Technical Report No. 9815, Department of Statistics, University of Toronto

Markov Chain Sampling Methods for Dirichlet Process Mixture Models

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
http://www.cs.utoronto.ca/∼radford/
radford@stat.utoronto.ca

1 September 1998

Abstract. Markov chain methods for sampling from the posterior distribution of a Dirichlet process mixture model are reviewed, and two new classes of methods are presented. 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.

1 Introduction

Modeling a distribution as a mixture of simpler distributions is useful both as a non-parametric 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

(1994) use 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) have devised an exact approach to handling non-conjugate priors that utilizes 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) apply 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 paper, 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.

2 Dirichlet process mixture models

Dirichlet process mixture models[1] go back to Antoniak (1974) and Ferguson (1983). The have recently been developed as practical methods by Escobar and West (1995), MacEachern and Müller (1998), and others.

The basic model applies to data $y_1, ..., y_n$ which we regard as exchangeable, 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 θ being G. We let the prior for this mixing distribution be a Dirichlet process (Ferguson 1973), with concentration parameter α and base distribution G_0 (ie, with base measure αG_0). This gives the following model:

$$\begin{aligned} y_i \mid \theta_i &\sim F(\theta_i) \\ \theta_i \mid G &\sim G \\ G &\sim D(G_0, \alpha) \end{aligned} \tag{1}$$

Often, F and G_0 will depend on additional hyperparameters not mentioned above, which, along with α, may be given priors at a higher level. The computational methods discussed in this paper extend easily to these more complex models, as briefly discussed in Section 7.

[1] 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.

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 (1), to obtain a representation of the prior distribution of the θ_i in terms of successive conditional distributions of the following form (Blackwell and MacQueen 1973):

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \quad \sim \quad \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta(\theta_j) \; + \; \frac{\alpha}{i-1+\alpha} G_0 \tag{2}$$

Here, $\delta(\theta)$ is the distribution concentrated at the single point θ.

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:

$$\begin{aligned} y_i \mid c, \phi \quad & \sim \quad F(\phi_{c_i}) \\ c_i \mid p \quad & \sim \quad \text{Discrete}\,(p_1, \ldots, p_K) \\ \phi_c \quad & \sim \quad G_0 \\ p_1, \ldots, p_K \quad & \sim \quad \text{Dirichlet}\,(\alpha/K, \ldots, \alpha/K) \end{aligned} \tag{3}$$

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 ϕ_c determine the distribution of observations from that class. The mixing proportions for the classes, p_c, are given a symmetric Dirichlet prior, with concentration parameter written as α/K, so that it approaches zero as K goes to infinity.

By integrating over the mixing proportions, p_c, we can write the prior for the c_i as the product of conditional probabilities of the following form:

$$P(c_i = c \mid c_1, \ldots, c_{i-1}) \quad = \quad \frac{n_{i,c} + \alpha/K}{i - 1 + \alpha} \tag{4}$$

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, we find that the conditional probabilities defining the prior for the c_i reach the following limits:[2]

$$\begin{aligned} P(c_i = c \mid c_1, \ldots, c_{i-1}) \quad & \rightarrow \quad \frac{n_{i,c}}{i - 1 + \alpha} \\ P(c_i \neq c_j \text{ for all } j < i \mid c_1, \ldots, c_{i-1}) \quad & \rightarrow \quad \frac{\alpha}{i - 1 + \alpha} \end{aligned} \tag{5}$$

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 (3) as $K \rightarrow \infty$ is equivalent to the Dirichlet process mixture model (1), due to the correspondence between the conditional probabilities for the θ_i in equation (2) and those implied by (5).

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

[2]Some readers may be disturbed by the failure of countable additivity for these limiting probabilities, but the limiting distribution of the observable quantities and the limiting forms of the algorithms based on this model are both well defined as K goes to infinity.

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 paper, though I do not give details of the algorithms for finite mixtures.

Yet another way of formulating a model equivalent to a Dirichlet process mixture 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, Chapter 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, 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 the points $\theta_{n+1}^{(t)}$ are drawn from the distribution $(n+\alpha)^{-1}\sum_{i=1}^{n}\delta(\theta_i^{(t)}) + \alpha(n+\alpha)^{-1}G_0$ (see equation (2)), where $\theta_1^{(t)},\ldots,\theta_n^{(t)}$ is the t'th point in a sample from the posterior distribution of the θ_i.

We can sample from the posterior distribution of $\theta_1,\ldots,\theta_n$ by simulating a Markov chain that has this 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 (1) is to repeatedly draw values for each θ_i from its conditional distribution given both the data and the θ_j for $j\neq i$ (written as θ_{-i}). This conditional distribution is obtained by combining the likelihood, written $F(y_i,\theta_i)$, and the prior conditional on θ_{-i}, which is

$$\theta_i \mid \theta_{-i} \;\sim\; \frac{1}{n-1+\alpha}\sum_{j\neq i}\delta(\theta_j) \;+\; \frac{\alpha}{n-1+\alpha}\,G_0 \tag{6}$$

This can be derived from equation (2) by imagining that i is the last observation, as we may, since the observations are exchangeable. When combined with the likelihood, this yields the following conditional distribution for use in Gibbs sampling:

$$\theta_i \mid \theta_{-i},\, y_i \;\sim\; \sum_{j\neq i} q_{i,j}\,\delta(\theta_j) \;+\; r_i\,H_i \tag{7}$$

Here, H_i is the posterior distribution for θ 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 as

$$q_{i,j} \;=\; b\,F(y_i,\theta_j) \tag{8}$$

$$r_i \;=\; b\,\alpha\int F(y_i,\theta)\,dG_0(\theta) \tag{9}$$

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 $\theta_1, \ldots, \theta_n$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Draw a new value from $\theta_i \mid \theta_{-i}, y_i$ as defined by equation (7).

