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

1 Estimating the Model via Data Augmentation: Parameterization Issues

A well known method to estimate the DLM is via data augmentation (DA) often using forward filtering backward sampling (FFBS), as in 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.

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

The first block runs a simulation smoother which draws the latent states from their conditional posterior distribution given the model parameters. This can be accomplished in a number of ways. FFBS is the original method proposed by Frühwirth-Schnatter [1994] and Carter and Kohn [1994], but alternatives include Koopman [1993], De Jong and Shephard [1995] and McCausland et al. [2011]. The second block draws $\phi = (V, W)$ from their joint conditional posterior which in this model turns out to be independent inverse Wishart distributions. In particular

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

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

The main problem with the state sampler is that in some regions of the parameter space the Markov chain mixes poorly for some of the parameters. For example, in the univariate local level model ($F_t = G_t = 1$ for $t = 1, 2, \cdots, T$) and similar models it is known that if the time constant variance of the latent states, W, is too small, mixing will be poor for W 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. Papaspiliopoulos et al. [2007] for example suggest alternating between the two augmentations within a Gibbs sampler. Some work focuses on using partially noncentered parameterizations that are a sort of bridge between the CP and NCP, e.g. 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.

Another method of combining the two DAs is through what Yu and Meng [2011] call interweaving. The idea is pretty simple: suppose that ϕ denotes the parameter vector, θ denotes one augmented data vector, γ denotes another augmented data vector, and y denotes the data. Then an MCMC algorithm that *interweaves* between θ and γ performs the following steps in a single iteration to obtain the $k+1$'st draw, $\phi^{(k+1)}$, from the k'th draw, $\phi^{(k)}$:

Algorithm 2.

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

Notice that an additional step is added to algorithm 1, and the final step now draws ϕ conditional on γ instead of θ. This is the intuition behind the name "interweaving"—the draw of the second augmented data vector is weaved in between the draws of θ and ϕ. This particular method of interweaving is called a *global* interweaving strategy (GIS) since interweaving occurs globally across the entire parameter vector. It's possible to define a *componentwise* interweaving strategy (CIS) that interweaves within specific steps of a Gibbs sampler as well. Step two of the GIS algorithm is typically accomplished by sampling $\phi|\theta, y$ and then $\gamma|\theta, \phi, y$. In addition, γ and θ are often, but not always, one-to-one transformations of each other conditional on (ϕ, y), i.e. $\gamma = M(\theta; \phi, y)$. Where $M(.; \phi, y)$ is a one-to-one function. In this case, the algorithm becomes:

Algorithm 3.

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

When γ is a one-to-one transformation of θ, step 4 is an update $\gamma = M(\theta; \phi, y)$. The GIS algorithm is directly comparable to the *alternating* algorithm suggested by Papaspiliopoulos et al. [2007]. Given the same two DAs, θ and γ, and parameter vector ϕ, the alternating algorithm for sampling from $p(\phi|y)$ is as follows:

Algorithm 4.

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

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

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 particular, their Theorem 1 suggests that the weaker the dependence between two data augumentations in the posterior, the more efficient the GIS algorithm. In the limit of a posteriori indepedent data augmentations, the GIS algorithm will even obtain iid draws from the posterior of the model parameter. This motivates their focus on ASIS — conditional on the model parameter, a SA and an AA are independent, which suggests that the dependence between the two DAs will be limited in the posterior. In fact, when the prior on ϕ is nice in some sense, Yu and Meng 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 and interweaving generally is a promising approach to improve the speed of MCMC in a variety of models no matter what region of the parameter space the posterior is concentrated.

