THE UNIVERSITY OF
WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

This paper is made available online in accordance with
publisher policies. Please scroll down to view the document
itself. Please refer to the repository record for this item and our
policy information available from the repository home page for
further information.

To see the final version of this paper please visit the publisher's website.
Access to the published version may require a subscription.

Author(s): Papaspiliopoulos O
Article Title: A note on posterior sampling from Dirichlet mixture models
Year of publication: 2008
Link to published article:
http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2008/08-20wv2.pdf

Publisher statement: None

A note on posterior sampling from Dirichlet mixture models

By OMIROS PAPASPILIOPOULOS

Department of Economics, Universitat Pompeu Fabra, Barcelona

omiros.papaspiliopoulos@upf.edu

SUMMARY

In this note we observe that the recent MCMC methods of Papaspiliopoulos & Roberts (2008) and Walker (2007) for Dirichlet mixture models are intrinsically connected and can be naturally combined to yield an algorithm which is better (in terms of mixing), faster (in terms of execution time) and easier (in terms of implementation and coding) than either of them.

Some keywords: Retrospective sampling; Slice sampling; Augmentation schemes; Label switching; ; Stick-breaking priors; blocking strategies

1. Introduction

MCMC-assisted posterior inference for Dirichlet mixture models based on a so-called conditional augmentation scheme is becoming increasingly popular. This augmentation scheme gives added flexibility in complex hierarchical structures and consists of augmenting rather than integrating out the random measure and updating it in the MCMC algorithm.

Two methods have been proposed in this context which achieve posterior simulation without resorting to approximations; the retrospective MCMC method of Papaspiliopoulos & Roberts (2008) and the slice sampling method of Walker (2007). Aiming to demonstrate how they are connected, and to devise a new composite algorithm, we first consider the toy (but very illustrative) problem of sampling from the Dirichlet Process Prior (DPP).

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

2. Representations and simulation from the DPP

Let $X = (X_1, \ldots, X_n)$ be a sample from the DPP. We can represent the distribution of X with the following hierarchy

CONDITIONAL AUGMENTATION A

$$\begin{aligned} X_i &= Z_{K_i} \\ K_i \mid p &\sim \sum_{j=1}^{\infty} p_j \delta_j(\cdot) \\ Z_j \mid \Theta &\sim H_\Theta, \ j = 1, 2, \ldots \\ p_1 = V_1, &\qquad p_j = (1 - V_1)(1 - V_2) \cdots (1 - V_{j-1}) V_j, \ j \geq 2 \\ V_j &\sim \mathrm{Be}(1, \alpha)\,. \end{aligned} \tag{1}$$

Here $V = (V_1, V_2, \ldots)$ and $Z = (Z_1, Z_2, \ldots)$ are vectors of independent variables and are independent of each other, the K_is are independent given $p = (p_1, p_2, \ldots)$. Let $U_i, i = 1, \ldots, n$ be independent draws from a uniform distribution on $(0, 1)$. Then we set $K_i = j$ if and only if

$$\sum_{l=0}^{j-1} p_l < U_i \leq \sum_{l=1}^{j} p_l \ , \tag{2}$$

where we define $p_0 = 0$. This can be done in finite time retrospectively by first simulating the U_i and then pairs of (V_j, Z_j) until (2) is satisfied (see Algorithm 1 of Papaspiliopoulos & Roberts, 2008). Sampling in finite time from an infinite mixture is facilitated by retrospective sampling, the Markovian structure of the p_j's and the independence of the Z_j's.

