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

1 Introduction

Ever since the seminal article by **?**, data augmentation has been a common strategy used to construct MCMC algorithms in order to approximate probability distributions. Suppose $p(\phi|y)$ is the target density, in this case the posterior distribution of some parameter ϕ given data y. We use $p(.)$ to denote the probability density of the enclosed random variables. Then the data augmentation (DA) algorithm adds an augmented data vector θ with joint distribution $p(\phi, \theta|y)$ such that $\int_\Theta p(\phi, \theta|y)d\theta = p(\phi|u)$. Then the DA algorithm is a Gibbs sampler which constructs a Markov chain for ϕ that obtains the $k+1$'st state of ϕ from the k'th state as follows (we implicitly condition on the data y in all algorithms and only superscript the previous and new draws of the model parameters — intermediate draws of a DA or the model parameter are not superscripted since they are not part of the Markov chain for ϕ):

Algorithm 1 (Data Augmentation)**.**

$$[\theta|\phi^{(k)}] \to [\phi^{(k+1)}|\theta]$$

Here $[\theta|\phi^{(k)}]$ means a draw of θ from $p(\theta|\phi^{(k)}, y)$ and $[\phi^{(k+1)}|\theta]$ means a draw from $p(\phi^{(k+1)}|\theta, y)$ — implicitly we condition on the data unless otherwise noted. The augmented data vector, θ, need not be interesting in any scientific sense — it can be viewed purely as a computational construct. But for cases where the natural DA is intrinsically interesting, the DA algorithm does incidentally obtain joint draws from $p(\phi, \theta|y)$ but we will view θ as a nuisance parameter.

The EM algorithm of **?** and its variants are closely analogous to DA algorithms — the DA algorithm can be viewed as a stochastic version of the EM algorithm. In fact there is a long history of using methods typically used to speed up one algorithm to speed up the other; **?** shows how much overlap in the two literatures exists. The main advantage of DA and EM algorithms is their ease of implementation, but much of this work is necessary because DA and EM algorithms can often be prohibitively slow. Most of this work has focused on multilevel models — e.g. **?** and **?**, but relatively little attention has been paid to time series models despite strong similarities between some time series models and the hierarchical models typically studied. We seek to improve DA schemes in a particular class of models — linear, Gaussian statespace models, a.k.a. dynamic linear models (DLMs).

One particularly recent advance in the DA literature, from **?**, is the notion of interweaving two separate DAs together. Suppose that we have a second agumented data vector γ with a full joint distribution $p(\phi, \theta, \gamma|y)$ such that $\int_\Theta \int_\Gamma p(\phi, \theta, \gamma|y)d\gamma d\theta = p(\phi|y)$. Then a GIS or general interweaving strategy obtains $\phi^{(k+1)}$ from $\phi^{(k)}$ as follows:

Algorithm 2. *GIS*

$$[\theta|\phi^{(k)}] \to [\gamma|\theta] \to [\phi^{(k+1)}|\gamma].$$

? show that when θ is a sufficient augmentation (SA, a.k.a. centered augmentation), i.e. $p(y|\theta, \phi) = p(y|\theta)$, and γ is an ancillary augmentation (AA, a.k.a. non-centered augmentation), i.e. $p(\gamma|\phi) = p(\gamma)$, then under some weak conditions this "ancillary-sufficient" interweaving strategy, or ASIS, is equivalent to the the optimal PX-DA algorithm, e.g. in **?**, **?**, **?** and **?**. Our main purpose is to apply this idea to a particular class of DLMs.

The generic DLM can be defined as follows:

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

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

for $t = 1, 2, \cdots, T$, and $v_{1:T}$, $w_{1:T}$ independent. Equation (1) is called the *observation equation* and equation (2) is called the *system equation*. Similarly, $v_{1:T}$ are called the *observation errors*, $V_{1:T}$ are called the

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

The subclass of DLMs we will focus on forces the variances to be time independent and treats $F_{1:T}$ and $G_{1:T}$ as known. Typically additional model structure is used to learn about $V_{1:T}$ and $W_{1:T}$ if time dependence is enforced anyway – e.g. a stochastic volatility prior which would require a statespace model describing the $V_{1:T}$'s and $W_{1:T}$'s as data. However, many of our results may be useful in more complicated time-varying variance models. So we set $V_t = V$ and $W_t = W$ for $t = 1, 2, \cdots, T$. Forcing $F_{1:T}$ and $G_{1:T}$ to be constant on the one hand is not a big constraint since relaxing it will simply add one or more Gibbs steps to the algorithms we explore so long as no parameter that enters any F_t or G_t also enters V or W. On the other hand, the behavior of these algorithms will depend on the precise structure of the unknown components of $F_{1:T}$ and $G_{1:T}$ and in one of the data augmentations that we discuss there is a bit more housekeeping associated with $F_{1:T}$ depending on an unknown parameter (Section **??**). In any case, when $\phi = (V, W)$ is our unkown parameter vector and we can write the model as

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

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

We show that in this particular class of DLMs, the usual DA vector, $\theta_{0:T}$,, is neither a SA nor an AA. Furthermore, we show that given some reasonable assumptions about the any SA, generically θ, drawing from $p(\theta|\phi, y)$ is just as difficult as drawing from the target posterior distribution $p(\phi|y)$, which defeats the purpose of looking for a SA to begin with. We do, however, find two separate AAs — the scaled disturbances, $\gamma_{0:T}$, defined by $\gamma_0 = \theta = 0$ and $\gamma_t = L_W^{-1}w_t$ for $t = 1, 2, \cdots, T$ where L_W is the Cholesky decomposition of W, and the scaled errors, $\psi_{0:T}$, defined by $\psi_0 = \theta_0$ and $\psi_t = L_V^{-1}v_t$ for $t = 1, 2, \cdots T$ where L_V is the Cholesky decomposition of V. The former has been used in both the multilevel models and time series literature and is essentially the standard non-centered augmentation, but the latter is novel.