To gain some intuition about why interweaving works, recall that a typical problem with slow MCMC is that there is high autocorrelation in the Markov chain for ϕ, $\{\phi^{(k)}\}_{k=1}^{K}$, leading to imprecise estimates of $\mathrm{E}[f(\phi)]$ for some function f. Our goal is to reduce this dependence. In the usual DA algorithm, e.g. algorithm 1, when ϕ and θ are highly dependent in the joint posterior, the draws from $p(\theta|\phi, y)$ and then from $p(\phi|\theta, y)$ will hardly move the chain which results in high autocorrelation. Interweaving helps break this autocorrelation in two ways. First, by inserting the extra step, e.g. steps 2 and 3 together in 3, the

chain gets an additional chance to move in a single iteration thereby weakening the autocorrelation. This is a feature of an alternating algorithm as well, but Yu and Meng show that the corresponding interweaving algorithm is often even more efficient. The key is the second point — when the posterior dependence between the two DAs is low, steps 2 and 3 in 3, i.e. step 2 in 2, is enough to almost completely break the dependence in the chain. For the alternating algorithm, it is typically not feasible to find a data augmentation such that step 2 or step 3 of 4 completely breaks the depedence in the chain — this would require finding a DA such that the model parameter and the DA are essentially independent which, in turn, would likely mean that drawing from the conditional posterior of the parameter given the DA is nearly as difficult as drawing from the marginal posterior of the model parameter.

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

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 is in the observation equation so that $\theta_{0:T}$ is not a SA for (V, W) while W is in the system equation so that $\theta_{0:T}$ is not an AA for (V, W) either. In order to find a SA we need to somehow move V from the observation equation (**??**) to the system equation (**??**) while leaving W in the system equation. We also need to find an AA by somehow moving W from the system equation to the observation equation while leaving V in the observation equation. A naive thing to try is to condition on the disturbances instead of the states and see if the disturbances for a SA or an AA for (V, W). The disturbances $w_{0:T}$ are defined by $w_t = \theta_t - G_t\theta_{t-1}$ for $t = 1, , 2, \cdots, T$ and and $w_0 = \theta_0$. However the DA algorithm based on the w_t's is identical to the algorithm based on the θ_t because it turns out that the conditional distributions $p(V, W|\theta_{0:T}, y_{1:T})$ and $p(V, W|w_{0:T}, y_{1:T})$ are identical.

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

To define the scaled disturbances in the general DLM, let L_W denote the Cholesky decomposition of W, i.e. the lower triangle matrix L_W such that $L_W L_W' = W$ Then we'll define the scaled disturbances $\gamma_{0:T}$ by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$. There are actually $p!$ different versions of the scaled disturbances depending on how we order the elements of θ_t, as Meng and Van Dyk [1998] note in a different class of models. We will sidestep the issue of the best ordering of the latent states. No matter which ordering is chosen, we can confirm our intuition that the scaled disturbances are an AA for V and W jointly. The reverse transformation is defined recursively by $\theta_0 = \gamma_0$ and $\theta_t = L_W\gamma_t + G_t\theta_{t-1}$ for $t = 1, 2, \cdots, T$. Then the Jacobian is block lower triangular with the identity matrix and T copies of L_W along the diagonal blocks,

so $|J| = |L_W|^T = |W|^{T/2}$. Then from (**??**) we can write the full joint distribution of $(V, W, \gamma_{0:T}, y_{1:T})$ as

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

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

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

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

The DA algorithm based on $\gamma_{0:T}$ draws $\gamma_{0:T}$ from its conditional posterior and then (V, W) from their joint conditional posterior given $\gamma_{0:T}$. There are a couple methods of performing this draw, including applying one of the simulation smoothers directly to drawing $\gamma_{0:T}$, if possible, or using one of them to draw the latent states $\theta_{0:T}$ before transformting the states to the sclaed disturbances. The draw from the joint conditional posterior of (V, W) is tricky because it is not a known density. We will illustrate how to accomplish it in a worked example in Section 1.4.

