Bayesian Analysis (2007) **2**, Number 3, pp. 479–482

Comment on Article by Jain and Neal

C.P. Robert*

From a stylistic point of view, I think this paper reads very much like a sequel to the important paper Jain and Neal (2004) and therefore it is not exactly self-contained since the main bulk of the paper is a commentary of the program provided in Section 4.2. Instead of the current version, I would thus have preferred a truly self-contained version with a more user-friendly introduction, for instance when reading and re-reading Sections 3 and 4.1...[1]

The central point of the paper is to extend Jain and Neal (2004) so that the lack of complete conjugacy of the prior does not prevent the algorithm from being run. Indeed, in Jain and Neal (2004), the model parameters are completely hidden in that the likelihood and the prior only depend on the cluster index vector $\mathbf{c}$, which means working in a finite set. The difficulty with priors G_0 that do not lead to closed form marginals is that the parameters must take part in the simulation process. The idea at the core of the current paper is to take advantage of the conditional conjugacy, i.e. the fact that the prior on a given parameter is still conjugate and thus manageable, conditional on all the other parameters, so that a Gibbs sampling version can be implemented.

At this stage, I understand the rationale of the partial conjugacy for the Metropolis-Hastings ratio to be computed (Section 4.1) but I wonder how difficult it would be to extend the idea to any type of prior distribution. I also note that at both split and merge stages the algorithm simulates new values of the parameter from the *prior* distribution, rather than from a more adapted distribution. This is as generic as it can be, but simulating from vague priors usually slows down algorithms and it is of course impossible for improper priors. It thus seems to me that the factor t directing the number of intermediate Gibbs (or Metropolis-Hastings) iterations in Step 3 must be influential in the overall behaviour of the algorithm and that large values of t may be necessary to overcome the dependence on the starting value.

More generally, I also wonder why a more global tempering strategy would not fare better than the local split-merge proposals used in the paper. For illustration purposes, I implemented below the regular Gibbs sampler in the [BetaBinomial] Example 1 of Jain and Neal (2004) and compared it with a naïve tempered version where the tempered likelihood L_τ is made of a product of $\tau \geq 1$ (sub)likelihoods based on a partition of the observations in τ random clusters, τ being itself uniform on $\{1, \ldots, n/2\}$. (The advantages of using this form of tempering are (a) that the same Gibbs sampler can be used for the sublikelihoods and (b) that the normalising constant of the tempered version is still available, as opposed to the choice of a power of the likelihood. The acceptance probability at the end of the tempered moves is then function of the likelihood ratio $L(\theta|x)/L_\tau(\theta|x)$ and can be directly computed.) As shown on Figure 1 *(bottom)*,

*CREST and CEREMADE, Uni. Paris Dauphine, France, mailto:xian@ceremade.dauphine.fr

[1]This may explain why the following reads more like an eloped referee's report than like a true discussion!

explained below, the mixing and the exploration of various likelihood values is quite improved with this tempered scheme, since no column sticks to a single colour theme.

Since Dirichlet mixtures are closely related to mixtures, I would have liked to read some discussion on the label switching phenomenon (see, e.g., Stephens 2000; Marin et al. 2005; Jasra et al. 2005). Indeed, while the original model of Jain and Neal (2004) is somehow impervious to the issue of label switching, since the clustering parameterisation only focus on class allocations, the introduction of the parameter in the game means that a proper exploration of the posterior requires the reproduction of the symmetry in the various components of the mixture. Using a split-merge basis for this exploration may then prove to be insufficiently powerful for this task.

In fact, it is close to impossible to judge of the overall convergence performances from the simulation output, which solely concentrates on the cluster sizes. Additional graphical summaries would be welcome, like the "allocation map" advertised in Robert and Casella (2004) and represented on both Figures 1 and 2. The pixelised lines on the pictures represent the cluster index via different colours for all observations, the index on the first axis being the index of the observation. The second axis corresponds to the iteration index. Long vertical stripes of similar colours indicate poor mixing of the algorithm.

In this illustration, we see clearly that the 5 equal groups of Example 1 of Jain and Neal (2004) are identified by the Gibbs sampler–as signalled by the homogeneous columns $1-20$, $21-40$, $41-60$, $61-80$ and $81-100$—and, furthermore, that label switching does occur, even if at a very slow pace—as shown by columns $61-80$ for instance.

