Ancillarity–Sufficiency or not

Interweaving to improve MCMC estimation of DLMs

Matthew Simpson
Iowa State University, Departments of Statistics and Economics
themattsimpson@gmail.com

PLACEHOLDER LOGO

Rewrite as it seems like interweaving has already been used to improve DA chains for DLMs.

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. Recently ancillarity-sufficiency interweaving has been introduced as a method to take advantage of alternate parameterizations in multilevel models in order to improve the mixing and convergence properties of the chain. Focusing on the local level DLM, we explore alternate parameterizations and various interweaving algorithms through simulation in order to improve mixing.

do what.

which chain?

Model and priors

The local level model is a univariate DLM with data y_t for $t = 1, 2, \cdots, T$ and univariate latent state θ_t for $t = 0, 2, \cdots, T$ that satisfies

$$y_t|\theta_{0:T} \overset{ind}{\sim} N(\theta_t, V)$$
$$\theta_t|\theta_{0:t-1} \sim N(\theta_{t-1}, W).$$

The unknown parameter vector is $\phi = (V, W)$. For priors, we'll take $\theta_0 \sim N(m_0, C_0)$, $V \sim IG(\alpha_V, \beta_V)$, and $W \sim IG(\alpha_W, \beta_W)$ where (θ_0, V, W) are mutually dependent, and m_0, $C_0 > 0$, $\alpha_V > 0$, $\beta_V > 0$, $\alpha_W > 0$, and $\beta_W > 0$ are some known constants.

The standard MCMC approach and its limitations

The standard MCMC sampling algorithm for the local level model and DLMs generally is a data augmentation (DA) algorithm that uses $\theta_{0:T}$ as an augmented data vector and forward filtering backward sampling (FFBS) in order to draw from $p(\theta_{0:T}|V, W, y_{1:T})$, see Frühwirth-Schnatter [1994] and Carter and Kohn [1994]. Call this algorithm the *state sampler*. One problem with this algorithm is that in some regions of the parameter space the Markov chain mixes poorly for some of the parameters. A well known method of improving mixing and convergence in MCMC samplers is reparameterizing the model, e.g. Papaspiliopoulos et al. [2007]. A recent approach that builds on reparameterizing is the interweaving strategy of Yu and Meng [2011].

Interweaving — GIS, ASIS, and CIS

Interweaving samplers are a class of algorithms for MCMC by Yu and Meng that rely on using multiple DAs within one iteration. In the general case, let y denote the data vector, ϕ the parameter vector, θ a DA, and γ another DA. Global interweaving strategy (GIS) samplers obtain $\phi^{(k+1)}$ from $\phi^{(k)}$ by:

$$[\theta|\phi^{(k)}, y] \to [\phi^{(k+0.5)}|\theta, y] \to [\gamma|\theta, \phi^{(k+0.5)}, y] \to [\phi^{(k+1)}|\gamma, y]$$

When γ is a one-to-one transformation of θ, step 3 is an update using that transformation. The GIS algorithm is directly comparable to an alternating algorithm that draws γ from $p(\gamma|\phi, y)$ in step 3 but is otherwise identical. If necessary, each of these steps can be broken down into separate Gibbs steps.

An ancillary sufficient interweaving strategy (ASIS) uses, in Yu and Meng's words, a sufficient augmentation (SA) and an ancillary augmentation (AA). The DA θ is a SA for ϕ if $p(y|\theta, \phi) = p(y|\theta)$, while θ is an AA for ϕ if $p(\theta|\phi) = p(\theta)$. The advantage of ASIS over a generic GIS is that the SA and the AA are typically a "beauty and the beast" pair – in all regions of the parameter space, the DA algorithm based on either the SA or the AA will have good mixing and convergence properties, allowing the ASIS algorithm to take advantage and have good properties.

This line is too long

Sometimes it's hard to find a SA or an AA or a given model. In that case, a componentwise inteweaving strategy (CIS) may be superior. Suppose $\phi = (\phi_1, \phi_2)$. A CIS algorithm uses a GIS algorithm in a Gibbs step for ϕ_1, and another GIS algorithm in a Gibbs step for ϕ_2, potentially using a different pair of DAs. In this setting, componentwise ASIS is possible — we need a SA–AA pair of DAs for ϕ_1, and another pair for ϕ_2.

which setting?

New DAs and algorithms for the local level model

conditional on w?

The states, $\theta_{0:T}$, are a SA for V and an AA for W. We construct two more DAs for the local level model. Define the scaled disturbances as $\gamma_0 = \theta_0$ and $\gamma_t = (\theta_t - \theta_{t-1})/\sqrt{W}$ for $t = 1, 2, \cdots, T$. The *scaled disturbance sampler* is the DA algorithm based on the scaled disturbances, except it draws V and W in separate Gibbs steps instead of jointly like in the state sampler. It turns out that $\gamma_{0:T}$ is an AA for (V, W). Frühwirth-Schnatter uses the analogue of this augmentation in the context of a dynamic regression model.

2004?