Walker (2007) makes the neat observation that, in the spirit of slice sampling, rather than treating $U = (U_1, \ldots, U_n)$ as auxiliary variables used in the simulation, we can

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

augment them directly in the hierarchical model, and write the DPP as

CONDITIONAL AUGMENTATION B

$$
\begin{aligned}
X_i &= Z_{K_i} \\
f(K_t, U_t \mid p) &= \sum_{j:p_j > U_t} \delta_j(\cdot) = \sum_{j=1}^{\infty} 1[U_t < p_j]\delta_j(\cdot) \qquad (3) \\
Z_j \mid \Theta &\sim H_\Theta,\ j = 1, 2, \ldots \\
p_1 = V_1, &\qquad p_j = (1 - V_1)(1 - V_2) \cdots (1 - V_{j-1})V_j,\ j \geq 2 \\
V_j &\sim \mathrm{Be}(1, \alpha)\,.
\end{aligned}
$$

Note that in (3) we specify the joint density of K_i and U_i. This is a standard representation for random variables, applied here to a discrete random variable. For unimodal densities it relates to Khinchine's theorem (see Section 6.2 of Devroye, 1986). Note that trivially the marginal for K_i is (1). However, the marginal for U is not uniform, it has a monotonically decreasing density, which is typical in this type of construction, and it will be typically intractable. We make two key observations here.

1. In terms of simulating from the DPP the first augmentation is much more convenient, since we can first simulate U_i from its marginal and consecutively K_i from its conditional using retrospective sampling. This is not feasible in the second augmentation, where neither of the marginals in (3) are tractable. On the other hand it is trivial to sample from this density using a Gibbs sampler since the conditionals have a very simple structure.

2. The parametrization of the DPP in terms of (K, Z, V) in scheme A is obtained as a marginal of the parametrisation in terms of (K, Z, V, U). This is key for the new algorithm we propose in §4.

3. MCMC sampling for Dirichlet mixture models

According to a Dirichlet mixture model (MDP hereon which stands for mixture of Dirichlet processes) the data depend on the DPP in the following way. Let $f(y \mid z, \lambda)$ be a

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

parametric density with parameters z and λ, then

$$Y_i \mid (Z, K) \sim f(Y_i \mid Z_{K_i}, \lambda),\ i = 1, \ldots, n \tag{4}$$

and the data are conditionally independent given (K, Z). The aim is to infer about the posterior distribution of (K, V, Z) given Y. We are typically interested in inferring the hyperparameters, in particular α, and we will discuss this at the end of §4.

Samples from the posterior distribution (K, V, Z) are obtained by Gibbs sampling from the corresponding conditionals. Papaspiliopoulos & Roberts (2008) work under Conditional Augmentation A. Their algorithm is based on Proposition 1 of that paper which establishes that conditionally on $(Y, K, \alpha, \lambda, \Theta)$, Z and V are independent of each other and consist of independent variables with distributions easy to simulate from. The complexity in their algorithm arises in the simulation of K from its full conditional. Conditionally on (Y, V, Z, λ), K is independent of (Θ, α) and it consist of conditionally independent elements with

$$\mathrm{pr}\{K_i = j \mid Y, V, Z, \lambda\} \propto p_j f(Y_i \mid Z_j, \lambda),\ j = 1, 2, \ldots . \tag{5}$$

Direct simulation from this distribution is difficult due to the intractability of the normalising constant, which involves an infinite summation. In the article two methods are suggested for simulation from this conditional; a Metropolis-Hastings step, which requires a careful calculation of the acceptance probability, and a direct simulation step which involves a retrospective simulation based on upper and lower bounds for (6).

Summarising, the algorithm of Papaspiliopoulos & Roberts (2008) involves very simple steps for the updates of Z and V, but a considerable harder step for K. In that article it is also emphasized the importance of label switching moves and two such moves are suggested.

On the other hand, Walker (2007) performs Gibbs sampling under Conditional Augmentation B, i.e. sampling of (V, Z, K, U) from their full conditionals. Conditionally on the rest, U consists of conditional independent elements with truncated uniform distributions. The conditional of Z is the same as in the algorithm of Papaspiliopoulos & Roberts

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

(2008). The augmentation of U simplifies considerably the conditional distribution of K which is now concentrated on a finite set;

$$\text{pr}\{K_i = j \mid Y, V, Z, U, \lambda\} \propto f(Y_i \mid Z_j, \lambda),\ j : p_j > U_i\,. \tag{6}$$

Walker (2007) gives a sufficient condition for identifying all those js such that $p_j > U_i$: they belong to the set $\{1, \ldots, j_i^*\}$ where j_i^* is the smallest l such that $\sum_{j=1}^{l} p_j > 1 - u_i$. Crucially, from a computational point of view, $j^{*(n)} := \max_{1 \le i \le n} j_i^*$ grows only logarithmically with n.

