



---

Markov Chain Sampling Methods for Dirichlet Process Mixture Models

Author(s): Radford M. Neal

Source: *Journal of Computational and Graphical Statistics*, Jun., 2000, Vol. 9, No. 2 (Jun., 2000), pp. 249-265

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America

Stable URL: <https://www.jstor.org/stable/1390653>

#### REFERENCES

Linked references are available on JSTOR for this article:

[https://www.jstor.org/stable/1390653?seq=1&cid=pdf-reference#references\\_tab\\_contents](https://www.jstor.org/stable/1390653?seq=1&cid=pdf-reference#references_tab_contents)

You may need to log in to JSTOR to access the linked references.

---

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



Taylor & Francis, Ltd., American Statistical Association, and Institute of Mathematical Statistics are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Computational and Graphical Statistics*

JSTOR

# Markov Chain Sampling Methods for Dirichlet Process Mixture Models

Radford M. NEAL

This article reviews Markov chain methods for sampling from the posterior distribution of a Dirichlet process mixture model and presents two new classes of methods. One new approach is to make Metropolis–Hastings updates of the indicators specifying which mixture component is associated with each observation, perhaps supplemented with a partial form of Gibbs sampling. The other new approach extends Gibbs sampling for these indicators by using a set of auxiliary parameters. These methods are simple to implement and are more efficient than previous ways of handling general Dirichlet process mixture models with non-conjugate priors.

**Key Words:** Auxiliary variable methods; Density estimation; Latent class models; Monte Carlo; Metropolis–Hasting algorithm.

## 1. INTRODUCTION

Modeling a distribution as a mixture of simpler distributions is useful both as a nonparametric density estimation method and as a way of identifying latent classes that can explain the dependencies observed between variables. Mixtures with a countably infinite number of components can reasonably be handled in a Bayesian framework by employing a prior distribution for mixing proportions, such as a Dirichlet process, that leads to a few of these components dominating. Use of countably infinite mixtures bypasses the need to determine the “correct” number of components in a finite mixture model, a task which is fraught with technical difficulties. In many contexts, a countably infinite mixture is also a more realistic model than a mixture with a small number of components.

Use of Dirichlet process mixture models has become computationally feasible with the development of Markov chain methods for sampling from the posterior distribution of the parameters of the component distributions and/or of the associations of mixture components with observations. Methods based on Gibbs sampling can easily be implemented for models based on conjugate prior distributions, but when non-conjugate priors are used, as is appropriate in many contexts, straightforward Gibbs sampling requires that an often difficult numerical integration be performed. West, Müller, and Escobar

---

Radford M. Neal is Associate Professor, Department of Statistics and Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (E-mail: [radford@stat.utoronto.ca](mailto:radford@stat.utoronto.ca); Web: <http://www.cs.utoronto.ca/~radford/>).

©2000 American Statistical Association, Institute of Mathematical Statistics,  
and Interface Foundation of North America  
*Journal of Computational and Graphical Statistics*, Volume 9, Number 2, Pages 249–265

(1994) used a Monte Carlo approximation to this integral, but the error from using such an approximation is likely to be large in many contexts.

MacEachern and Müller (1998) devised an exact approach for handling non-conjugate priors that uses a mapping from a set of auxiliary parameters to the set of parameters currently in use. Their “no gaps” and “complete” algorithms based on this approach are widely applicable, but somewhat inefficient. Walker and Damien (1998) applied a rather different auxiliary variable method to some Dirichlet process mixture models, but their method appears to be unsuitable for general use, as it again requires the computation of a difficult integral.

In this article, I review this past work and present two new approaches to Markov chain sampling. A very simple method for handling non-conjugate priors is to use Metropolis–Hastings updates with the conditional prior as the proposal distribution. A variation of this method may sometimes sample more efficiently, particularly when combined with a partial form of Gibbs sampling. Another class of methods uses Gibbs sampling in a space with auxiliary parameters. The simplest method of this type is very similar to the “no gaps” algorithm of MacEachern and Müller, but is more efficient. This approach also yields an algorithm that resembles use of a Monte Carlo approximation to the necessary integrals, but which does not suffer from any approximation error.

I conclude with a demonstration of the methods on a simple problem, which confirms that the new algorithms improve on the previous “no gaps” algorithm. Which of the several new algorithms introduced is best likely depends on the model and dataset to which they are applied. Further experience with these algorithms in a variety of contexts will be required to assess their relative merits.

## 2. DIRICHLET PROCESS MIXTURE MODELS

Dirichlet process mixture models go back to Antoniak (1974) and Ferguson (1983). [Note: Dirichlet process mixture models are sometimes also called “mixture of Dirichlet process models,” apparently because of Antoniak’s (1974) characterization of their posterior distributions. Since models are not usually named for the properties of their posterior distributions, this terminology is avoided here.] These models have recently been developed as practical methods by Escobar and West (1995), MacEachern and Müller (1998), and others.

The basic model applies to data  $y_1, \dots, y_n$  which we regard as part of an indefinite exchangeable sequence, or equivalently, as being independently drawn from some unknown distribution. The  $y_i$  may be multivariate, with components that may be real-valued or categorical. We model the distribution from which the  $y_i$  are drawn as a mixture of distributions of the form  $F(\theta)$ , with the mixing distribution over  $\theta$  being  $G$ . We let the prior for this mixing distribution be a Dirichlet process (Ferguson 1973), with concentration parameter  $\alpha$  and base distribution  $G_0$  (i.e., with base measure  $\alpha G_0$ ). This gives the following model:

$$\begin{aligned} y_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim \text{DP}(G_0, \alpha). \end{aligned} \tag{2.1}$$

Here, “ $X \sim S$ ” means “ $X$  has the distribution  $S$ ”, so the right side is a specification of

a distribution (e.g.,  $N(\mu, \sigma^2)$ ), not of a density function. DP is the Dirichlet process, a distribution over distributions. Here and later, the obvious independence properties (e.g., given the  $\theta_i$ , the  $y_i$  are independent of each other and of  $G$ ) are silently assumed.

Often, the distributions  $F$  and  $G_0$  will depend on additional hyperparameters not mentioned above, which, along with  $\alpha$ , may be given priors at a higher level, as illustrated, for example, by Escobar and West (1998). The computational methods discussed in this article extend easily to these more complex models, as briefly discussed in Section 7.

Since realizations of the Dirichlet process are discrete with probability one, these models can be viewed as countably infinite mixtures, as pointed out by Ferguson (1983). This is also apparent when we integrate over  $G$  in model (2.1), to obtain a representation of the prior distribution of the  $\theta_i$  in terms of successive conditional distributions of the following form (Blackwell and MacQueen 1973):

$$\theta_i \mid \theta_1, \dots, \theta_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) + \frac{\alpha}{i-1+\alpha} G_0. \tag{2.2}$$

Here,  $\delta(\theta)$  is the distribution concentrated at the single point  $\theta$ . Notation of the form  $pR + (1-p)S$ , where  $R$  and  $S$  are distributions, represents the distribution that is the mixture of  $R$  and  $S$ , with proportions  $p$  and  $1-p$ , respectively.

