Bayesian Analysis (0000) **00**, Number 0, pp. 1

Discussion of "Unified Framework for De-Duplication and Population Size Estimation" by Tancredi et al.

Nianqiao Ju[*,§] , Niloy Biswas[*], Pierre E. Jacob [*], Gonzalo Mena [*,†],
John O'Leary[*] and Emilia Pompe[‡]

We congratulate the authors on their important contribution to the record linkage literature, performing data de-duplication in the presence of population size uncertainty. The authors' method involves a mix of discrete parameters, such as a latent population size N and an unobserved partition over $\{1,\ldots,n\}$, and continuous ones like the vectors β_0, β' and θ. Estimating this model entails significant computational challenges, resembling those associated with mixture models and Bayesian non-parametric methods. These arise as much in the design of a sampling algorithm as in the assessment of its convergence.

We would like to highlight a few tools that can be used to build confidence in MCMC results under such conditions. Many of these involve Markov Chain couplings, as in Johnson (1996, 1998), and more recently in Glynn and Rhee (2014); Nikooienejad et al. (2016); Jacob et al. (2020); Biswas et al. (2019). These methods allow for the assessment and removal of the impact of the starting distribution. They apply when it is possible to run multiple chains that evolve marginally according to the proposed algorithm and jointly so that they meet after a random number of iterations.

In the following, we describe how to generate chains $X_t^{(1)} = (\eta^{(1)}, \beta_0^{(1)}\beta'^{(1)}, \theta^{(1)}, N^{(1)})$ and $X_t^{(2)} = (\eta^{(2)}, \beta_0^{(2)}, \beta'^{(2)}, \theta^{(2)}, N^{(2)})$ which follow the Gibbs sampler of Section 5 of this article and which meet exactly at a random time. A basic strategy for coupling Gibbs methods involves coupling each conditional update. The full conditional distribution of label indicators η is a Multinomial distribution on $\{1,\ldots,n\}$ with the vector of probabilities computed as in (5.5) of the article. To couple these updates, we compute such probabilities for both chains and implement a maximal coupling to obtain two labels which will be identical with the maximal probability. For the continuous variables β_0, β' and θ, which are updated with Metropolis–Hasting steps, we can employ maximal couplings of the Normal or Dirichlet proposal distributions, and use common Uniform draws to accept or reject them. Finally we can update N with an exact Gibbs step, truncating N to a very large integer, and implement a maximal coupling of this step.

However, an interesting difficulty arises, reminiscent of the infamous label switching issue (Stephens, 2000). Suppose that a first chain has $\eta^{(1)} = (1,4,3,4,2)$ and the second $\eta^{(2)} = (4,3,2,3,1)$, in a simple example with $n = 5$. These labels correspond to

[*] Department of Statistics, Harvard University
[†] Harvard Data Science Initiative
[‡] Department of Statistics, University of Oxford
[§] nju@g.harvard.edu

DOI: 0000

the same partition, and yet the η-components of the chains are different and thus the chains cannot coincide. Judging from our toy experiments, the associated meeting time would be long. This can be alleviated by an additional relabeling step, to be performed after the update of the components of η. A simple strategy, for example, is to relabel η according to the order of the occurrences of new blocks, from component 1 to n. That is, both the labels $\eta^{(1)} = (1, 4, 3, 4, 2)$ and $\eta^{(2)} = (4, 3, 2, 3, 1)$ would be relabelled $(1, 2, 3, 2, 4)$. This relabeling needs to be accompanied by an adequate reshuffling of the associated parameters, namely the β'-components in the notation of Section 5. Other, more sophisticated relabeling strategies could be devised, perhaps inspired by the literature on label-switching issues in mixture models (Stephens, 2000; Marin et al., 2005; Frühwirth-Schnatter, 2011).

We will make some R scripts implementing a coupling of the proposed Gibbs sampler available at `https://github.com/EmiliaPompe/discussion_unified_framework`, along with some simple numerical experiments on the synthetic dataset `RLdata500` from the `R` package `RecordLinkage` analysed in Section 6 of the paper.

References

Biswas, N., Jacob, P. E., and Vanetti, P. (2019). "Estimating convergence of Markov chains with L-lag couplings." In *Advances in Neural Information Processing Systems*, 7389–7399. 1

Frühwirth-Schnatter, S. (2011). "Label switching under model uncertainty." *Mixtures: Estimation and Application*, 213–239. 2

Glynn, P. W. and Rhee, C.-h. (2014). "Exact estimation for Markov chain equilibrium expectations." *Journal of Applied Probability*, 51(A): 377–389. 1

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). "Unbiased Markov chain Monte Carlo with couplings." *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (with discussion) (to appear)*. 1

Johnson, V. E. (1996). "Studying Convergence of Markov Chain Monte Carlo Algorithms Using Coupled Sample Paths." *Journal of the American Statistical Association*, 91(433): 154–166. 1

— (1998). "A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms." *Journal of the American Statistical Association*, 93(441): 238–248. 1

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). "Bayesian modelling and inference on mixtures of distributions." *Handbook of statistics*, 25: 459–507. 2

Nikooienejad, A., Wang, W., and Johnson, V. E. (2016). "Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors." *Bioinformatics*, 32(9): 1338–1345.
URL `https://doi.org/10.1093/bioinformatics/btv764` 1

Stephens, M. (2000). “Dealing with label switching in mixture models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809. 1, 2