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

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

$(V, W, \psi_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \psi_{0:T}, y_{1:T}) \propto & \exp\left[-\frac{1}{2}(\psi_0 - m_0)' C_0^{-1}(\psi_0 - m_0)\right] \\ & \times |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \psi_t' \psi_t\right] \\ & \times |W|^{-(\delta_t + k + T + 2)/2} \exp\left[-\frac{1}{2}\left(tr\left(\Lambda_W W^{-1}\right) + (y_t - \mu_t)'(F_t W F_t')^{-1}(y_t - \mu_t)\right)\right] \end{aligned} \tag{3}$$

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

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

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

The DA algorithm based on $\psi_{0:T}$ is similar to that of $\gamma_{0:T}$ except we note that simiulation smoothing can be accomplished by directly applying the algorithm of McCausland et al. [2011] because the precision matrix of $\psi_{0:T}$ retains the necessary tridiagonal structure. Also we mention in passing that there is a bit of symmetry here — the joint conditional posterior of (V, W) given $\gamma_{0:T}$ is from the same family of densities as that of (W, V) given $\psi_{0:T}$ so that V and W essentially switch places. The upshot is that if we can draw from one we can draw from the other, so this part of our work has been essentially halved.

1.3 The elusive search for a sufficient augmentation

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

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

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

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

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

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

$$\mu = \begin{bmatrix} F_1 G_1 \theta_0 \\ F_2 G_2 G_1 \theta_0 \\ \vdots \\ F_T G_T G_{T-1} \cdots G_1 \theta_0 \end{bmatrix}, \quad \tilde{V}_{k\times k} = \begin{bmatrix} V & 0 & 0 & \ddots & 0 \\ 0 & V & 0 & \ddots & 0 \\ 0 & 0 & V & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \ddots & V \end{bmatrix}, \quad \tilde{W}_{k\times k} = \begin{bmatrix} F_1 R_1 \\ F_2 R_2 \\ \vdots \\ F_T R_T \end{bmatrix} W \begin{bmatrix} R_1' F_1' & R_2' F_2' & \dots & R_T' F_T' \end{bmatrix}.$$

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

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

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

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

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

which implies

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

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

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

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

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

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

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

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

The fundamental problem is that once we find a sufficient augmentation, in order to use it we must obtain draws from a density that appears just as hard to sample from as the posterior density we are already trying to approximate. We did treat θ_0 as a model parameter instead of an element of the data augmentation above, but changing this only makes the resulting conditional posterior of ϕ more complicated. The logic above does not rule out a useful sufficient augmentation, but it does suggest that it will be difficult to find one. We run into a similar problem while trying to find two data augmentations that are independent in the posterior — after making sensible sounding assumptions about the nature of those DAs (i.e. joint with the data they are normally distributed and some dependence assumptions for simplicity), the conditional posterior of ϕ ends of being identical to or just as complicated as the marginal posterior of ϕ.

1.4 The "wrongly scaled" DAs

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

First consider $\tilde{\gamma}_{0:T}$. Notice that for $t = 1, 2, \cdots, T$, $\tilde{\gamma}_t = L_V^{-1} L_W \gamma_t$ while $\tilde{\gamma}_0 = \gamma_0$. The reverse transformation is then $\gamma_t = L_W^{-1} L_V \tilde{\gamma}_t$. The Jacobian is then block diagonal with $L_W^{-1} L_V$ along the diagonal. Thus $|J| = |L_W|^{-T}|L_V|^T = |W|^{-T/2}|V|^{T/2}$. Then from (1) we can write the joint distribution of $(V, W, \tilde{\gamma}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \tilde{\gamma}_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\tilde{\gamma}_0 - m_0)' C_0^{-1}(\tilde{\gamma}_0 - m_0)\right] |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \\ &\times |W|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)' V^{-1} \left(y_t - F_t\theta_t(\tilde{\gamma}_{0:T})\right)\right] \\ &\times |W|^{-(\lambda_W + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_W W^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\gamma}_t' (L_V^{-1} W (L_V^{-1})')^{-1} \tilde{\gamma}_t\right] \end{aligned} \tag{4}$$