Equivalent models can also be obtained by taking the limit as  $K$  goes to infinity of finite mixture models with  $K$  components having the following form:

$$\begin{aligned} y_i \mid c_i, \phi &\sim F(\phi_{c_i}) \\ c_i \mid \mathbf{p} &\sim \text{Discrete}(p_1, \dots, p_K) \\ \phi_c &\sim G_0 \\ \mathbf{p} &\sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K). \end{aligned} \tag{2.3}$$

Here,  $c_i$  indicates which ‘‘latent class’’ is associated with observation  $y_i$ , with the numbering of the  $c_i$  being of no significance. For each class,  $c$ , the parameters  $\phi_c$  determine the distribution of observations from that class; the collection of all such  $\phi_c$  is denoted by  $\phi$ . The mixing proportions for the classes,  $\mathbf{p} = (p_1, \dots, p_K)$ , are given a symmetric Dirichlet prior, with concentration parameter written as  $\alpha/K$ , so that it approaches zero as  $K$  goes to infinity.

By integrating over the mixing proportions,  $\mathbf{p}$ , we can write the prior for the  $c_i$  as the product of conditional probabilities of the following form:

$$\begin{aligned} P(c_i = c \mid c_1, \dots, c_{i-1}) \\ = P(c_1, \dots, c_{i-1}, c_i = c) / P(c_1, \dots, c_{i-1}) \end{aligned} \tag{2.4}$$

$$= \frac{\int p_{c_1} \dots p_{c_{i-1}} p_c \Gamma(\alpha) \Gamma(\alpha/K)^{-K} p_1^{(\alpha/K)-1} \dots p_K^{(\alpha/K)-1} d\mathbf{p}}{\int p_{c_1} \dots p_{c_{i-1}} \Gamma(\alpha) \Gamma(\alpha/K)^{-K} p_1^{(\alpha/K)-1} \dots p_K^{(\alpha/K)-1} d\mathbf{p}} \tag{2.5}$$

$$= \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha}, \tag{2.6}$$

where  $n_{i,c}$  is the number of  $c_j$  for  $j < i$  that are equal to  $c$ .

If we now let  $K$  go to infinity, the conditional probabilities in Equation (2.6), which define the prior for the  $c_i$ , reach the following limits:

$$\begin{aligned}
P(c_i = c \mid c_1, \dots, c_{i-1}) &\rightarrow \frac{n_{i,c}}{i - 1 + \alpha} \\
P(c_i \neq c_j \text{ for all } j < i \mid c_1, \dots, c_{i-1}) &\rightarrow \frac{\alpha}{i - 1 + \alpha}.
\end{aligned}
\tag{2.7}$$

[Note: Some readers may be disturbed by the failure of countable additivity for these limiting probabilities, in which  $P(c_i \neq c_j \text{ for all } j < i) > 0$  even though  $P(c_i = c) = 0$  for any specific  $c$  that is not equal to some  $c_j$  with  $j < i$ . However, the limiting distribution of the observable quantities (the  $y_i$ ), and the limiting forms of the algorithms based on this model, are both well defined as  $K$  goes to infinity.]

Since the  $c_i$  are significant only in so far as they are or are not equal to other  $c_j$ , the above probabilities are all that are needed to define the model. If we now let  $\theta_i = \phi_{c_i}$  we can see that the limit of model (2.3) as  $K \rightarrow \infty$  is equivalent to the Dirichlet process mixture model (2.1), due to the correspondence between the conditional probabilities for the  $\theta_i$  in Equation (2.2) and those implied by (2.7).

I have previously used this limiting process to define a model which (unknown to me at the time) is equivalent to a Dirichlet process mixture (Neal 1992). This view is useful in deriving algorithms for sampling from the posterior distribution for Dirichlet process mixture models. Conversely, an algorithm for Dirichlet process mixture models will usually have a counterpart for finite mixture models. This is the case for the algorithms discussed in this article, though I do not give details of the algorithms for finite mixtures.

Yet another way of formulating the equivalent of a Dirichlet process mixture model is in terms of the prior probability that two observations come from the same mixture component (equal to  $1/(1+\alpha)$  in the models above). This approach has been used by Anderson (1990, chap. 3) in formulating a model for use as a psychological theory of human category learning.

### 3. GIBBS SAMPLING WHEN CONJUGATE PRIORS ARE USED

Exact computation of posterior expectations for a Dirichlet process mixture model is infeasible when there are more than a few observations. However, such expectations can be estimated using Monte Carlo methods. For example, suppose we have a sample of  $T$  points from the posterior distribution for  $\theta = (\theta_1, \dots, \theta_n)$ , with the  $t$ th such point being  $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_n^{(t)})$ . Then using Equation (2.2), the predictive distribution for a new observation,  $y_{n+1}$ , can be estimated by  $(1/T) \sum_{t=1}^T F(\theta_{n+1}^{(t)})$ , where  $\theta_{n+1}^{(t)}$  is drawn from the distribution  $(n+\alpha)^{-1} \sum_{i=1}^n \delta(\theta_i^{(t)}) + \alpha(n+\alpha)^{-1} G_0$ .

We can sample from the posterior distribution of  $\theta = (\theta_1, \dots, \theta_n)$  by simulating a Markov chain that has the posterior as its equilibrium distribution. The simplest such methods are based on Gibbs sampling, which when conjugate priors are used can be done in three ways.

The most direct approach to sampling for model (2.1) is to repeatedly draw values for each  $\theta_i$  from its conditional distribution given both the data and the  $\theta_j$  for  $j \neq i$  (written as  $\theta_{-i}$ ). This conditional distribution is obtained by combining the likelihood for  $\theta_i$  that results from  $y_i$  having distribution  $F(\theta_i)$ , which will be written as  $F(y_i, \theta_i)$ ,

and the prior conditional on  $\theta_{-i}$ , which is

$$\theta_i \mid \theta_{-i} \sim \frac{1}{n-1+\alpha} \sum_{j \neq i} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} G_0. \tag{3.1}$$

This conditional prior can be derived from Equation (2.2) by imagining that  $i$  is the last of the  $n$  observations, as we may, since the observations are exchangeable. When combined with the likelihood, this yields the following conditional distribution for use in Gibbs sampling:

$$\theta_i \mid \theta_{-i}, y_i \sim \sum_{j \neq i} q_{i,j} \delta(\theta_j) + r_i H_i. \tag{3.2}$$

Here,  $H_i$  is the posterior distribution for  $\theta$  based on the prior  $G_0$  and the single observation  $y_i$ , with likelihood  $F(y_i, \theta)$ . The values of the  $q_{i,j}$  and of  $r_i$  are defined by

$$q_{i,j} = b F(y_i, \theta_j) \tag{3.3}$$

$$r_i = b \alpha \int F(y_i, \theta) dG_0(\theta), \tag{3.4}$$

where  $b$  is such that  $\sum_{j \neq i} q_{i,j} + r_i = 1$ . For this Gibbs sampling method to be feasible, computing the integral defining  $r_i$  and sampling from  $H_i$  must be feasible operations. This will generally be so when  $G_0$  is the conjugate prior for the likelihood given by  $F$ .