Furthermore, we employ the componentwise interweaving strategy (CIS) of **?**. A CIS algorithm for $\phi = (\phi_1, \phi_2)$ essentially employs interweaving for each block of ϕ spearately, e.g.

Algorithm 3. *CIS*

$$\begin{array}{lllll} [\theta_1|\phi_1^{(k)}, \phi_2^{(k)}] & \rightarrow & [\gamma_1|\phi_2^{(k)}, \theta_1] & \rightarrow & [\phi_1^{(k+1)}|\phi_2^{(k)}, \gamma_1] \quad \rightarrow \\ [\theta_2|\phi_1^{(k+1)}, \phi_2^{(k)}, \gamma_1] & \rightarrow & [\gamma_2|\phi_1^{(k+1)}, \theta_2] & \rightarrow & [\phi_2^{(k+1)}|\phi_1^{(k+1)}, \gamma_2] \end{array}$$

where θ_i and γ_i are distinct data augmentations for $i = 1, 2$, but potentially $\gamma_1 = \theta_i$ for $i = 1, 2$ or $\gamma_2 = \theta_i$ for $i = 1, 2$. The first line draws ϕ_1 conditional on ϕ_2 using interweaving in a Gibbs step, while the second line does the same for ϕ_2 condtional on ϕ_1. The algorithm can easily be extended to greater than two blocks within ϕ. The advantage with CIS is that it is often easier to find an AA–SA pair of DAs for ϕ_1 conditional on ϕ_2 and another pair for ϕ_2 conditional on ϕ_1 than for $\phi = (\phi_1, \phi_2)$ jointly. We construct a CIS algorithm for the subclass of DLMs we consider, updating V and W in separate blocks, based on the "wrongly scaled" versions of the scaled errors and scaled disturbances, i.e. we create an AA-SA pair for W using $\gamma_{0:T}$ and $\tilde{\gamma}_{0:T}$ where $\tilde{\gamma}_0 = \theta_0$ and $\tilde{\gamma}_t = L_V^{-1}w_t$ for $t = 1, 2, \cdots, T$ and similarly for V using $\psi_{0:T}$ and $\tilde{\psi}_{0:T}$, analogously defined. Further, we that $\tilde{\gamma}_{0:T}$ and $\tilde{\psi}_{0:T}$ can both be replaced with $\theta_{0:T}$ without changing the algorithm, despite the fact that $\theta_{0:T}$ is a SA for W conditional on V but not for V conditional on W. Even further, we show that this algorithm is the same as a GIS algorithm that interweaves between $\gamma_{0:T}$ and $\psi_{0:T}$, except with the steps arranged in a different order.

In the context of a particular DLM, the local level model with univariate y_t, univariate θ_t, and $F_t = G_t = 1$, we conduct a simulation study in order to explore the properties of the various MCMC algorithms derived for the general case. In the process we find that we have to give up drawing V and W jointly when conditioning on any DA other than the states, and in doing so we draw from two classes of densities closely related to the generalized inverse Gaussian distribution (see e.g. **?**). These densities take the form

$$p(x) \propto x^{-\alpha-1} \exp\left[-ax + bx^{1/2} - cx^{-1}\right]$$

and

$$p(x) \propto x^{-\alpha-1} \exp\left[-ax + bx^{-1/2} - cx^{-1}\right]$$

Both densities contain the generalized inverse Gaussian as a special case when $b = 0$ and in our problem both are often log-concave.

In our simulations, we find that the true signal-to-noise ratio, $R = W/V$, determines the behavior of all of the standard DA algorithms based on the various DAs we construct, and the behavior of the GIS algorithms can be traced to the behavior fo the base DA algorithms for the DAs used in the GIS algorithm. In particular we find that the scaled disturbance DA algorithm works best for $R < 1$ while the scaled error DA algorithm works best for $R > 1$, while GIS algorithm that interweaves between the scaled disturbances and the scaled errors works well as long as R is not too close to one, though for larger sample sizes all algorithms have problems in increasingly large regions of the parameter space.

The rest of the paper is organized as follows. In Section **??** we discuss data augmentation methods, introduce several new data augmentations for this class of DLMs and show that it is unlikely that a useful sufficient augmentation (centered augmentation) exists when we consider all model parameters as unknown. Section **??** contains several interweaving algorithms based on the various data augmentations we discuss in Section **??** along with some results about when various CIS and GIS algorithms are equivalent. In Section **??** we work out an example with the local level model and report the results of fitting the model to different simulated datasets using several of the MCMC samplers we construct. Finally Section **??** discusses our results and concludes.