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

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

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

Now consider $\tilde{\psi}_t = L_W^{-1} L_V \psi_t$ for $t = 1, 2, \cdots, T$ where again $\tilde{\psi}_0 = \psi_0 = \theta_0$ Then $\psi_t = L_V^{-1} L_W \tilde{\psi}_t$ and the Jacobian is block diagonal with $L_V^{-1} L_W$ along the diagonal. So $|J| = |V|^{-T/2}|W|^{T/2}$ and from (3) we can write the joint distribution of $(V, W, \tilde{\psi}_{0:T}, y_{1:T})$ as

$$\begin{aligned} p(V, W, \tilde{\psi}_{0:T}, y_{1:T}) &\propto \exp\left[-\frac{1}{2}(\tilde{\psi}_0 - m_0)' C_0^{-1}(\tilde{\psi}_0 - m_0)\right] \\ &\times |V|^{-(\lambda_V + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_V V^{-1}\right)\right] \exp\left[-\frac{1}{2}\sum_{t=1}^{T} \tilde{\psi}_t' (L_W^{-1} V (L_W^{-1})')^{-1} \tilde{\psi}_t\right] \\ &\times |W|^{-(\lambda_W + k + 2)/2} \exp\left[-\frac{1}{2} tr\left(\Lambda_W W^{-1}\right)\right] |V|^{-T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T} (y_t - \tilde{\mu}_t)' (F_t W F_t')^{-1} (y_t - \tilde{\mu}_t)\right] \end{aligned} \tag{5}$$

where we define $\tilde{\mu}_1 = L_W \tilde{\psi}_1 - F_1 G_1 \tilde{\psi}_0$ and for $t = 2, 3, \cdots, T$ $\tilde{\mu}_t = L_W \tilde{\psi}_t - F_t G_t F_{t-1}^{-1}(y_{t-1} - L_W \tilde{\psi}_{t-1})$ In terms of $\tilde{\psi}_{0:T}$, the model is then:

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

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

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

2 Interweaving in the DLM: Global and Componentwise

We now have five DAs for the generic DLM with known F_t's and G_t's. For simplicity we'll assume that $dim(y_t) = dim(\theta_t)$ and F_t invertible for $t = 1, 2, \cdots, T$ so that the scaled errors are easy to work with. The five DAs are the states, $\theta_{0:T}$, the scaled disturbances $\gamma_{0:T}$, the scaled errors $\psi_{0:T}$, the wrongly scaled disturbances $\tilde{\gamma}_{0:T}$, and the wrongly scaled errors $\tilde{\psi}_{0:T}$. This allows us to construct several GIS algorithms based on algorithm 3. The main algorithms we consider are the state-dist, state-error, dist-error, and triple interweaving algorithms. The names should be intuitive, but for example the state-dist algorithm interweaves between the states $\theta_{0:T}$ and the scaled disturbances $\gamma_{0:T}$. 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. To illustrate, algorithm 5 is the state-dist GIS algorithm:

Algorithm 5.

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$
2. *Draw* $(V^{(k+0.5)}, W^{(k+0.5)})$ *from* $p(V, W|\theta_{0:T}, y_{1:T})$
3. *Update* $\gamma_{0:T}^{(k+1)}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = (L_W^{(k+0.5)})^{-1}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$
4. *Draw* $(V^{(k+1)}, W^{(k+1)})$ *from* $p(V, W|\gamma_{0:T}, y_{1:T})$

where again $L_W^{(k+0.5)}$ is the Cholesky decomposition of $W^{(k+0.5)}$. In practice we may want to break up step 4 into two steps if it is easier to draw from the full conditionals of V and W rather than drawing them jointly.

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

Algorithm 6.

1. *Draw* $\theta_{0:T}$ *from* $p(\theta_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$
2. *Draw* $V^{(k+1)}$ *from* $p(V|W^{(k)}, \theta_{0:T}, y_{1:T})$
3. *Draw* $W^{(k+0.5)}$ *from* $p(W|V^{(k+1)}, \theta_{0:T}, y_{1:T})$
4. *Update* $\gamma_{1:T}^{(k+1)}$ *from* $\gamma_0 = \theta_0$ *and* $\gamma_t = (L_W^{(k+0.5)})^{-1}(\theta_t - G_t\theta_{t-1})$ *for* $t = 1, 2, \cdots, T$
5. *Draw* $W^{(k+1)}$ *from* $p(W|V^{(k+1)}, \gamma_{0:T}, y_{1:T})$

This algorithm is actually the same as a version of the state-dist interweaving algorithm with some of the steps rearranged, specifically algorithm 5. So it should be similar in performance to a GIS algorithm.