We may summarize this method as follows:

**Algorithm 1.** Let the state of the Markov chain consist of  $\theta = (\theta_1, \dots, \theta_n)$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ : Draw a new value from  $\theta_i \mid \theta_{-i}, y_i$  as defined by Equation (3.2).

This algorithm was used by Escobar (1994) and by Escobar and West (1995). It produces an ergodic Markov chain, but convergence to the posterior distribution may be rather slow, and sampling thereafter may be inefficient. The problem is that there are often groups of observations that with high probability are associated with the same  $\theta$ . Since the algorithm cannot change the  $\theta$  for more than one observation simultaneously, a change to the  $\theta$  values for observations in such a group can occur only rarely, as such a change requires passage through a low-probability intermediate state in which observations in the group do not all have the same  $\theta$  value.

This problem is avoided if Gibbs sampling is instead applied to the model formulated as in (2.3), with the mixing proportions,  $\mathbf{p}$ , integrated out. When  $K$  is finite, each Gibbs sampling scan consists of picking a new value for each  $c_i$  from its conditional distribution given  $y_i$ , the  $\phi_c$ , and the  $c_j$  for  $j \neq i$  (written as  $c_{-i}$ ), and then picking a new value for each  $\phi_c$  from its conditional distribution given the  $y_i$  for which  $c_i = c$ . The required conditional probabilities for  $c_i$  can easily be computed:

$$P(c_i = c \mid c_{-i}, y_i, \phi) = b F(y_i, \phi_c) \frac{n_{-i,c} + \alpha/K}{n-1+\alpha}, \tag{3.5}$$

where  $n_{-i,c}$  is the number of  $c_j$  for  $j \neq i$  that are equal to  $c$ , and  $b$  is the appropriate normalizing constant. This expression is found by multiplying the likelihood,  $F(y_i, \phi_c)$ ,

by the conditional prior, which is derived from Equation (2.6) by imagining that  $i$  is the last observation. (Note that the denominator  $n-1+\alpha$  could be absorbed into  $b$ , but here and later it is retained for clarity.) The conditional distribution for  $\phi_c$  will generally be easy to sample from when the priors used are conjugate, and even when Gibbs sampling for  $\phi_c$  is difficult, one may simply substitute some other update that leaves the required distribution invariant. Note that when a new value is chosen for  $\phi_c$ , the values of  $\theta_i = \phi_{c_i}$  will change simultaneously for all observations associated with component  $c$ .

When  $K$  goes to infinity, we cannot, of course, explicitly represent the infinite number of  $\phi_c$ . We instead represent, and do Gibbs sampling for, only those  $\phi_c$  that are currently associated with some observation. Gibbs sampling for the  $c_i$  is based on the following conditional probabilities (with  $\phi$  here being the set of  $\phi_c$  currently associated with at least one observation):

$$\begin{aligned} \text{If } c = c_j \text{ for some } j \neq i: P(c_i = c \mid c_{-i}, y_i, \phi) &= b \frac{n_{-i,c}}{n-1+\alpha} F(y_i, \phi_c) \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i, \phi) &= b \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi) dG_0(\phi). \end{aligned} \quad (3.6)$$

Here,  $b$  is the appropriate normalizing constant that makes the above probabilities sum to one. The numerical values of the  $c_i$  are arbitrary, as long as they faithfully represent whether or not  $c_i = c_j$ —that is, the  $c_i$  are important only in that they determine what has been called the “configuration” in which the data items are grouped in accordance with shared values for  $\theta$ . The numerical values for the  $c_i$  may therefore be chosen for programming convenience, or to facilitate the display of mixture components in some desired order. When Gibbs sampling for  $c_i$  chooses a value not equal to any other  $c_j$ , a value for  $\phi_{c_i}$  is chosen from  $H_i$ , the posterior distribution based on the prior  $G_0$  and the single observation  $y_i$ .

We can summarize this second Gibbs sampling method as follows:

**Algorithm 2.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$  and  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ : If the present value of  $c_i$  is associated with no other observation (i.e.,  $n_{-i,c_i} = 0$ ), remove  $\phi_{c_i}$  from the state. Draw a new value for  $c_i$  from  $c_i \mid c_{-i}, y_i, \phi$  as defined by Equation (3.6). If the new  $c_i$  is not associated with any other observation, draw a value for  $\phi_{c_i}$  from  $H_i$  and add it to the state.
- For all  $c \in \{c_1, \dots, c_n\}$ : Draw a new value from  $\phi_c \mid$  all  $y_i$  for which  $c_i = c$ —that is, from the posterior distribution based on the prior  $G_0$  and all the data points currently associated with latent class  $c$ .

This is essentially the method used by Bush and MacEachern (1996) and later by West, Müller, and Escobar (1994). As was the case for the first Gibbs sampling method, this approach is feasible if we can compute  $\int F(y_i, \phi) dG_0(\phi)$  and sample from  $H_i$ , as will generally be the case when  $G_0$  is the conjugate prior.

Finally, in a conjugate context, we can often integrate analytically over the  $\phi_c$ , eliminating them from the algorithm. The state of the Markov chain then consists only of

the  $c_i$ , which we update by Gibbs sampling using the following conditional probabilities:

$$\begin{aligned}
 \text{If } c = c_j \text{ for some } j \neq i: P(c_i = c \mid c_{-i}, y_i) &= b \frac{n_{-i,c}}{n-1+\alpha} \int F(y_i, \phi) dH_{-i,c}(\phi) \\
 P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i) &= b \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi) dG_0(\phi).
 \end{aligned}
 \tag{3.7}$$

Here,  $H_{-i,c}$  is the posterior distribution of  $\phi$  based on the prior  $G_0$  and all observations  $y_j$  for which  $j \neq i$  and  $c_j = c$ .

This third Gibbs sampling method can be summarized as follows:

**Algorithm 3.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ : Draw a new value from  $c_i \mid c_{-i}, y_i$  as defined by Equation (3.7).

This algorithm is presented by MacEachern (1994) for mixtures of normals and by myself (Neal 1992) for models of categorical data.

