Ancillarity-Sufficiency or not; Interweaving to improve MCMC estimation of the DLM

Matthew Simpson[1], Jarad Niemi[2], Vivekananda Roy[2]

May 2, 2014

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1] Iowa State University, Depts. of Statistics and Economics, Ames, IA, USA
simpsonm@iastate.edu

[2] Iowa State University, Dept. of Statistics, Ames, IA, USA
<niemi,vroy>@iastate.edu

Abstract

In dynamic linear models (DLMs), MCMC sampling can often be very slow for estimating the posterior density — especially for longer time series. In particular, in some regions of the parameter space the standard data augmentation algorithm can mix very slowly. Using some of the insights from the data augmentation for multilevel models literature, we explore several alternative data augmentations for a general class of DLMs and we show that no "practical" sufficient augmentation exists. In addition, we utilize these augmentations to construct several interweaving algorithms — though we cannot construct an ancillary-sufficient interweaving algorithm (ASIS) since no sufficient augmentation exists, we find two ancillary augmentation and are able to construct a componentwise interweaving algorithm that uses ASIS for each model parameter conditional on the rest. Using the local level DLM, we show how to construct several of these algorithms and conduct a simulation study in order to discern their properties. We find that several algorithms that outperform the usual "state sampler" for many values of the population parameters, though there is room for improvement because one or more steps of the more efficient algorithms involves an often inefficient rejection sampling draw from a class of density that contains the generalized inverse Gaussian as a special case.

Keywords: data augmentation; time series; interweaving

1 Introduction

In dynamic linear models (DLMs), MCMC sampling can often be very slow for estimating the posterior density — especially for longer time series. In particular, in some regions of the parameter space the standard data augmentation algorithm can mix very slowly. Using some of the insights from the data augmentation for multilevel models literature, we explore several alternative data augmentations for a general class of DLMs. The generic DLM can be defined as follows:

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

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

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 is called the *observation* *variance*, $w_{1:T}$ are called the *system disturbances* and W is called the *system variance*. 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, and W are all functions of ϕ, but we will assume $\phi = (V, W)$. To complete the model specification in a Bayesian context, we need priors on θ_0, V, and W. We'll use the standard approach and assume that they are mutually independent a priori and that $\theta_0 \sim N(m_0, C_0)$, $V \sim IW(\Lambda_V, \lambda_V)$ and $W \sim IW(\Lambda_W, \lambda_W)$ where m_0, C_0, Λ_V, λ_V, Λ_W, and λ_W are known hyperparameters and $IW(\Lambda, \lambda)$ denotes the inverse Wishart distribution with degrees of freedom λ and positive definite scale matrix Λ.

2 Results for the General DLM

The standard method of estimating this model is via data augmentation (DA), 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, $(\phi^{(k+1)}, theta^{(k+1)})$, from the k'th state, $\phi^{(k)}$ as follows:

Algorithm 1. *DA algorithm.*

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

We construct two more data augmentations, the scaled disturbances, given by $\gamma_0 = \theta_0$ and $\gamma_t = L_W^{-1}(\theta_t - G_t\theta_{t-1})$ for $t = 1, 2, \cdots, T$ where L_W is the Cholesky factor of W, and the scaled errors, given by $\psi_0 = \theta_0$ and $\psi_t = L_V^{-1}(y_t - F_t\theta_t)$ for $t = 1, 2, \cdots, T$ where L_V is the Cholesky factor of V. The scaled disturbances are well known in the time series and multilevel models literature, e.g. Frühwirth-Schnatter [2004], Papaspiliopoulos et al. [2007], and Van Dyk and Meng [2001], as the "non-centered augmentation" but the scaled errors are novel. Both DAs are ancillary augmentations (AAs) — generic DA θ is an AA if $p(\theta|\phi) = p(\phi)$ where ϕ is the model parameter. If we can find a sufficient augmentation (SA), i.e. a DA θ such that $p(y|\theta, \phi) = p(y|\theta)$ where y is the data, then we can construct an ancillary–sufficient interweaving algorithm (ASIS) from Yu and Meng [2011]. We show that any SA would be impractical to use, but nevertheless construct several interweaving algorithms including GIS and CIS algorithms. A GIS algorithm, or general interweaving strategy, based on two DAs θ and γ has the form

