, W_{1:T}^{(k+1)})$ *from* $p(V_{1:T}, W_{1:T}|\gamma_{0:T}, y_{1:T})$

where again $L_t^{(k+0.5)}$ is the Cholesky decomposition of $(W_t^{(k+0.5)})^{-1}$, i.e. $(L_t^{(k+0.5)})'L_t^{(k+0.5)} = (W_t^{(k+0.5)})^{-1}$. Steps 1 and 2 are the same as steps 1 and 2 in algorithm 2, and step 4 is the same as step 2 of algorithm 6. The triple interweaving algorithm is the same as algorithm 8 except it adds two more steps at the end: an update of $\psi_{0:T}$ from $\gamma_{0:T}$ and the draw of $(V_{1:T}, W_{1:T})$ in step 4, and then a draw from $(V_{1:T}, W_{1:T}|\psi_{0:T}, y_{1:T})$. In practice we may want to break up step 4 into two steps if it's easier to draw from the full conditionals of $V_{1:T}$ and $W_{1:T}$ rather than drawing them jointly. Algorithm 9 below is exactly this.

Algorithm 9.

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$

2. *Draw* $(V_{1:T}, W_{1:T})$ *from* $p(V_{1:T}, W_{1:T}|\theta_{0:T}, y_{1:T})$

3. *Update* $\gamma_{0:T}^{(k+1)}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = L_t^{-1}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$

4. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}, \gamma_{0:T}, y_{1:T})$

5. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$

Step 4 actually draws $V_{1:T}$ from the same density as in step 2, but only the last of the two draws is used for anything in the algorithm. As a result, we can either draw only $W_{1:T}$ in step 2 or step 4 can be ommitted.

None of these GIS algorithms are ASIS algorithms — none of the DAs are a SA for $(V_{1:T}, W_{1:T})$. The states, $\theta_{0:T}$, are a SA for $W_{1:T}|V_{1:T}$ though, so this motivates a CIS algorithm. A partial CIS algorithm is immediate:

Algorithm 10.

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$

2. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \theta_{0:T}, y_{1:T})$

3. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \theta_{0:T}, y_{1:T})$

4. *Update* $\gamma_{1:T}^{(k+1)}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = L_t^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$

5. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V^{(k+1)}, \gamma_{0:T}, y_{1:T})$

This algorithm is actually the same as a version of the state-dist interweaving algorithm, specifically algorithm 9. So we can construct a partial CIS algorithm, but it's actually the exact same algorithm as a GIS algorithm.

With a little more work, we can also construct a full CIS algorithm that also turns out to be the same as another GIS algorithm. Recall that $\gamma_t = L_t^{-1}(\theta_t - G_t\theta_{t-1})$ and $\psi_t = K_t^{-1}(y_t - \theta_t)$ for $t = 1, 2, \cdots, T$ where $L_t'L_t = W_t$ and $K_t'K_t = V_t$. Now define $\tilde{\gamma}_t = K_t^{-1}(\theta_t - G_t\theta_{t-1})$ and $\tilde{\psi}_t = L_t^{-1}(y_t - \theta_t)$ 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. Now we'll show that $\gamma_{0:T}$ and $\tilde{\gamma}_{0:T}$ are an AA-SA pair for $W_{1:T}|V_{1:T}$ while $\psi_{0:T}$ and $\tilde{psi}_{0:T}$ are an AA-SA pair for $V_{1:T}|W_{1:T}$. We've already shown that both $\psi_{0:T}$ and $\gamma_{0:T}$ are AAs for $(V_{1:T}, W_{1:T})$, so we just need to show that $\tilde{\gamma}_{0:T}$ is a SA for $W_{1:T}|V_{1:T}$ and that $\tilde{\psi}_{0:T}$ is a SA for $V_{1:T}|W_{1:T}$.

First consider $\tilde{\gamma}_{0:T}$. If we define $L_0 = K_0 = I_k$ where I_k is the $k \times k$ identity matrix, then $\tilde{\gamma}_t = K_t^{-1}L_t\gamma_t$ for $t = 0, 1, 2, \cdots, T$. The reverse transformation is then $\gamma_t = L_t^{-1}K_t\tilde{\gamma}_t$. The Jacobian is then block diagonal