### 4. EXISTING METHODS FOR HANDLING NON-CONJUGATE PRIORS

Algorithms 1 to 3 cannot easily be applied to models where  $G_0$  is not the conjugate prior for  $F$ , as the integrals in Equations (3.4), (3.6), and (3.7) will usually not be analytically tractable. Sampling from  $H_i$  may also be hard when the prior is not conjugate.

West, Müller, and Escobar (1994) suggested using either numerical quadrature or a Monte Carlo approximation to evaluate the required integral. If  $\int F(y_i, \phi) dG_0(\phi)$  is approximated by an average over  $m$  values for  $\phi$  drawn from  $G_0$ , one could also approximate a draw from  $H_i$ , if required, by drawing from among these  $m$  points with probabilities proportional to their likelihoods, given by  $F(y_i, \phi)$ . Though their article is not explicit, it appears that West, Müller, and Escobar’s non-conjugate example uses this approach with  $m = 1$  (see MacEachern and Müller 1998).

Unfortunately, this approach is potentially quite inaccurate. Often,  $H_i$ , the posterior based on  $y_i$  alone, will be considerably more concentrated than the prior,  $G_0$ , particularly when  $y_i$  is multidimensional. If a small to moderate number of points are drawn from  $G_0$ , it may be that none are typical of  $H_i$ . Consequently, the probability of choosing  $c_i$  to be a new component can be much lower than it would be if the exact probabilities of Equation (3.6) were used. The consequence of this is not just slower convergence, since on the rare occasions when  $c_i$  is in fact set to a new component, with an appropriate  $\phi$  typical of  $H_i$ , this new component is likely to be discarded in the very next Gibbs sampling iteration, leading to the wrong stationary distribution. This problem shows that the usual Gibbs sampling procedure of forgetting the current value of a variable before sampling from its conditional distribution will have to be modified in any valid scheme that uses values for  $\phi$  drawn from  $G_0$ .

MacEachern and Müller (1998) presented a framework that does allow auxiliary values for  $\phi$  drawn from  $G_0$  to be used to define a valid Markov chain sampler. I will



explain their idea as an extension of Algorithm 2 of Section 3. There, the numerical values of the  $c_i$  were regarded as significant only in so far as they indicate which observations are associated with the same component. MacEachern and Müller considered more specific schemes for assigning distributions to the  $c_i$ , which serve to map from a collection of values for  $\phi_c$  to values for the  $\theta_i$ . Many such schemes will produce the same distribution for the  $\theta_i$ , but lead to different sampling algorithms.

The “no gaps” algorithm of MacEachern and Müller arises when the  $c_i$  for  $i = 1, \dots, n$  are required to cover the set of integers from 1 to  $k$ , with  $k$  being the number of distinct  $c_i$ , but are not otherwise constrained. By considering Gibbs sampling in this representation, they derive the following algorithm:

**Algorithm 4.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$  and  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ : Let  $k^-$  be the number of distinct  $c_j$  for  $j \neq i$ , and let these  $c_j$  have values in  $\{1, \dots, k^-\}$ . If  $c_i \neq c_j$  for all  $j \neq i$ , then with probability  $k^- / (k^- + 1)$  do nothing, leaving  $c_i$  unchanged. Otherwise, label  $c_i$  as  $k^- + 1$  if  $c_i \neq c_j$  for all  $j \neq i$ , or draw a value for  $\phi_{k^-+1}$  from  $G_0$  if  $c_i = c_j$  for some  $j \neq i$ . Then draw a new value for  $c_i$  from  $\{1, \dots, k^- + 1\}$  using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi_1, \dots, \phi_{k^-+1}) = \begin{cases} b n_{-i,c} F(y_i, \phi_c) & \text{if } 1 \leq c \leq k^- \\ b [\alpha / (k^- + 1)] F(y_i, \phi_c) & \text{if } c = k^- + 1 \end{cases}$$

where  $b$  is the appropriate normalizing constant. Change the state to contain only those  $\phi_c$  that are now associated with an observation.

- For all  $c \in \{c_1, \dots, c_n\}$ : Draw a new value from  $\phi_c \mid y_i$  such that  $c_i = c$ , or perform some other update to  $\phi_c$  that leaves this distribution invariant.

This algorithm can be applied to any model for which we can sample from  $G_0$  and compute  $F(y_i, \theta)$ , regardless of whether  $G_0$  is the conjugate prior for  $F$ . However, there is a puzzling inefficiency in the algorithm’s mechanism for setting  $c_i$  to a value different from all other  $c_j$ —that is, for assigning an observation to a newly created mixture component. The probability of such a change is reduced from what one might expect by a factor of  $k^- + 1$ , with a corresponding reduction in the probability of the opposite change. As will be seen in Section 6, a similar algorithm without this inefficiency is possible.

MacEachern and Müller also developed an algorithm based on a “complete” scheme for mapping from the  $\phi_c$  to the  $\theta_i$ . It requires maintaining  $n$  values for  $\phi$ , which may be inefficient when  $k \ll n$ . The approach that will be presented in Section 6 allows more control over the number of auxiliary parameter values used.

Another approach to handling non-conjugate priors was devised by Walker and Damien (1998). Their method avoids the integrals needed for Gibbs sampling, but requires instead that the probability under  $G_0$  of the set of all  $\theta$  for which  $F(y_i, \theta) > u$  be computable, and that one be able to sample from  $G_0$  restricted to this set. Although these operations are feasible for some models, they will in general be quite difficult, especially when  $\theta$  is multidimensional.

Finally, Green and Richardson (in press) developed a Markov chain sampling method based on splitting and merging components that is applicable to non-conjugate models. Their method is considerably more complex than the others discussed in this article, since it attempts to solve the more difficult problem of obtaining good performance in situations where the other methods tend to become trapped in local modes that are not easily escaped with incremental changes. Discussion of this issue is beyond the scope of this article.

### 5. METROPOLIS–HASTINGS UPDATES AND PARTIAL GIBBS SAMPLING

Perhaps the simplest way of handling non-conjugate priors is by using the Metropolis–Hastings algorithm (Hastings 1970) to update the  $c_i$ , using the conditional prior as the proposal distribution.

Recall that the Metropolis–Hastings algorithm for sampling from a distribution for  $x$  with probabilities  $\pi(x)$ , using a proposal distribution  $g(x^*|x)$ , updates the state  $x$  as follows:

Draw a candidate state,  $x^*$ , according to the probabilities  $g(x^*|x)$ . Compute the acceptance probability

$$a(x^*, x) = \min \left[ 1, \frac{g(x|x^*) \pi(x^*)}{g(x^*|x) \pi(x)} \right]. \tag{5.1}$$

With probability  $a(x^*, x)$ , set the new state,  $x'$ , to  $x^*$ . Otherwise, let  $x'$  be the same as  $x$ .

