Bayesian Analysis (2010) **5**, Number 3, pp. 461–464

Rejoinder

Ioanna Manolopoulou*, Cliburn Chan† and Mike West‡

We thank the discussants, Fabio Rigat and Nick Whiteley, for their insightful and positive comments. They suggest a number of potential directions for extension of the work and raise connections with other research. We address the points they raise in connection with broader modeling and communication considerations, followed by specific aspects and details of computational strategy.

1 Modeling and Communication

Discussion comments on general questions of applied statistical modeling relate to the need for attention to a balance between contextual/applied interests and statistical modeling refinements motivated by an application. A good deal of time and effort in collaborations and applied work is spent on communication of the relevance and roles of complex Bayesian models to non-statistical disciplinary scientists.

The specific setting here is that of non-parametric Bayesian mixture models. These models are nowadays standard and widely accepted by statistical and machine learning communities. Their demonstrated success in applications in many areas in the last decade or so has done much to foster understanding and appreciation among disciplinary scientists. In our current applied context of cell subtype characterisation in flow cytometry studies, mixture models are established (e.g. Chan et al. 2008; Pyne et al. 2009). For the purposes of communication we have promoted non-parametric DP mixtures as really just direct extensions of standard mixtures that allow for uncertainty about the (practically effective) number of components. That is easily communicated and the remaining technical aspect of note is just the use of effectively standard class of priors over component parameters. A substantial practical modeling bridge in our work in these applications is the clustering of subsets of Gaussian DP mixture components according to concentration around inferred locals modes in the distribution, and putative interpretation of some of these clustered components as defining (resulting non-Gaussian) subpopulations of biological interest.

On the specific question of inference on the concentration hyperparameter α, which again has been standard in the literature since the early 1990s, we note that this hierarchical model specification has the usual goals and attributes of inducing a degree of robustness while incurring negligible additional computational (Escobar and West 1995; Ishwaran and James 2002). Although the number of components has no immediate biological interpretation (other than as a gross upper bound on the number of subtypes) and so the role of α in its impact on the number of components is only of technical

*Department of Statistical Science, Duke University, Durham, NC, mailto:im30@stat.duke.edu
†Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, mailto:cliburn.chan@duke.edu
‡Department of Statistical Science, Duke University, Durham, NC, mailto:mw@stat.duke.edu

 DOI:10.1214/10-BA517REJ

interest, inference on such hyper-parameters does have practical utility in, particularly, studies comparing data sets, as it does in other, related areas of applications of DP mixtures when comparisons are of interest (e.g. Ji et al. 2009).

2 Computational Strategies and Details

Whiteley is interested in specific formalizations of our adaptive selection strategy as an optimization problem. Such an approach would certainly be of theoretical interest, and could lead to a broader view of adaptive data sampling problems in a Bayesian decision theoretic framework. To develop such an interpretation – and define a class of utility/objective functions that underlie it – seems challenging; as we point out in the Appendix of the paper, maximizing the amount of information about a single component of interest results in an intractable optimization problem and is strongly affected by the particular structure of each data set (reflected through π, μ, Σ). However, the strategy is inherently derived from contextual goals/utilities so future work to explore a formal decision-theoretic treatments is to be encouraged.

The stopping rule for data selection is clearly, as noted in the discussions, a central element of the analysis. The stopping rule amounts to sampling until the estimated number of "relevant" observations left to resample is deemed "small enough". The specific form of the stopping criterion used in our examples may seem ad-hoc, but it stems directly from concern about the practical concordance between sample size and the infinite population structure, which in practice is the assumption that few observations carry almost all the weight in the targeted resampling. The specific formulation allows for the number of markers to scale additively and thus requires a relatively small p in order to ensure a wide enough distribution of weight over the unsampled observations. It has also been honed and refined in multiple empirical studies. However, this is one specific choice for stopping rule and we are certainly open to generalisations or alternatives. A more sophisticated $c_{threshold}$ might similarly incorporate τ such that $c_{threshold} = \exp(-\sum_{i=1}^{p}(4\tau_{ii})^{-1})$, for example. Again, a broader decision theoretic view of the stopping rule may also be worth exploring.

Whiteley asks about choices of the size of the random subsample (n^R), the size of the targeted subsample batches (B), as well as MCMC/SMC run lengths and number of chains. On n^R, an inherent trade-off is involved. A larger n^R is expected to improve M-H acceptance rates and precision of inferences on targeted components, but at a computational cost. In the examples presented in the paper, we used relatively small values of n^R in order to demonstrate the potential of our methods even in cases where the size of the initial subsample is small. With advances in computation for MCMC (and other algorithms) in mixtures (e.g. Suchard et al. 2010) larger sample sizes can be entertained so that larger values of n^R can be used. The batch size B might perhaps be best viewed as a function of n^R, e.g., $B = bn^R$ for some small fraction b. This leads to the expectation of relatively modest changes in the SMC updated posteriors at each batch processing step, requiring a smaller number of Metropolis-Hastings updates. A large number of Metropolis-Hastings steps at the end of the SMC sampler ensures

convergence to the final target distribution, which can be checked through the effective sample size, for example.

Choices of L and run lengths are standard questions in all MCMC and SMC/M-H implementations, and there is really nothing new to add to the existing, generic MC theoretical insights overlaid by cumulated experience and wisdom of the simulation-based Bayesian community. The number of chains L is linked to the dimension of the data p and the size of the initial subsample n_U, both of which determine the dimensionality of the parameters μ, Σ and z^R. The final size of the targeted subsample n_B will affect the efficiency of the proposals and hence the effective sample size of the particles, and thus the number of chains L required.

We note also that resampling of the particles is not possible unless the configuration indicators z^R are mutated in the SMC steps as well. In our current formulation in the paper, samples of z^R are fixed for each particle, so that resampling would result in poor coverage of the z^R space. Instead, larger values of n^R and efficient proposals are crucial. Emerging ideas of adaptive variants of SMC methods, such as suggested and referenced by Whiteley, can also help to reduce particle degeneracy, and are very worth considering for extensions/modifications of the current approach; here, a schedule of adaptively choosing B at each SMC iteration could help with controlling the difference between target distributions at each step, for example.

References

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. (2008). "Statistical mixture modeling for cell subtype identification in flow cytometry." *Cytometry A*, 73: 693–701. 461

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. 461

Ishwaran, H. and James, L. (2002). "Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information." *Journal of Computational and Graphical Statistics*, 11: 508–532. 461

Ji, C., Merl, D., Kepler, T., and West, M. (2009). "Spatial mixture modelling for unobserved point processes: Application to immunofluorescence histology." *Bayesian Analysis*, 4: 297–316. 462

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jagera, P. L., and Mesirova, J. P. (2009). "Automated high-dimensional flow cytometric data analysis." *Proceedings of the National Academy of Sciences*, 106(21): 8519. 461

Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010). "Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures." *Journal of Computational and Graphical Statistics*, 19: 419–438. 462