with $L_t^{-1}K_t$'s along the diagonal. Thus $|J| = \prod_{t=0}^{T} |L_t|^{-1}|K_t| = \prod_{t=1}^{T} |W_t|^{-1/2}|V_t|^{1/2}$. Then from (1) we can write the joint distribution of $(V_{1:T}, W_{1:T}, \tilde{\gamma}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V_{1:T}, W_{1:T}, \tilde{\gamma}_{0:T}, y_{1:T}) \propto & \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1}(\tilde{\gamma}_0 - m_0)\right] \prod_{t=1}^{T} |V_t|^{-(\eta_t+k+2)/2} \exp\left[-\frac{1}{2} tr\left(\Psi_t V_t^{-1}\right)\right] \\ & \times |W_t|^{-1/2} \exp\left[-\frac{1}{2}\left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)' V_t^{-1}\left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)\right] \\ & \times |W_t|^{-(\delta_t+k+2)/2} \exp\left[-\frac{1}{2} tr\left(\Omega_t W_t^{-1}\right)\right] \exp\left[-\frac{1}{2}\tilde{\gamma}_t'(K_t^{-1}W_t(K_t^{-1})')^{-1}\tilde{\gamma}_t\right] \end{aligned} \tag{4}$$

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

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

for $t = 1, 2, \cdots, T$. Since K_t is the Cholesky decomposition of V_t, the observation equation doesn't contain W_t. So $\tilde{\gamma}_{0:T}$ is a SA for $W_{1:T}|V_{1:T}$ and thus $\gamma_{0:T}$ and $\tilde{\gamma}_{0:T}$ form an AA-SA pair for $W_{1:T}|V_{1:T}$. Note also that since W_t and K_t are both in the system equation, $\tilde{\gamma}_{0:T}$ is not an AA for $V_{1:T}$ nor for $W_{1:T}$.

Now consider $\tilde{\psi}_t = L_t^{-1}K_t\psi_t$ for $t = 0, 1, 2, \cdots, T$ where, again, $L_0 = K_0 = I_k$, the $k \times k$ identity matrix. Then $\psi_t = K_t^{-1}L_t\tilde{\psi}_t$ and the Jacobian is block diagonal with $K_t^{-1}L_t$'s along the diagonal. So $|J| = \prod_{t=1}^{T} |V_t|^{-1/2}|W_t|^{1/2}$ and from (3) we can write the joint distribution of $(V_{1:T}, W_{1:T}, \tilde{\psi}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V_{1:T}, W_{1:T}, \tilde{\psi}_{0:T}, y_{1:T}) \propto & \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)' C_0^{-1}(\tilde{\psi}_0 - m_0)\right] \\ & \times \prod_{t=1}^{T} |V_t|^{-(\eta_t+k+2)/2} \exp\left[-\frac{1}{2} tr\left(\Psi_t V_t^{-1}\right)\right] \exp\left[-\frac{1}{2}\tilde{\psi_t}'(L_t^{-1}V_t(L_t^{-1})')^{-1}\tilde{\psi}_t\right] \\ & \times |W_t|^{-(\delta_t+k+2)/2} \exp\left[-\frac{1}{2} tr\left(\Omega_t W_t^{-1}\right)\right] |V_t|^{-1/2} \exp\left[-\frac{1}{2}(y_t - \tilde{\mu}_t)'(F_t W_t F_t')^{-1}(y_t - \tilde{\mu}_t)\right] \end{aligned} \tag{5}$$

where we define $\tilde{\mu}_t = L_t\psi_t + F_t G_t F_{t-1}(y_{t-1} - L_{t-1}\tilde{\psi}_{t-1})$ with, again, $y_0 = 0$ and $L_0 = K_0 = I_k$. In terms of $\tilde{\psi}_{0:T}$, the model is then:

$$\begin{aligned} y_t|V_{1:T}, W_{1:T}, \tilde{\psi}_{0:T}, y_{1:t-1} &\sim N(\tilde{\mu}_t, F_t'W_tF_t) \\ \tilde{\psi}_t &\overset{ind}{\sim} N(0, L_t^{-1}V_t(L_t^{-1})') \end{aligned}$$

for $t = 1, 2, \cdots, T$. Since $\tilde{\mu}_t$ only depends on W_t (through L_t) and not on V_t, $V_{1:T}$ is absent from the observation equation. Thus $\tilde{\psi}_{0:T}$ is a SA for $V_{1:T}|W_{1:T}$ and as a result $\psi_{0:T}$ and $\tilde{\psi}_{0:T}$ form an AA-SA pair for $V_{1:T}|W_{1:T}$. Again that both W_t and V_t are in the system equation so $\tilde{\psi}_{0:T}$ is not an AA for either $V_{1:T}$ or $W_{1:T}$. Now we can construct a full CIS algorithm:

Algorithm 11.