Similarly, define the scaled errors as $\psi_0 = \theta_0$ and $\psi_t = (y_t - \theta_t)/\sqrt{V}$ for $t = 1, 2, \cdots, T$. The *scaled error sampler* is the DA algorithm based on the scaled errors. The scaled error sampler also draws V and W in separate Gibbs steps and, like the scaled disturbances, it turns out that the scaled errors are also an AA for (V, W). We aren't able to find a SA for (V, W) however, so as far as we can tell there is no ASIS algorithm available.

This gives four separate GIS algorithms — the *state – disturbance GIS algorithm* based on the states and the scaled disturbances, the *state – error GIS algorithm* based on the states and the scaled errors, the *disturbance – error GIS algorithm*, based on the scaled disturbances and the scaled errors, and the *triple GIS algorithm* based on all three augmented data vectors. In all four algorithms, we use the various agumented data vectors in the order $\theta_{0:T}$, $\gamma_{0:T}$, then $\psi_{0:T}$. Other orders could change the properties of the algorithms, but Yu and Meng report that this choice doesn't seem to matter much in general.

(2011)

We also construct a full CIS algorithm that uses $\theta_{0:T}$ and $\gamma_{0:T}$ in the Gibbs step for W while using $\theta_{0:T}$ and $\psi_{0:T}$ in the Gibbs step for V. This full CIS algorithm can be shown to be identical to a componentwise ASIS algorithm and is also the same as the disturbance – error GIS algorithm except with the order of some of the steps changed.

Simulation setup

In order to investigate these algorithms, we simulated time series with length $T = 10, 100, 1000$ over a grid of V–W space. Then for each dataset, we fit the local level model using each algorithm mentioned above. We used the same priors for each dataset: $\theta_0 \sim N(0, 10^7)$, $V \sim IG(5, 4\tilde{V})$, and $W \sim IG(5, 4\tilde{W})$, mutually independent where $(\tilde{V}, \tilde{W})$ are the true values of V and W used to simulate the time series.

For each model and each sampler, we obtained 2500 draws after a burn in of 500 with each chain started at the true values used to simulate the time series. Define the effective sample proportion (ESP) for a scalar component of the chain as the effective sample size (ESS) of the component divided by the actual sample size, i.e. $ESP = ESS/n$. An $ESP = 1$ indicates that the Markov chain is behaving as if it obtains iid draws from the posterior.

Results

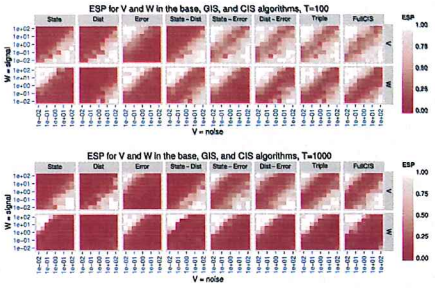

Figure 1

Figure 1 contains plots of ESP for V and W in each chain of each of the samplers for $T = 100$, and $T = 1000$. We'll focus on $T = 100$ first. The state sampler has a low ESP for V and a high ESP for W when the signal-to-noise ratio, W/V, is larger than one and a low ESP for W and a high ESP for V when W/V is less than one. The particular values of V and W don't seem to matter at all — just their relative values. Moving up any diagonal on the plots for V and W in the state sampler, W/V is constant and the ESS appears roughly constant. The basic lesson here is that the state sampler has mixing issues for whichever of V or W is smaller.

The figure tells a different story for the scaled disturbance sampler. When W/V is less than one, ESPs for both V and W are nearly one while when W/V is greater than one ESPs for both V and W become small. The scaled error sampler has essentially the opposite properties. When W/V is large, it has ESP near one for both V and W. On the other hand, when W/V is small is has a low ESP for both V and W. None of these conclusions change as T increases, though this exacerbates all mixing problems.

This suggests that the disturbance–error GIS will be better than any of the other two DA GIS algorithms because it has the "beauty and the beast" property over a wide ranger of the parameter space. This is true and in fact, the triple GIS and full CIS algorithms don't seem to do any better in any region of the parameter space. For all three of these samplers, ESPs for both V and W are high when W/V is farther away from one, but lower when W/V is near one.

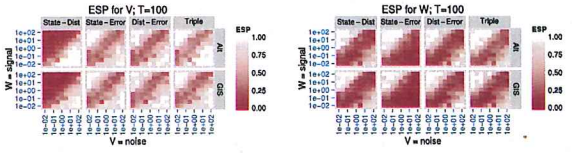

Figure 2

Figure 2 shows the same plots for the full set of GIS algorithms and for their corresponding alternating algorithms, but only for $T = 100$ in order to illustrate that the interweaving algorithm doesn't improve mixing over the corresponding alternating algorithm. However, it is a bit less computationally costly since the third step is a transformation instead of a draw from the joint distribution of a full augmented data vector.

only a bit?

Implications for general DLMs

The scaled disturbances always exist as a possible DA in any DLM as a one-to-one transformation of the states. The scaled errors, on the other hand, are only a one-to-one transformation of the states in some special models. In other models they can be defined, but determining the relevant conditional distributions may be more difficult. So the disturbance – error GIS algorithm can be defined, but it is not entirely clear at this point what quantity will play the role of the signal-to-noise ratio when V and W are (possibly time dependent) covariance matricies.

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.

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

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.

Acknowledgements

Jarad Niemi, Vivekananda Roy

why are you doing burnin when the chains are started at the true value?