Nevertheless, the conditioning on U creates global constraints on the Vs:

$$p_j > U_i\,, \forall i = 1, \ldots, n\,. \tag{7}$$

The easiest way to simulate from this constrained distribution is by single site Gibbs sampling of the V_js. (Note that the constraint applies only to $j \le \max_i K_i$).

Summarising, the algorithm of Walker (2007) leads to a very easy update of K, but at the expense of a single site Gibbs sampling of V. Additionally, due to the constraints (8) it is difficult to add label switching moves.

4. A new algorithm: Exact Block Gibbs sampler

The fact that Conditional Augmentation A is a marginal of Conditional Augmentation B provides us with a natural way to combine both schemes. Consider the following algorithm:

EXACT BLOCK GIBBS SAMPLER

Give an initial allocation K. Iterate the following steps.

Step 1. *Simulated jointly (V, U) conditionally on the rest.*

Step 1.1 *Simulate V from the marginal (w.r.t U) conditional distribution, as described in Papaspiliopoulos & Roberts (2008).*

Step 1.2 *Simulate U from its full conditional, as described in Walker (2007).*

Step 2. *Simulate Z_j from its full conditional.*

Step 3. *Simulate K from its full conditional as described in Walker (2007).*

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

Note that this block Gibbs sampler combines the advantages of both algorithms. The update of V is simple, since conditioning upon U is avoided. Additionally, the update of K is easy since the conditioning on U removes the problem of intractable normalising constant. In this setup label-switching moves can also be added, provided they are also blocked with U, as in Step 1.1.

Figure 1 contains a small simulation study of the performance of the different algorithms. This is by no means a thorough investigation, which is underway. We work with the 'bimod 100' dataset of Papaspiliopoulos & Roberts (2008) which consists of 100 draws Y_i from the bimodal mixture, $0.5N(-1, 0.5^2) + 0.5N(1, 0.5^2)$. We fit a non-conjugate MDP model, where f is a Gaussian likelihood, and Z_j consists of the mean and the precision of the Gaussian density. Details are given in Section 4 of Papaspiliopoulos & Roberts (2008). We take $\alpha = 1$, and contrast four algorithms: the Retrospective MCMC of Papaspiliopoulos & Roberts (2008) (Retro), the slice sampler of Walker (2007) (Slice), our new Block Gibbs algorithm (Block), and the Block Gibbs algorithm with label switching moves (Block label). The relative computational times for an iteration of each algorithm "in stationarity" are Retro 2.48, Slice 1, Block 0.98, Block label 1. One can see that there are also big computational gains from the combination of the algorithms. This experiment does not bring out the importance of label switching moves, but the extensive study of Papaspiliopoulos & Roberts (2008) shows that such moves are very important for larger data sets and different values of α.

Updates of hyperparameters can be easily added. Of particular interest is inference for α, which can be done as discussed in Walker (2007). The conditional approach can be very easily extended to other stick-breaking models, in particular the two-parameter Poisson-Dirichlet process. The Exact Block Gibbs algorithm can be immediately exported in these more general setups.

References

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag.

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

Papaspiliopoulos, O. & Roberts, G. O. (2008). Retrospective mcmc for dirichlet process hierarchical models. *Biometrika* 95 169–186.

Walker, S. (2007). Sampling the dirichlet mixture model with slices. *Comm. Statist. Sim. Comput.* 36 45–54.