This update from  $x$  to  $x'$  leaves  $\pi$  invariant. When  $x$  is multidimensional, proposal distributions that change only one component of  $x$  are often used. Updates based on several such proposals, along with updates of other types, can be combined in order to construct an ergodic Markov chain that will converge to  $\pi$ .

This approach can be applied to model (2.3) for finite  $K$ , with the  $p_c$  integrated out, using Metropolis–Hastings updates for each  $c_i$  in turn, along with Gibbs sampling or other updates for the  $\phi_c$ . When updating just  $c_i$ , we can ignore those factors in the posterior distribution that do not involve  $c_i$ . What remains is the product of the likelihood for observation  $i$ ,  $F(y_i, \phi_{c_i})$ , and the conditional prior for  $c_i$  given the other  $c_j$ , which is

$$P(c_i = c \mid c_{-i}) = \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha}, \tag{5.2}$$

where, as before,  $n_{-i,c}$  is the number of  $c_j$  for  $j \neq i$  that are equal to  $c$ . This can be obtained from Equation (2.6) by imagining that  $i$  is the last observation. If we now choose to use this conditional prior for  $c_i$  as the proposal distribution, we find that this factor cancels when computing the acceptance probability of Equation (5.1), leaving

$$a(c_i^*, c_i) = \min \left[ 1, \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right]. \tag{5.3}$$

This approach continues to work as we let  $K \rightarrow \infty$  in order to produce an algorithm

for a Dirichlet process mixture model. The conditional prior for  $c_i$  becomes

$$\begin{aligned} \text{If } c = c_j \text{ for some } j: P(c_i = c \mid c_{-i}) &= \frac{n_{-i,c}}{n-1+\alpha} \\ P(c_i \neq c_j \text{ for all } j \mid c_{-i}) &= \frac{\alpha}{n-1+\alpha}. \end{aligned} \tag{5.4}$$

If we use this as the proposal distribution for an update to  $c_i$ , we will need to draw an associated value for  $\phi$  from  $G_0$  if the candidate,  $c_i^*$ , is not in  $\{c_1, \dots, c_n\}$ . Note that if the current  $c_i$  is not equal to any other  $c_j$ , the probability of choosing  $c_i^*$  to be the same as  $c_i$  is zero—that is, when  $c_i^*$  is chosen to be different from the other  $c_j$  it will always be a new component, not the current  $c_i$ , even when that also differs from the other  $c_j$ . (The method would be valid even if a new component were not created in this situation, but this is the behavior obtained by taking the  $K \rightarrow \infty$  limit of the algorithm for finite  $K$ .)

We might wish to perform more than one such Metropolis–Hastings update for each of the  $c_i$ . With this elaboration, the algorithm can be summarized as follows:

**Algorithm 5.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$  and  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ , repeat the following update of  $c_i$   $R$  times: Draw a candidate,  $c_i^*$ , from the conditional prior for  $c_i$  given by Equation (5.4). If this  $c_i^*$  is not in  $\{c_1, \dots, c_n\}$ , choose a value for  $\phi_{c_i^*}$  from  $G_0$ . Compute the acceptance probability,  $a(c_i^*, c_i)$ , as in Equation (5.3), and set the new value of  $c_i$  to  $c_i^*$  with this probability. Otherwise let the new value of  $c_i$  be the same as the old value.
- For all  $c \in \{c_1, \dots, c_n\}$ : Draw a new value from  $\phi_c \mid y_i$  such that  $c_i = c$ , or perform some other update to  $\phi_c$  that leaves this distribution invariant.

If  $R$  is greater than one, it is possible to save computation time by reusing values of  $F$  that were previously computed. An evaluation of  $F$  can also be omitted when  $c_i^*$  turns out to be the same as  $c_i$ . The number of evaluations of  $F$  required to update one  $c_i$  is thus no more than  $R+1$ . For comparison, the number of evaluations of  $F$  needed to update one  $c_i$  for Gibbs sampling and the “no gaps” algorithm is approximately equal to one plus the number of distinct  $c_j$  for  $j \neq i$ .

If the updates for the  $\phi_c$  in the last step of Algorithm 5 are omitted, the algorithm can be rephrased in terms of the  $\theta_i = \phi_{c_i}$ , with the following result:

**Algorithm 6.** Let the state of the Markov chain consist of  $\theta = (\theta_1, \dots, \theta_n)$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ , repeat the following update of  $\theta_i$   $R$  times: Draw a candidate,  $\theta_i^*$ , from the following distribution:

$$\frac{1}{n-1+\alpha} \sum_{j \neq i} \delta(\theta_j) + \frac{\alpha}{n-1+\alpha} G_0.$$

Compute the acceptance probability

$$a(\theta_i^*, \theta_i) = \min[1, F(y_i, \theta_i^*) / F(y_i, \theta_i)].$$

Set the new value of  $\theta_i$  to  $\theta_i^*$  with this probability; otherwise let the new value of  $\theta_i$  be the same as the old value.

This might have been justified directly as a Metropolis–Hastings algorithm, but the fact that the proposal distribution for  $\theta_i^*$  is a mixture of continuous and discrete distributions introduces conceptual, or at least notational, difficulties. Note that this algorithm suffers from the same problem of not being able to change several  $\theta_i$  simultaneously as was discussed for Algorithm 1.

The behavior of the Metropolis–Hastings methods (Algorithms 5 and 6) differs substantially from that of the corresponding Gibbs sampling methods (Algorithms 2 and 1) and the “no gaps” method (Algorithm 4). These other methods consider all mixture components when deciding on a new value for  $c_i$ , whereas the Metropolis–Hastings method is more likely to consider changing  $c_i$  to a component associated with many observations than to a component associated with few observations. Also, the probability that the Metropolis–Hastings method will consider changing  $c_i$  to a newly created component is proportional to  $\alpha$ . (Of course, the probability of actually making such a change depends on  $\alpha$  for all methods; here the issue is whether such a change is even considered.)

It is difficult to say which behavior is better. Algorithm 5 does appear to perform adequately in practice, but since small values of  $\alpha$  (around one) are often used, one might wonder whether an algorithm that could consider the creation of a new component more often might be more efficient.

We can produce such an algorithm by modifying the proposal distribution for updates to the  $c_i$ . In particular, whenever  $c_i = c_j$  for some  $j \neq i$ , we can propose changing  $c_i$  to a newly created component, with associated  $\phi$  drawn from  $G_0$ . In order to allow the reverse change, in which a component disappears, the proposal distribution for “singleton”  $c_i$  that are not equal to any  $c_j$  with  $j \neq i$  will be confined to those components that are associated with other observations, with probabilities proportional to  $n_{-i,c}$ . Note that when the current  $c_i$  is not a singleton, the probability of proposing a new component is a factor of  $(n-1+\alpha) / \alpha$  greater than the conditional prior, while when  $c_i$  is a singleton, the probability of proposing any existing component is a factor of  $(n-1+\alpha) / (n-1)$  greater than its conditional prior. The probability of accepting a proposal must be adjusted by the ratio of these factors.

