Referee Report for Streamlined Computing for Variational Inference with Higher Level Random Effects

This article extends the variational message passing approach proposed by Wand (2017) to handle multilevel random effects models. This paper deals, in particular, with up-to two levels of nesting. For example, one can have students nested in schools and schools nested in counties, for three levels of sampling units. The ideas presented here are enough to permit going to arbitrary levels of nesting, but this introduces enough extra bookkeeping that the manuscript does not cover this.

A few things are unclear to me about this manuscript. The main idea ultimately is to use sparse matrix algorithms to facilitate the variational message passing algorithm. The sparse matrix algorithms in Section 3, which form the basis of the methodology, are noted to be the same as those given by Pinheiro and Bates (2000). With this in mind, I have the following comments:

- I am very unclear on the benefit of adopting the factor graph formalism. At the optimal q-densities we have (for example)

  $$\log q(\beta, u) = \text{const} + \mathbb{E}_q \left\{ \log p(\beta, u \mid \text{everything else}) \mid \beta, u \right\}.$$

  Hence, updating the q's via coordinate ascent just requires computing full-conditional distributions, and taking a q-expectation. The relationship between these updates, the use of conjugate priors, and sufficient statistics in exponential families, are well known; in particular, for exponential families with conjugate priors, general expressions for the q-updates in terms of sufficient statistics are available. I find this fact much easier to wrap my head around, so I don't see why one would prefer factor graphs over just using a DAG. The use of DAGs to compute full-conditionals only requires looking at the parents and children of a node, and is done (for example) in `WinBUGS`.

The authors know all of this, so to reiterate, I think the paper would be improved if the authors can state clearly why they phrase things in terms of message passing.

- If I'm correct, and this all boils down to computing full-conditionals, then it seems like the key distinguishing features of the two-level and three-level models are the need to compute the full-conditional (β, u). I think the manuscript would be improved if the authors can explain why their streamlining algorithms do not apply to allow streamlining of a blocked Gibbs sampler, where the blocks are defined by the factorization in (3) and/or (4). This may clear up why we want to use the message-passing/factor-graph formalism in the first place for me. In general, I think a better comparison would be with a blocked Gibbs sampler (taking advantage of all sparse-matrix operations) rather than `STAN`.

- The authors present timing results comparing their method to a "naive" method. I'm a little unclear on what is being done. The manuscript says that (24) is being computed "directly." Does this mean it is being computed without leveraging sparsity? If that's the case, the method is so naive that I don't think anyone would ever do that. Having fit multilevel models, I know for certain that I would never invert these gigantic block matrices directly. I imagine that my first try would be to use a relatively simple Gibbs sampler, and I should be able to avoid needing to invert gigantic matrices if I use blocks that separate $\boldsymbol{u}$ from $\boldsymbol{\beta}$. Maybe that doesn't mix so well, although perhaps the ideas here would allow for a streamlined Gibbs sampler that blocks (u, β)?

- My opinion is that the paper is too long. The algorithms are certainly useful, but I think it is reasonable to sacrifice some amount of detail in the algorithms for a higher focus on the high-level ideas. My personal preference would be for many of the algorithms to be relegated to the appendix.

- I am very surprised at how close the variational Bayes intervals match the "exact" intervals. Can the authors comment on why this is the case? It is certainly not the case in general; in general, one expects the variational Bayes intervals to be far too narrow. Roughly speaking, the q-posterior is as concentrated as a full-conditional would be. One worries that readers might be invited to the casual conclusion that variational Bayes intervals have good coverage properties. Even in this case, we can see that the intervals are slightly too narrow. I also would like to see a principled assessment of the coverage properties of these intervals via simulation.

References

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-Plus*. Springer-Verlag, New York.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association*, 112(517):137–168.