Algorithm 2. *GIS Algorithm.*

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

while a CIS algorithm, or componentwise interweaving strategy, essentially runs a GIS step for ϕ_1 given ϕ_2 and a separate GIS step for $\phi_2|\phi_1$ where $\phi = (\phi_1, \phi_2)$. Much like the Gibbs sampler, this can be extended to multiple GIS steps and some of them may even be standard Gibbs steps. In particular, each GIS step can be an ASIS step when the two DAs used in that step form an AA-SA pair for $\phi_i|\phi_{-i}$. We show that the following CIS algorithm for the DLM is equivalent to a CIS algorithm with ASIS steps:

Algorithm 3. *Full CIS for DLMs, based on states.*

$$\begin{array}{lllllllll}
[\theta_{0:T}|V^{(k)}, W^{(k)}] & \to & [V|W^{(k)}, \theta_{0:T}] & \to & [\psi_{0:T}|V, W^{(k)}, \theta_{0:T}] & \to & [V^{(k+1)}|W^{(k)}, \psi_{0:T}] & \to \\
[\theta_{0:T}|V^{(k+1)}, W^{(k)}, \psi_{0:T}] & \to & [W|V^{(k+1)}, \theta_{0:T}] & \to & [\gamma_{0:T}|V^{(k+1)}, W, \theta_{0:T}] & \to & [W^{(k+1)}|V^{(k+1)}, \gamma_{0:T}]. &
\end{array}$$

Furthermore, we show that this algorithm is the same as the following Dist-Error GIS algorithm with some steps rearranged and V and W drawn separately instead of jointly:

Algorithm 4. *Dist-Error GIS for DLMs.*

Parameter	State	Dist	Error	State-Dist	State-Error	Dist-Error	Full CIS
V	$R < 1$	$R < 1$	$R > 1$	$R < 1$	$R \not\approx 1$	$R \not\approx 1$	$R \not\approx 1$
W	$R > 1$	$R < 1$	$R > 1$	$R \not\approx 1$	$R > 1$	$R \not\approx 1$	$R \not\approx 1$

Table 1: Rule of thumb for when each algorithm has a high effective sample size for each variable as a function of the true signal-to-noise ratio, $R = W/V$.

$$\begin{array}{lclclcl} [\gamma_{0:T}|V^{(k)}, W^{(k)}] & \rightarrow & [V|W^{(k)}, \gamma_{0:T}] & \rightarrow & [W|V, \gamma_{0:T}] & & \rightarrow \\ [\psi_{0:T}|V, W, \gamma_{0:T}] & \rightarrow & [V^{(k+1)}|W, \psi_{0:T}] & \rightarrow & [W^{(k+1)}|V^{(k+1)}, \psi_{0:T}]. & & \end{array}$$

3 Simulation Study in the Local Level Model

We apply these algorithms in a worked example using the local level model where $G_t = F_t = 1$ for $t = 1, 2, \cdots, T$ with one difference — V and W are sampled separately instead of jointly when conditioning on the scaled disturbances or the scaled errors. In doing so, when we draw $W|V, \gamma_{0:T}$ or $V|W, \psi_{0:T}$, we draw from the following density

$$p(x) \propto x^{-\alpha-1} \exp\left[-ax + b\sqrt{x} - \beta/x\right]$$

where $\alpha, a, \beta > 0$ and $b \in \Re$. This density contains the generalized inverse Gaussian as a special case when $b = 0$, but is difficult to sample from efficiently, which hurts any algorithm based on the "scaled" DAs. We use adaptive rejection sampling (Gilks and Wild [1992]) when possible, but otherwise use rejection sampler with a Cauchy proposal for $\log(x)$.

