Bayesian Analysis (2010) **5**, Number 3, pp. 457–460

Comment on Article by Manolopoulou et al.

Nick Whiteley*

This article addresses the problem of developing efficient methods for performing inference when faced with very large data sets. The authors focus on a mixture modelling problem arising in biology. Here the mixture model is used to classify and discriminate between cell sub-types, the main interest being in the parameters associated with a low-probability mixture component, with the latter identified by placing an ordering constraint on the unobserved mixture weights. Computational methods for sampling from the full Bayesian posterior in mixture models have been studied at great length. Due to the model structure and the relative simplicity of its implementation, Gibbs sampling is a popular choice, although it is widely recognized that this type of approach can suffer from very poor mixing characteristics and there are various alternatives which can be much more effective.

In any case, when the total number of data points is large, the cost of function evaluations required as part of each MCMC iteration can be rather high. The authors propose a method to avoid some of this cost, performing an approximation of full Bayesian inference via a combination of Monte Carlo methods. Their idea is to avoid processing all observations and concentrate computational effort on those data which are, in some sense, most relevant to the mixture component of interest. This is an intriguing idea. In the role of a discussant I will take this opportunity to pose some questions for the authors regarding the principle of their method and to highlight some characteristics of the Monte Carlo methods they employ.

In my understanding, there are two conceptual components to the proposed method:

- On the basis of an initial subset of the data, design an adaptive, sequential data selection scheme, with the data subsets entering the definition of a sequence of approximate posterior distributions.
- Use Monte Carlo methods to sample from this sequence of distributions, whilst also updating the data selection scheme.

1 Principles of the approach

As the authors stress in the abstract and elsewhere in the article, their main concern is over the trade-off between computational cost arising from the amount of data processed and information obtained about particular quantities of interest. Upon reading the abstract, my first impression was that this is naturally approached as a type of optimal design problem: one is faced with a choice between different data collection strategies (indexed in the present context by values of the weighting function parameters) and

*Department of Mathematics, University of Bristol, Bristol, U.K., mailto:nick.whiteley@bristol.ac.uk

 DOI:10.1214/10-BA517B

there is a stated requirement for the chosen strategy to fulfill certain criteria, in the presence of uncertainty. However, this is not the approach which the authors adopt. Most notably, they do not express a specific, quantitative characterization of the information which they seek, or their preference for its accuracy. My first question to the authors is: why did they not pursue an explicit formulation as an optimization problem, with a utility function, optimality criterion, etc.? The authors hint at sequential design in section 6 and I wondered if they had a formal optimization procedure in mind when constructing their method, whether they can comment on what a procedure of this sort would involve, or if they have reasons to consider such an approach inappropriate.

Apart from being a central component of an explicit optimization approach to the problem, specification of a quantitative measure of the gain in information associated with a particular data sampling strategy is important for assessing the effectiveness of the proposed method, at least empirically. In the example of section 5.1, the authors comment on the performance of their method in terms of the similarity in concentration and location between an approximate marginal distribution they obtain and a corresponding marginal of the full posterior. On the other hand, in the case of the example of section 5.2, and the discussion of appendix A, the authors state:

"Here it's not possible to draw a direct comparison between the posterior distributions of the component k^*, because the component structure in the random, targeted and full data set case changes significantly."

These two examples and the lack of a clear criterion for performance leave me unsure as to the precise aims of the method and with no clear way to empirically characterize the effectiveness of the proposed approach. The numerical examples rely on a comparison with the full posterior, and whilst I realize that this is natural for purposes of exposition, what reassurances can the authors provide about the approximate posteriors in the general case? As with any form of approximation, it is natural to ask what it is we are losing.

One of the key ideas of the article is the initial random selection of a subset of the data, on the basis of which to construct a mechanism for subsequent observation selection. As the authors comment in the discussion of section 6, the method they propose is highly sensitive to the size of this initial subset. Can the authors suggest some form of guideline as to how to choose this size? A requirement to approaching this in a principled manner points again to the topic of optimal design and the need for a quantitative performance criterion.

2 Monte Carlo scheme

In section 5 of the article the authors introduce a form of sequential Monte Carlo method. My understanding is that this involves sampling a collection of processes, each conditioned on a single draw of z^R from the posterior associated with the initial random sample. The authors adopt the distributional approximation displayed in equation (20) in order avoid the computational expense which would be required to well-approximate

the joint posterior of the parameters, z^R and z^T, updated as more observations are considered.

Conditional on these draws from the initial posterior, each of the processes evolves according to an inhomogeneous sequence of MCMC kernels, where the inhomogeneity arises from the variation of the kernels' target distributions, adapted to the joint history of the processes and the auxiliary stochasticity arising from the random target data selection. This is a non-standard algorithm and I found some of the indexing used in the specification of the algorithm confusing. For example, j seems to be used to index the processes (the particles), but also seems to index time steps. Is this what the authors intended?

Regarding the conditioning of each sampled process on a single draw from the initial posterior, there seems to be some algorithmic inefficiency here as, at least in the implementation which the authors have made publicly available, this leaves many of the initial samples unused. Isn't this wasteful?

Another design issue which arises when using the proposed sequential method is how to choose B, the size of the block of observations incorporated at each time step. An intimately related issue, not mentioned explicitly in the article, is the number of iterations of the MCMC kernel to employ at each time step. For simplicity, consider the weighting function as fixed, assume a non-random ordering of the targeted observations for incorporation and condition on the draws from the initial posterior distribution. In this case, the proposed algorithm amounts to running a collection of independent, inhomogeneous Markov chains.

The chains are not interacting via a resampling mechanism: the ability of these chains to "keep up" with the changing target distributions (and therefore ultimately reach stationarity with respect to the final distribution of interest), relies solely on the ergodicity properties of the kernels employed. Changing the target distributions too rapidly (roughly corresponding to large B), or using too few MCMC iterations at each time step, can result in poor performance. This is a recognized phenomenon in the context of standard sequential Monte Carlo methods and it seems reasonable to conjecture that the proposed adaptive method is susceptible to the same issue.

In the case of more standard Sequential Monte Carlo methods employing a resampling mechanism, a technique for adaptively choosing the sequence of distributions, as the simulation progresses, has recently emerged (Jasra et al., 2010). This involves choosing the sequence of distributions so as to prevent the effective sample size of the particle population (computed in terms of a collection of importance weights) falling below a given threshold. Whilst the theoretical consequences of this form of adaptation are yet to be characterized, this is a natural method for automated selection of the sequence of target distributions and resampling times. This approach does not transfer to algorithms without resampling. In the algorithm proposed in the article under consideration, the authors appear to choose B manually. Can they suggest any guidelines?

Overall, I think that the general topic which the article addresses is an interesting

one. However I am not convinced that the proposed methods are entirely well justified. I found a quantitative specification of the aim of the method lacking, this seems important for both motivating the method and assessing its performance. Guidelines for choosing various tuning parameters of the method would be very welcome.

References

Jasra, A., Stephens, D.A., Doucet, A., and Tsagaris, T. (2010). "Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo", *Scandinavian Journal of Statistics.* (To appear).