With a little more work, we can also construct a full CIS algorithm that also turns out to be essentially the same as another GIS algorithm. Here we employ the wrongly scaled disturbances and errors. Recall that the wrongly scaled disturbances are $\tilde{\gamma}_0 = \gamma_0 = \theta_0$ and for $t = 1, 2, \cdots, T$, $\tilde{\gamma}_t = L_V^{-1}(\theta_t - G_t\theta_{t-1})$ and the wrongly scaled errors are $\tilde{\psi}_0 = \psi_0 = \theta_0$ and for $t = 1, 2, \cdots, T$, $\tilde{\psi}_t = L_W^{-1}(y_t - F_t\theta_t)$. Now we already now

that $\gamma_{0:T}$ is an AA for $W|V$ and $\tilde{\gamma}_{0:T}$ is a SA for $W|V$, so the two form an AA-SA pair for $W|V$. Similarly, $\psi_{0:T}$ is an AA for $V|W$ while $\tilde{psi}_{0:T}$ is a SA for $V|W$ so together they form an AA-SA pair for $V|W$. Now we can construct a full CIS algorithm:

Algorithm 7.

1. *Draw* $\tilde{\psi}_{0:T}$ *from* $p(\tilde{\psi}_{0:T}|V^{(k)}, W^{(k)}, y_{1:T})$.
2. *Draw* $V^{(k+0.5)}$ *from* $p(V|W^{(k)}, \tilde{\psi}_{0:T}, y_{1:T})$
3. *Update* $\psi_{0:T}$ *from* $\psi_0 = \tilde{\psi}_0$ *and* $\psi_t = (L_V^{(k+0.5)})^{-1} L_W^{(k)} \tilde{\psi}_t$ *for* $t = 1, 2, \cdots, T$.
4. *Draw* $V^{(k+1)}$ *from* $p(V|W^{(k)}, \psi_{0:T}, y_{1:T})$.
5. *Update* $\tilde{\gamma}_{0:T}$ *from* $\psi_{0:T}$, $W^{(k)}$, *and* $V^{(k+1)}$.
6. *Draw* $W^{(k+0.5)}$ *from* $p(W|V^{(k+1)}, \tilde{\gamma}_{0:T}, y_{1:T})$
7. *Update* $\gamma_{0:T}$ *from* $\gamma_0 = \tilde{\gamma}_0$ *and* $\gamma_t = (L_W^{(k+0.5)})^{-1} L_V^{(k+1)} \tilde{\gamma}_t$ *for* $t = 1, 2, \cdots, T$.
8. *Draw* $W^{(k+1)}$ *from* $p(W|V^{(k+1)}, \gamma_{0:T}, y_{1:T})$

Steps 1-4 constitute a Gibbs step for V and steps 5-8 constitute a Gibbs step for W. It turns out that $p(W|V, \tilde{\gamma}_{0:T}, y_{1:T})$ and $p(W|V, \theta_{0:T}, y_{1:T})$ are the same density, and also that $p(V|W, \tilde{\psi}_{0:T}, y_{1:T})$ and $p(V|W, \theta_{0:T}, y_{1:T})$ are the same density. The upshot is that step 1 of algorithm 7 can be replaced with a draw from $p(\theta_{0:T}|V, W, y_{1:T})$, and any time we condition on one of the "wrongly scaled" variables, we can condition on $\theta_{0:T}$ instead. It can be shown that this algorithm and the dist-error algorithm can each have their steps rearranged so that the two algorithms are the same. This suggests that we should expect the two algorithms to perform similarly.

References

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

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

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

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

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

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

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

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

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

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

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

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

Yaming Yu and Xiao-Li Meng. To center or not to center: That is not the question - an ancillarity–sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20(3):531–570, 2011.