Referee's report on

Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

Submitted to JASA, manuscript #JASA-T&M-2017-0147

This paper studies data augmentation algorithms for MCMC. The authors motivate their contribution by noting that standard data augmentation Gibbs samplers deteriorate in effiency as the data/sample size (n) grows, and that this poor mixing can be explained by a lack of control over Gibbs step sizes: as the sample size grows, Gibbs steps converge to zero more quickly than the width of the high-probability region of the posterior (i.e., they become too small, in a relative sense). To address this problem, the paper develops a new family of data-augmentation–type proposal distributions, indexed by location parameters $(b_i)_{i=1}^n$ and scale parameters $(r_i)_{i=1}^n$. These proposal distributions are used in a Metropolis scheme to sample the original posterior. The resulting algorithm is thus no longer a Gibbs sampler, but it is more flexible.

The paper discusses methods for tuning (r_i) and (b_i) and presents specific instantiations of the algorithm for probit regression, logistic regression, a hierarchical binomial model, and a Poisson log-normal model. (Tuning is adaptive but finite in length, thus sidestepping convergence questions associated with adaptive MCMC.) Strong empirical performance is demonstrated in all of the examples.

Overall I think this is an interesting approach, offering impressive gains in computational efficiency for an important and widely used class of Bayesian models. The paper is well written and the results are convincing. But I have several questions on the tuning of the algorithm and the broader rationale for using this approach (given that it already departs from Gibbs). Specific comments are as follows:

1. Since the algorithm already involves Metropolization and requires the tuning of up to $2n$ proposal parameters, a natural question is how it would compare to the following (naïve!) approach: compute a standard Gibbs proposal for θ (essentially using $Q_{r,b}$ in (3) with $r = 1$ and $b = 0$) and then simply *scale* this proposal. The scaling factor would certainly depend on n, and perhaps be as fine-grained as one scaling factor per component of θ (i.e, p distinct factors). The approach proposed in the paper is more structured, but how does it differ in spirit or effect from this simple alternative? In other words, if you're already resorting to regular M-H and a large number of tuning parameters, what is the "special" advantage of using the proposed modified Gibbs update to generate the proposal?

2. The tuning of r and b seems to have two parts. First, some questions about r tuning:

   - Is there ever a situation where r cannot be chosen to set the $\Delta_{\mathcal{I}}$ factors (page 14, lines 34–36) to zero? In the examples this seemed always to be feasible.

- If Δ can always be brought to zero, then the choice of norm in its definition does not much matter. But if it is nonzero then we should care about *how* Δ is small. To that end, the Frobenius norm does not seem particularly satisfactory in the context of symmetric positive definite (SPD) matrices. Why not use instead Rao's distance between Gaussian distributions of the same mean, which corresponds to the natural metric on the cone of SPD matrices? Among other advantages, the distance between two SPD matrices A and B would then be the same as the distance between their inverses, A^{-1} and B^{-1}.

With regard to b tuning: the choice proposed at the top of page 15 seems particularly local and thus subject to fluctuation, since it depends only on the current state of the chain θ_t. What happens when you then stop adaptation after a fixed number of steps?

It would be interesting to see trace plots of r and b adjustments during the tuning phase.

3. It's slightly unsatisfying that some ad hoc tuning of the variance parameters (page 15, line 21) must occasionally be done. Of course, it's also entirely understandable, but can you rationalize the approach or suggest a different alternative? Perhaps an empirical posterior covariance rather than Fisher information at the mode?

4. I like the analytical results on how r should scale with n in the simpler cases (e.g., probit). In the more complex cases where r and b are tuned using the Fisher information, some post-tuning scaling analysis could be illuminating. Empirically, how do the r values scale with n? Does b have any relationship to n? This would suggest repeating some of the examples for different n and identifying trends. But this exercise could yield rules of thumb that may be useful in practice.

5. Editorial comments:

   - Even though this is common jargon, it would be good to define n and p before they are used on page 3.
   - Define the acronym CDA when it is first used on page 6.
   - On page 5, is g in the link function the same as the g on lines 35–36? Presumably not.
   - In equation (2), shouldn't there be conditioning on y?
   - The sentence starting on line 8–9 of page 15 ("Thus, we use...") is difficult to parse.
   - In Figure 2(a) and similar trace plots, what variable is being plotted (vertical axis)?

(Note that page numbers above refer to page labels at the top of the review copy pdf. They are offset by one with respect to the page numbers in the footer of the manuscript.)