On their own, these updates are sufficient to produce a Markov chain that is ergodic, as can be seen from the fact that there is a nonzero probability that a single scan of the data items will result in a state where every data item is associated with a different component. Such a chain would often sample inefficiently, however, since it can move an observation from one existing component to another only by passing through a possibly unlikely state in which that observation is a singleton. Such changes can be made more likely by combining these Metropolis–Hastings updates with partial Gibbs sampling updates, which are applied only to those observations that are not singletons, and which are allowed to change  $c_i$  for such an observation only to a component associated with some other observation. In other words, these updates perform Gibbs sampling for the posterior distribution conditional on the set of components that are associated with at least one observation remaining the same as at present. No difficult integrations are required for this partial Gibbs sampling operation.

Combining the modified Metropolis–Hasting updates, the partial Gibbs sampling up-

dates, and the usual updates to  $\phi_c$  for  $c \in \{c_1, \dots, c_n\}$  produces the following algorithm:

**Algorithm 7.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$  and  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ , update  $c_i$  as follows: If  $c_i$  is a not a singleton (i.e.,  $c_i = c_j$  for some  $j \neq i$ ), let  $c_i^*$  be a newly created component, with  $\phi_{c_i^*}$  drawn from  $G_0$ . Set the new  $c_i$  to this  $c_i^*$  with probability

$$a(c_i^*, c_i) = \min \left[ 1, \frac{\alpha}{n-1} \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right].$$

Otherwise, when  $c_i$  is a singleton, draw  $c_i^*$  from  $c_{-i}$ , choosing  $c_i^* = c$  with probability  $n_{-i,c} / (n-1)$ . Set the new  $c_i$  to this  $c_i^*$  with probability

$$a(c_i^*, c_i) = \min \left[ 1, \frac{n-1}{\alpha} \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})} \right].$$

If the new  $c_i$  is not set to  $c_i^*$ , it is the same as the old  $c_i$ .

- For  $i = 1, \dots, n$ : If  $c_i$  is a singleton (i.e.,  $c_i \neq c_j$  for all  $j \neq i$ ), do nothing. Otherwise, choose a new value for  $c_i$  from  $\{c_1, \dots, c_n\}$  using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi, c_i \in \{c_1, \dots, c_n\}) = b \frac{n_{-i,c}}{n-1} F(y_i, \phi_c),$$

where  $b$  is the appropriate normalizing constant.

- For all  $c \in \{c_1, \dots, c_n\}$ : Draw a new value from  $\phi_c \mid y_i$  such that  $c_i = c$ , or perform some other update to  $\phi_c$  that leaves this distribution invariant.

## 6. GIBBS SAMPLING WITH AUXILIARY PARAMETERS

In this section, I show how models with non-conjugate priors can be handled by applying Gibbs sampling to a state that has been extended by the addition of auxiliary parameters. This approach is similar to that of MacEachern and Müller (1998), but differs in that the auxiliary parameters are regarded as existing only temporarily; this allows more flexibility in constructing algorithms.

The basic idea of auxiliary variable methods is that we can sample from a distribution  $\pi_x$  for  $x$  by sampling from some distribution  $\pi_{xy}$  for  $(x, y)$ , with respect to which the marginal distribution of  $x$  is  $\pi_x$ . We can extend this idea to accommodate auxiliary variables that are created and discarded during the Markov chain simulation. The permanent state of the Markov chain will be  $x$ , but a variable  $y$  will be introduced temporarily during an update of the following form:

1. Draw a value for  $y$  from its conditional distribution given  $x$ , as defined by  $\pi_{xy}$ .
2. Perform some update of  $(x, y)$  that leaves  $\pi_{xy}$  invariant.
3. Discard  $y$ , leaving only the value of  $x$ .

It is easy to see that this update for  $x$  will leave  $\pi_x$  invariant as long as  $\pi_x$  is the marginal distribution of  $x$  under  $\pi_{xy}$ . We can combine several such updates, which may involve different auxiliary variables, along with other updates that leave  $\pi_x$  invariant, to construct a Markov chain that will converge to  $\pi_x$ .

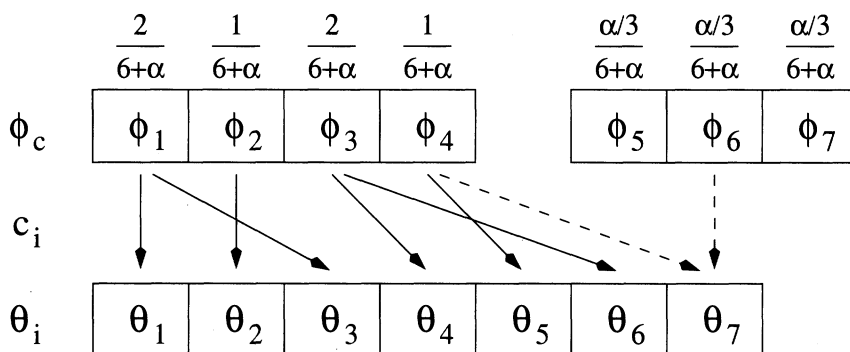


Figure 1. Representing the conditional prior distribution for a new observation using auxiliary parameters. The component for the new observation is chosen from among the four components associated with other observations plus three possible new components, with parameters,  $\phi_5, \phi_6, \phi_7$ , drawn independently from  $G_0$ . The probabilities used for this choice are shown at the top. The dashed arrows illustrate the possibilities of choosing an existing component, or a new component that uses one of the auxiliary parameters.

We can use this technique to update the  $c_i$  for a Dirichlet process mixture model without having to integrate with respect  $G_0$ . The permanent state of the Markov chain will consist of the  $c_i$  and the  $\phi_c$ , as in Algorithm 2, but when  $c_i$  is updated, we will introduce temporary auxiliary variables that represent possible values for the parameters of components that are not associated with any other observations. We then update  $c_i$  by Gibbs sampling with respect to the distribution that includes these auxiliary parameters.

Since the observations  $y_i$  are exchangeable, and the component labels  $c_i$  are arbitrary, we can assume that we are updating  $c_i$  for the last observation, and that the  $c_j$  for other observations have values in the set  $\{1, \dots, k^-\}$ , where  $k^-$  is the number of distinct  $c_j$  for  $j \neq i$ . We can now visualize the conditional prior distribution for  $c_i$  given the other  $c_j$  in terms of  $m$  auxiliary components and their associated parameters. The probability of  $c_i$  being equal to a  $c$  in  $\{1, \dots, k^-\}$  will be  $n_{-i,c}/(n-1+\alpha)$ , where  $n_{-i,c}$  is the number of times  $c$  occurs among the  $c_j$  for  $j \neq i$ . The probability of  $c_i$  having some other value will be  $\alpha/(n-1+\alpha)$ , which we will split equally among the  $m$  auxiliary components we have introduced. Figure 1 illustrates this setup for  $m = 3$ .