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism

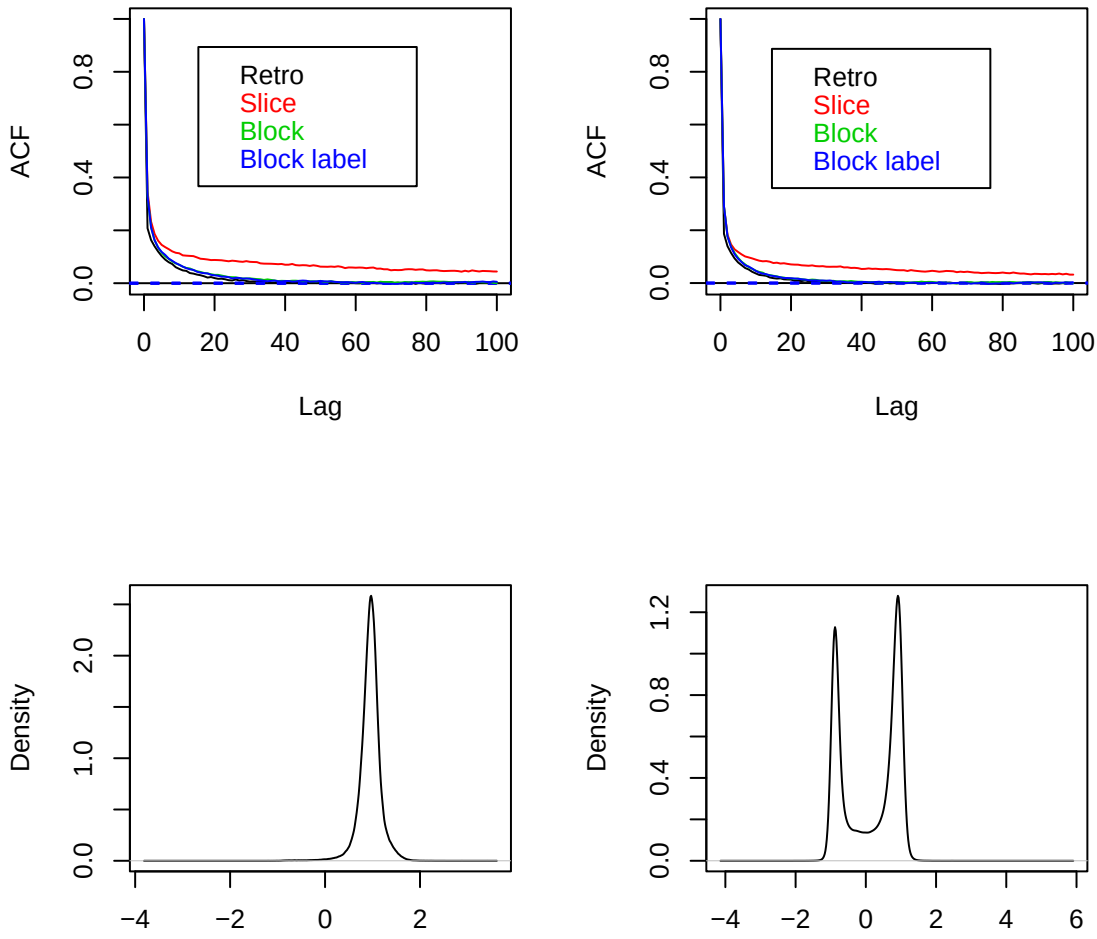

Figure 1: Top: MCMC summaries. Autocorrelation plots which correspond to Z_{K_i}, for $i = 3$ (left), and $i = 2$ (right), for the Retrospective MCMC (Retro), the slice sampler (Slice), our new Block Gibbs algorithm (Block), and the Block Gibbs algorithm with label switching moves (Block label). Bottom: the posterior densities from which the chains are sampling from, Z_{K_i}, for $i = 3$ (left), and $i = 2$ (right). This is the bimod-100 dataset of Papaspiliopoulos & Roberts (2008), with α fixed to 1.

CRiSM Paper No. 8-20v2, www.warwick.ac.uk/go/crism