Using this worked example, we simulated a fake dataset from the local level model for various choices of V, W, and T. We created a grid over V–W space with (V, W) ranging from $(10^{-2}, 10^{-2})$ to $(10^2, 10^2)$ and we simulated a dataset for all possible combinations of V and W with each of $T = 10, 100, 1000$. Then for each dataset, we fit the local level model using each DA algorithm, each GIS algorithm based on any two of the DAs, and the CIS algorithm. We used the same rule for constructing priors for each model: $\theta_0 \sim N(0, 10^7)$, $V \sim IG(5, 4V^*)$, and $W \sim IG(5, 4W^*)$, mutually independent where (V^*, W^*) are the true values of V and W used to simulate the time series. So the prior mean is equal to the true values of V and W so that both the prior and likelihood and thus the posterior roughly agree about the likely values of V and W. For each dataset and each sampler we obtained $n = 3000$ draws and threw away the first 500 as burn in. The chains were started at the true values used to simulated the time series, so we can examine the behavior of the chains to determine how well they mix but not how quickly they converge. We look at the effective sample size (ESS) (see e.g. Gelman et al. [2003]) of each component component in order to assess the MCMC efficiency of each sampler.

Table 1 summarizes the results for each MCMC sampler. We find that the state sampler, the DA algorithm based on the states, has a high ESS for V when the population signal-to-noise ratio $R = W^*/V^*$ is less than one, and a high ESS for W when R is greater than one. The scaled disturbance sampler has a high ESS for both V and W when $R < 1$ while the scaled error sampler has a high ESS for both V and W when $R > 1$. The GIS algorithms, i.e. State-Dist, State-Error, and Dist-Error, all have high ESS for either V or W when at least one of the two DA algorithms it is based on has a high ESS for that parameter, e.g. the Dist-Error algorithm has high ESS for both V and W as long as R is not too close to one. The CIS algorithm behaves essentially identically to the Dist-Error algorithm. One major caveat to this table is that as T, the length of the time series, increases, all ESS's decrease so that e.g. for the Dist-Error sampler to have high ESS's for V and W, R must be farther and farther from one.

Figure 1 contains a plot of the log time per 1000 effective draws, i.e.

$$\log \frac{\text{time in minutes}}{\text{effective sample size}}.$$

When computational time is taken into account, the Dist-Error GIS and Full CIS algorithms come out on

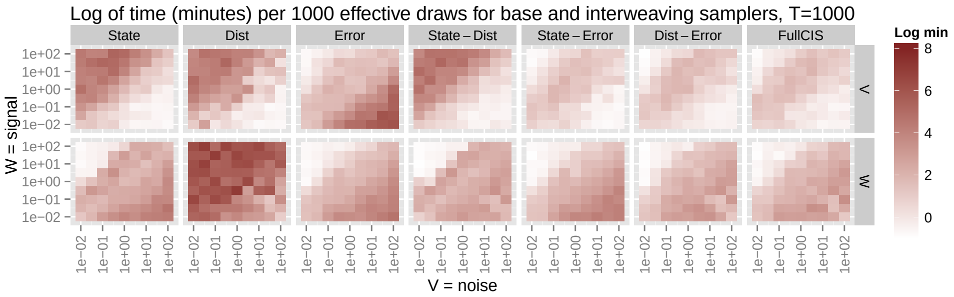

Figure 1: Log of the time in minutes per 1000 effective draws in the posterior sampler for V and W, and for a time series of length $T = 1000$ in the state, scaled disturbance and scaled error samplers and for all four interweaving samplers. Horizontal and vertical axes indicate the true values of V and W respectively for the simulated data. The signal-to-noise ratio is constant moving up any diagonal. In the upper left the signal is high, in the lower right the noise is high.

top, despite both algorithms having to inefficiently draw from the density mentioned above. With a better method of drawing from these densities, these two samplers will improve even more.

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.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis. 2003.

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

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.