This representation of the prior gives rise to a corresponding representation of the posterior, which also includes these auxiliary parameters. The first step in using this representation to update  $c_i$  is to sample from the conditional distribution of these auxiliary parameters given the current value of  $c_i$  and the rest of the state. If  $c_i = c_j$  for some  $j \neq i$ , the auxiliary parameters have no connection with the rest of the state, or the observations, and are simply drawn independently from  $G_0$ . If  $c_i \neq c_j$  for all  $j \neq i$  (i.e.,  $c_i$  is a singleton), then it must be associated with one of the  $m$  auxiliary parameters. Technically, we should select which auxiliary parameter it is associated with randomly, but since it turns out to make no difference, we can just let  $c_i$  be the first of these auxiliary components. The corresponding value for  $\phi$  must, of course, be equal to the existing  $\phi_{c_i}$ . The  $\phi$  values for the other auxiliary components (if any, there are none if  $m = 1$ ) are again drawn independently from  $G_0$ .

We now perform a Gibbs sampling update for  $c_i$  in this representation of the posterior distribution. Since  $c_i$  must be either one of the components associated with other

observations or one of the auxiliary components that were introduced, we can easily do Gibbs sampling by evaluating the relative probabilities of these possibilities. Once a new value for  $c_i$  has been chosen, we discard all  $\phi$  values that are not now associated with an observation.

This algorithm can be summarized as follows:

**Algorithm 8.** Let the state of the Markov chain consist of  $c = (c_1, \dots, c_n)$  and  $\phi = (\phi_c : c \in \{c_1, \dots, c_n\})$ . Repeatedly sample as follows:

- For  $i = 1, \dots, n$ : Let  $k^-$  be the number of distinct  $c_j$  for  $j \neq i$ , and let  $h = k^- + m$ . Label these  $c_j$  with values in  $\{1, \dots, k^-\}$ . If  $c_i = c_j$  for some  $j \neq i$ , draw values independently from  $G_0$  for those  $\phi_c$  for which  $k^- < c \leq h$ . If  $c_i \neq c_j$  for all  $j \neq i$ , let  $c_i$  have the label  $k^- + 1$ , and draw values independently from  $G_0$  for those  $\phi_c$  for which  $k^- + 1 < c \leq h$ . Draw a new value for  $c_i$  from  $\{1, \dots, h\}$  using the following probabilities:

$$P(c_i = c \mid c_{-i}, y_i, \phi_1, \dots, \phi_h) = \begin{cases} b \frac{n_{-i,c}}{n-1+\alpha} F(y_i, \phi_c) & \text{for } 1 \leq c \leq k^- \\ b \frac{\alpha/m}{n-1+\alpha} F(y_i, \phi_c) & \text{for } k^- < c \leq h \end{cases},$$

where  $n_{-i,c}$  is the number of  $c_j$  for  $j \neq i$  that are equal to  $c$ , and  $b$  is the appropriate normalizing constant. Change the state to contain only those  $\phi_c$  that are now associated with one or more observations.

- For all  $c \in \{c_1, \dots, c_n\}$ : Draw a new value from  $\phi_c \mid y_i$  such that  $c_i = c$ , or perform some other update to  $\phi_c$  that leaves this distribution invariant.

Note that the relabelings of the  $c_j$  above are conceptual; they may or may not require any actual computation, depending on the data structures used.

When  $m = 1$ , Algorithm 8 closely resembles Algorithm 4, the “no gaps” algorithm of MacEachern and Müller (1998). The difference is that the probability of changing  $c_i$  from a component shared with other observations to a new singleton component is approximately  $k^- + 1$  times greater with Algorithm 8, and the same is true for the reverse change. When  $\alpha$  is small, this seems to be a clear benefit, since the probabilities for other changes are affected only slightly.

In the other extreme, as  $m \rightarrow \infty$ , Algorithm 8 approaches the behavior of Algorithm 2, since the  $m$  (or  $m-1$ ) values for  $\phi_c$  drawn from  $G_0$  effectively produce a Monte Carlo approximation to the integral computed in Algorithm 2. However, the equilibrium distribution of the Markov chain defined by Algorithm 8 is exactly correct for any value of  $m$ , unlike the situation when a Monte Carlo approximation is used to implement Algorithm 2.

## 7. UPDATES FOR HYPERPARAMETERS

For many problems, it is necessary to extend the model to incorporate uncertainty regarding the value of  $\alpha$  or regarding the values of other hyperparameters that determine  $F$  and  $G_0$ . These hyperparameters can be included in the Markov chain simulation, as is briefly discussed here.

The conditional distribution of  $\alpha$  given the other parameters depends only on the number of distinct  $c_i$ . It can be updated by some Metropolis–Hastings method, or by methods discussed by Escobar and West (1995). Alternatively, one can eliminate  $\alpha$  from the model by integrating over it. As noted by MacEachern (1998), the moderate number of one-dimensional numerical integrations required for this can be done once, before the Markov chain simulation.

If  $F$  depends on hyperparameters  $\gamma$ , the conditional density for  $\gamma$  given the current  $\theta_i$  will be proportional to its prior density times the likelihood,  $\prod_{i=1}^n F(y_i, \theta_i, \gamma)$ . If  $G_0$  depends on hyperparameters  $\eta$ , the conditional density for  $\eta$  given the current  $c_i$  and  $\phi_c$  will be proportional to its prior density times  $\prod_c G_0(\phi_c, \eta)$ , where the product is over values of  $c$  that occur in  $\{c_1, \dots, c_n\}$ . Note that each such  $c$  occurs only once in this product, even if it is associated with more than one observation. The difficulty of performing Gibbs sampling or other updates for  $\gamma$  and  $\eta$  will depend on the detailed forms of these conditional distributions, but no issues special to Dirichlet process mixture models are involved.

One subtlety does arise when algorithms employing auxiliary  $\phi$  parameters are used. If  $\phi$  values not associated with any observation are retained in the state, the conditional distribution for  $\eta$  given the rest of the state will include factors of  $G_0(\phi, \eta)$  for these  $\phi$  as well as for the  $\phi$  values associated with observations. Since this will tend to slow convergence, it is desirable to discard all unused  $\phi$  values, regenerating them from  $G_0$  as needed, as is done for the algorithms in this article.

