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