1. *Draw* $\tilde{\psi}_{0:T}$ *from* $p(\tilde{\psi}_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.
2. *Draw* $V_{1:T}^{(k+0.5)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \tilde{\psi}_{0:T}, y_{1:T})$
3. *Update* $\psi_{0:T}$ *from* $\psi_0 = \tilde{\psi}_0$ *and* $\psi_t = (K_t^{-1})^{(k+0.5)}(L_t)^{(k)}\tilde{\psi}_t$ *for* $t = 1, 2, \cdots, T$.
4. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \psi_{0:T}, y_{1:T})$.
5. *Update* $\tilde{\gamma}_{0:T}$ *from* $\psi_{0:T}$, $W_{1:T}^{(k)}$, *and* $V_{1:T}^{(k+1)}$.

6. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \tilde{\gamma}_{0:T}, y_{1:T})$

7. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \tilde{\gamma}_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(K^{(k+1)})_t\tilde{\gamma}_t$ *for* $t = 1, 2, \cdots, T$.

8. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$

Steps 1-4 constitute a Gibbs step for $V_{1:T}$ and steps 5-8 constitute a Gibbs step for $W_{1:T}$. Step 1 can be accomplished by using FFBS to draw the states and transforming appropriately, or by using the disturbance smoother of Koopman [1993] and again transforming appropriately. Note that $L_t^{(k)}$ is the Cholesky decomposition of $W_t^{(k)}$ and $K_t^{(k)}$ is the Cholesky decomposition of $V_t^{(k)}$.

It turns out that $p(W_{1:T}|V_{1:T}, \tilde{\gamma}_{0:T}, y_{1:T})$ and $p(W_{1:T}|V_{1:T}, \theta_{0:T}, y_{1:T})$ are the same density, and $p(V_{1:T}|W_{1:T}, \tilde{\psi}_{0:T}, y_{1:T})$ and $p(V_{1:T}|W_{1:T}^{(k+1)}, \theta_{0:T}, y_{1:T})$ are also the same density. Since $\tilde{\gamma}_t = K_t^{-1}(\theta_t - G_t\theta_{t-1})$ is a one-to-one function of $\theta_{0:T}$ given $V_{1:T}$ with a diagonal Jacobian, the conditional distribution of $W_{1:T}$ does not depend on whether we condition on $\theta_{0:T}$ or $\tilde{\gamma}_{0:T}$. Similar reasoning applies to $W_{1:T}$ given either $\theta_{0:T}$ or $\tilde{\psi}_{0:T}$. The upshot is that step 1 of algorithm 11 can be replaced with a draw from $p(\theta_{0:T}|V_{1:T}, W_{1:T}, y_{1:T})$, and any time we condition on one of the "tilde" variables, we can condition on $\theta_{0:T}$ instead.

Now we can rewrite the full CIS algorithm in terms of $\theta_{0:T}$ instead of the tilde variables. We'll also rearrange the order in which $\theta_{0:T}$ and $\psi_{0:T}$ are used in the Gibbs step for $V_{1:T}$. This rearranging does change the algorithm, but it's still a full CIS algorithm.

Algorithm 12.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.

2. *Draw* $V_{1:T}^{(k+0.5)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \psi_{0:T}, y_{1:T})$.

3. *Update* $\theta_{0:T}$ *from* $\theta_0 = \psi_0$ *and* $\theta_t = y_t - (K_t^{(k+0.5)})\psi_t$ *for* $t = 1, 2, \cdots, T$.

4. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \theta_{0:T}, y_{1:T})$.

5. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \theta_{0:T}, y_{1:T})$.

6. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$.

7. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$.

There is no update step between the two Gibbs steps for $V_{1:T}$ and $W_{1:T}$, i.e. between steps 4 and 5, because the DA is alread in the proper form to draw $W_{1:T}^{(k+0.5)}$ in step 5. The only thing left to show now is that this is the same as a GIS algorithm. Consider the error–dist GIS algorithm that interweaves between the scaled errors and the scaled disturbances:

Algorithm 13.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.

2. *Draw* $(V_{1:T}^{(k+0.5)}, W_{1:T}^{(k+0.5)})$ *from* $p(V_{1:T}, W_{1:T}|\psi_{0:T}, y_{1:T})$.

3. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \psi_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2 \cdots, T$ *for* $\theta_0 = \psi_0$ *and* $\theta_t = y_t - (K_t^{(k+0.5)})\psi_t$ *for* $t = 1, 2, \cdots, T$.