## 8. A DEMONSTRATION

I tested the performance of Algorithms 4 through 8 on the following data  $(y_1, \dots, y_9)$ :

$$-1.48, -1.40, -1.16, -1.08, -1.02, +0.14, +0.51, +0.53, +0.78$$

A Dirichlet process mixture model was used with the component distributions having the form  $F(\theta) = N(\theta, 0.1^2)$ , the prior being  $G_0 = N(0, 1)$ , and the Dirichlet process concentration parameter being  $\alpha = 1$ . Although  $G_0$  is in fact conjugate to  $F$ , the algorithms for non-conjugate priors were used. However, this conjugacy was exploited in Algorithms 4, 5, 7, and 8 in order to implement the Gibbs sampling step where a new value for  $\phi_c$  is drawn from its posterior distribution given the data associated with component  $c$ . If the prior used were not conjugate, this Gibbs sampling update might be more difficult, or might have to be replaced by a Metropolis update, or by some other update leaving the conditional distribution invariant.

A state from close to the posterior distribution was found by applying 100 iterations of Algorithm 5 with  $R = 5$ . This state was then used to initialize the Markov chain for each of the algorithms, which were all run for 20,000 subsequent iterations (one iteration being one application of the operations in the descriptions given earlier).

The performance of each algorithm was judged by the computation time per iteration and by the “autocorrelation time” for two quantities:  $k$ , the number of distinct  $c_i$ , and  $\theta_1$ , the parameter associated with  $y_1$ . The autocorrelation time for a quantity, defined as one plus twice the sum of the autocorrelations at lags one and up, is the factor by which the sample size is effectively reduced when estimating the expectation of that quantity, as compared to an estimate based on points drawn independently from the posterior



Table 1. Performance of the Algorithms Tested

|  | <i>Time per iteration<br/>in microseconds</i> | <i>Autocorrelation<br/>time for <math>k</math></i> | <i>Autocorrelation<br/>time for <math>\theta_1</math></i> |
|--|---|--|---|
| Alg. 4 ("no gaps")                       | 7.6   | 13.7   | 8.5   |
| Alg. 5 (Metropolis–Hastings, $R = 4$ )   | 8.6   | 8.1  | 10.2  |
| Alg. 6 (M–H, $R = 4$ , no $\phi$ update) | 8.3   | 19.4   | 64.1  |
| Alg. 7 (mod M–H & partial Gibbs)         | 8.0   | 6.9  | 5.3   |
| Alg. 8 (auxiliary Gibbs, $m = 1$ )       | 7.9   | 5.2  | 5.6   |
| Alg. 8 (auxiliary Gibbs, $m = 2$ )       | 8.8   | 3.7  | 4.7   |
| Alg. 8 ( $m = 30$ , approximates Alg. 2) | 38.0  | 2.0  | 2.8   |

distribution (see Ripley 1987, sec. 6.3). It was estimated using autocorrelation estimates from the 20,000 iterations.

The Metropolis–Hastings methods (Algorithms 5 and 6) were run with  $R$ , the number of updates for each  $c_i$ , set to 4. This makes the computation time per iteration approximately equal to that for the other methods tested. Gibbs sampling with auxiliary parameters (Algorithm 8) was tested with  $m = 1$  and  $m = 2$ . It was also run with  $m = 30$ , even though this is clearly too large, because with a large value of  $m$ , this algorithm approximates the behavior of Algorithm 2 (apart, of course, from computation time). This lets us see how much the autocorrelation times for the algorithms are increased over what is possible when the prior is conjugate.

The results are shown in Table 1. They confirm that Algorithm 8 with  $m = 1$  is superior to the "no gaps" method. Setting  $m = 2$  decreases autocorrelation times further, more than offsetting the slight increase in computation time per iteration. The simple Metropolis–Hastings method (Algorithm 5) performs about as well as the "no gaps" method. The combination of Metropolis–Hastings and partial Gibbs sampling of Algorithm 7 performs about as well as Algorithm 8 with  $m = 1$ . As expected, performance is much worse when updates for the  $\phi_c$  are omitted, as in Algorithm 6.

The results for Algorithm 8 with  $m = 30$  show that there is a cost to using algorithms that do not rely on the prior being conjugate, but this cost is small enough to be tolerable when a non-conjugate prior is a more realistic expression of prior beliefs. Note that if Algorithm 2 were implemented for this model using analytic integration, the time per iteration would be roughly the same as for Algorithm 8 with a small value of  $m$  (ie, about nine microseconds), while the autocorrelation times would be about the same as those shown for Algorithm 8 with  $m = 30$ .

Although Algorithm 8 with  $m = 1$  can be seen to essentially dominate the "no gaps" method (Algorithm 4), due to its greater probability of changes involving singletons, the varying characteristics of Algorithms 5, 7, and 8, with various settings of  $R$  and  $m$ , are such that each algorithm can be expected to outperform the others on at least some data sets. The relative performance of the methods may also be affected by other aspects of the model, such as whether updates are also done for hyperparameters. The methods tested here are implemented in my software for flexible Bayesian modeling (the above results were obtained using the version of 1998-09-01). This software is available from my Web page (<http://www.cs.utoronto.ca/~radford/>), and can be used to experiment with these algorithms in conjunction with various models and datasets.

## ACKNOWLEDGMENTS

I thank Steve MacEachern for helpful comments on the manuscript. This research was supported by the Natural Sciences and Engineering Research Council of Canada, and by the Institute for Robotics and Intelligent Systems.

[Received February 1999. Revised November 1999.]

## REFERENCES

- Anderson, J. R. (1990), *The Adaptive Character of Thought*, Hillsdale, NJ: Erlbaum.
- Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.
- Blackwell, D., and MacQueen, J. B. (1973), "Ferguson Distributions via Pólya urn Schemes," *The Annals of Statistics*, 1, 353–355.
- Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285.
- Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277.
- Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.
- (1998), "Computing Nonparametric Hierarchical Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey et al., New York: Springer-Verlag, pp. 1–22.
- Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, pp. 209–230.
- (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, eds. H. Rizvi and J. Rustagi, New York: Academic Press, pp. 287–303.
- Green, P. J., and Richardson, S. (in press) "Modelling Heterogeneity With and Without the Dirichlet Process," *Scandinavian Journal of Statistics*.
- Hastings, W. K. (1970) "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.
- MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate Style Dirichlet Process Prior," *Communications in Statistics: Simulation and Computation*, 23, 727–741.
- (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey et al., New York: Springer-Verlag, pp. 23–43.
- MacEachern, S. N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.
- Neal, R. M. (1992), "Bayesian Mixture Modeling," in *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, 1991*, eds. C. R. Smith, G. J. Erickson, and P. O. Neudorfer, Dordrecht: Kluwer Academic Publishers, pp. 197–211.
- Ripley, B. D. (1987), *Stochastic Simulation*, New York: Wiley.
- Walker, S., and Damien, P. (1998), "Sampling Methods for Bayesian Nonparametric Inference Involving Stochastic Processes," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey et al., New York: Springer-Verlag, pp. 243–254.
- West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty*, eds. P. R. Freeman and A. F. M. Smith, New York: Wiley, pp. 363–386.