A point of detail (?) is that the algorithm must be (is) validated as a Gibbs procedure rather than as a Metropolis-Hastings algorithm, given that at any stage only a subset of the parameters and of the clustering indicators is updated. In addition, this is quite an interesting example of algorithmic bypassing the varying dimension pitfalls, since it avoids dealing with the measure theoretic subtleties encountered by reversible jump for instance (Green 1995) while being in a continuous varying dimension state space, contrary to the setup of Jain and Neal (2004).

References

Green, P. (1995). "Reversible jump MCMC computation and Bayesian model determination." *Biometrika*, 82(4): 711–732. 480

Jain, S. and Neal, R. (2004). "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model." *J. Computat. Graphical Statist.*, 13(1): 158–182. 479, 480, 481

Jasra, A., Holmes, C., and Stephens, D. (2005). "Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling." *Statistical Science*, 20(1): 50–67. 480

Marin, J., Mengersen, K., and Robert, C. (2005). "Bayesian Modelling and Inference

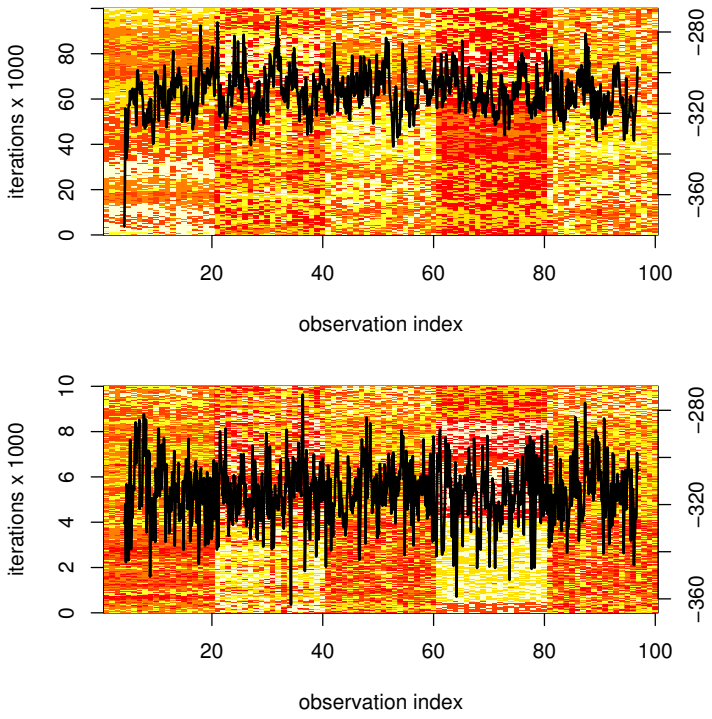

Figure 1: *(top)* Allocation map of the simulated cluster index vector $\mathbf{c}^{(t)}$ for $m = 6$, $n = 100$ observations and $T = 10^5$ Gibbs iterations (subsampled every 1000 iteration), in the setup of Example 1 of Jain and Neal (2004). The colours used in the graphs range from red (1) to white (6) and identify the labels of the cluster indicators c_i along the iterations. The superimposed graph is the corresponding sequence of likelihood values over the $T = 10^5$ Gibbs iterations, associated with the scale on the right hand side. *(bottom)* Same representation for a tempered version with $T = 10^3$ iterations made of $T_o = 10^2$ tempered moves.

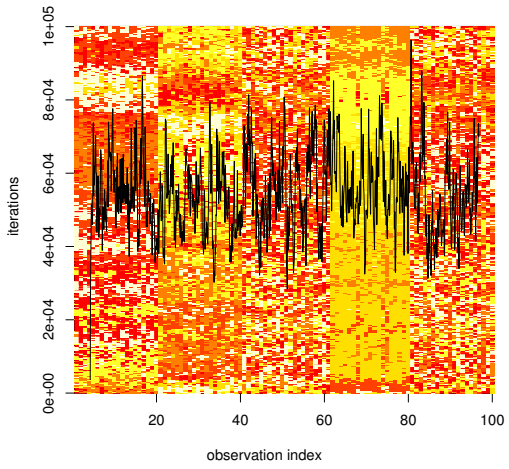

Figure 2: Same representation as Figure 1 for another run of the Gibbs sampler,

on Mixtures of Distributions." In Rao, C. and Dey, D. (eds.), *Handbook of Statistics*, volume 25. Springer-Verlag, New York. 480

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition. 480

Stephens, M. (2000). "Dealing with label switching in mixture models." *J. Royal Statist. Soc. Series B*, 62(4): 795–809. 480