4. *Draw* $(V_{1:T}^{(k+1)}, W_{1:T}^{(k+1)})$ *from* $p(V_{1:T}, W_{1:T}|\gamma_{0:T}, y_{1:T})$.

This algorithm samples $V_{1:T}$ and $W_{1:T}$ jointly in steps 2 and 4. If we instead sample them from each of their full conditionals, we get another variant of this algorithm:

Algorithm 14.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.
2. *Draw* $V_{1:T}^{(k+0.5)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \psi_{0:T}, y_{1:T})$.
3. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+0.5)}, \psi_{0:T}, y_{1:T})$.
4. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \psi_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2\cdots, T$ *with* $\theta_0 = \psi_0$ *and* $\theta_t = y_t - (K_t^{(k+0.5)})\psi_t$ *for* $t = 1, 2, \cdots, T$.
5. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}^{(k+0.5)}, \gamma_{0:T}, y_{1:T})$.
6. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$.

Step 4 can be broken up into two steps: a transformation from $\psi_{0:T}$ to $\theta_{0:T}$, and another transformation from $\theta_{0:T}$ to $\gamma_{0:T}$. This allows us to rewrite algorithm 14 as:

Algorithm 15.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.
2. *Draw* $V_{1:T}^{(k+0.5)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \psi_{0:T}, y_{1:T})$.
3. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+0.5)}, \psi_{0:T}, y_{1:T})$.
4. *Update* $\theta_{0:T}$ *from* $\theta_0 = \psi_0$ *and* $\theta_t = y_t - (K_t^{(k+0.5)})\psi_t$ *for* $t = 1, 2, \cdots, T$.
5. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2\cdots, T$.
6. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|W_{1:T}^{(k+0.5)}, \gamma_{0:T}, y_{1:T})$.
7. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$.

Now the draw of $W_{1:T}^{(k+0.5)}$ in step 3 could actually be drawn conditional on $\theta_{0:T}$ instead of $\psi_{0:T}$ since this does not change the conditional distribution of $W_{1:T}$, so the order of steps 3 and 4 doesn't matter. Similarly in step 6 $V_{1:T}$ could be drawn conditional on $\theta_{0:T}$ isntead of $\gamma_{0:T}$ without change the distribution from which it is drawn, so steps 6 and 5 can be interchanged. This allows us to rewrite algorithm 15 as:

Algorithm 16.

1. *Draw* $\psi_{0:T}$ *from* $p(\psi_{0:T}|V_{1:T}^{(k)}, W_{1:T}^{(k)}, y_{1:T})$.
2. *Draw* $V_{1:T}^{(k+0.5)}$ *from* $p(V_{1:T}|W_{1:T}^{(k)}, \psi_{0:T}, y_{1:T})$.
3. *Update* $\theta_{0:T}$ *from* $\theta_0 = \psi_0$ *and* $\theta_t = y_t - (K_t^{(k+0.5)})\psi_t$ *for* $t = 1, 2, \cdots, T$.
4. *Draw* $V_{1:T}^{(k+1)}$ *from* $p(V_{1:T}|\theta_{0:T}, y_{1:T})$.
5. *Draw* $W_{1:T}^{(k+0.5)}$ *from* $p(W_{1:T}|\theta_{0:T}, y_{1:T})$.
6. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = (L_t^{-1})^{(k+0.5)}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2\cdots, T$.
7. *Draw* $W_{1:T}^{(k+1)}$ *from* $p(W_{1:T}|V_{1:T}^{(k+1)}, \gamma_{0:T}, y_{1:T})$.

which is identical to algorithm 12. So one version of the full CIS algorithm based on $\psi_{0:T}$, $\tilde{\psi}_{0:T}$, $\gamma_{0:T}$, and $\tilde{\gamma}_{0:T}$ is identical to a GIS algorithm. As long as we believe Yu and Meng and don't think the order in which we use each of the DAs in a CIS or GIS algorithm matters much, there doesn't appear to be any benefit to using CIS for DLMs.

References

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

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.

Sylvia Frühwirth-Schnatter and Regina Tüchler. Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing*, 18(1):1–13, 2008.

Sylvia Frühwirth-Schnatter and Helga Wagner. Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. page 165, 2011.

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.

Jan R Magnus. Linear structures. 1988.

Jan R Magnus and H Neudecker. The elimination matrix: some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods*, 1(4):422–449, 1980.

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.

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.

Giovanni Petris, Patrizia Campagnoli, and Sonia Petrone. *Dynamic linear models with R*. Springer, 2009.

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.