This algorithm is 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 θ. Since the algorithm cannot change the θ for more than one observation simultaneously, changes to the θ values for observations in such a group can occur only rarely, as they require passage through a low-probability intermediate state in which observations in the group do not all have the same θ value.

This problem is avoided if Gibbs sampling is instead applied to the model formulated as in (3), with the mixing proportions, p_c, 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 ϕ_c, and the c_j for $j \neq i$ (written as c_{-i}), and then picking a new value for each ϕ_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:

$$P(c_i = c \mid c_{-i}, y_i, \phi) \quad = \quad b\, F(y_i, \phi_c)\, \frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha} \tag{10}$$

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. The last factor is derived from equation (4) 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 ϕ_c will also be easy to sample from when the priors used are conjugate, and even when Gibbs sampling for ϕ_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 ϕ_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 ϕ_c. We instead represent, and do Gibbs sampling for, only those ϕ_c that are currently associated with some observation. Gibbs sampling for the c_i is based on the following conditional probabilities (with ϕ here being the set of ϕ_c currently associated with at least one observation):

$$\begin{aligned} \text{If } c = c_j \text{ for some } j \neq i\text{:} \quad P(c_i = c \mid c_{-i}, y_i, \phi) &= b\, \frac{n_{-i,c}}{n-1+\alpha}\, F(y_i, \phi_c) \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i, \phi) &= b\, \frac{\alpha}{n-1+\alpha} \int F(y_i, \phi)\, dG_0(\phi) \end{aligned} \tag{11}$$

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 at they faithfully represent whether or not $c_i = c_j$; they may 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 ϕ_{c_i} is chosen from H_i, the posterior distribution of ϕ 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 $c_1, \ldots, c_n$ and $\phi = (\phi_c \;:\; c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$: If the present value of c_i is associated with no other observation (ie, $n_{-i,c_i} = 0$), remove ϕ_{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 (11). If the new c_i is not associated with any other observation, draw a value for ϕ_{c_i} from H_i and add it to the state.
> - For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$.

This is essentially the method used by Bush and MacEachern (1996) and 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 ϕ_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:

$$\begin{aligned} \text{If } c = c_j \text{ for some } j \neq i\text{:}\;\; P(c_i = c \mid c_{-i}, y_i) \;&=\; b\,\frac{n_{-i,c}}{n-1+\alpha} \int F(y_i, \phi)\, dH_{-i,c}(\phi) \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}, y_i) \;&=\; b\,\frac{\alpha}{n-1+\alpha} \int F(y_i, \phi)\, dG_0(\phi) \end{aligned} \tag{12}$$

Here, $H_{-i,c}$ is the posterior distribution of ϕ 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 $c_1, \ldots, c_n$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$: Draw a new value from $c_i \mid c_{-i}, y_i$ as defined by equation (12).

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 above cannot easily be applied to models where G_0 is not the conjugate prior for F, as the integrals in equations (9), (11), and (12) 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) suggest 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 ϕ 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 paper 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 chosing c_i to be a new component can be much lower than it would be if the exact probabilities of equation (11) 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 ϕ 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 ϕ drawn from G_0.

MacEachern and Müller (1998) present a framework that does allow auxiliary values for ϕ 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 consider more specific schemes for assigning distributions to the c_i, which serve to map from a collection of values for ϕ_c to values for the θ_i. Many such schemes will produce the same distribution for the θ_i, but lead to different sampling algorithms.

The "no gaps" algorithm of MacEachern and Müller arises when the c_i for $i = 1, \ldots, n$ 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 $c_1, \ldots, c_n$ and $\phi = (\phi_c \,:\, c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: Let k^- be the number of distinct c_j for $j \neq i$, and let these c_j have values in $\{1, \ldots, k^-\}$. 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 ϕ_{k^-+1} from G_0 if $c_i = c_j$ for some $j \neq i$. Then draw a new value for c_i from $\{1, \ldots, k^-+1\}$ using the following probabilities:

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

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

- For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to ϕ_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 — ie, 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 have also developed an algorithm based on a "complete" scheme for mapping from the ϕ_c to the θ_i. It requires maintaining n values for ϕ, 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 has recently been 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 θ 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 θ is multidimensional.

Finally, Green and Richardson (1998) have 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 paper, 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 aren't easily escaped with incremental changes. Discussion of this issue is beyond the scope of this paper.

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 with density $\pi(x)$ using proposals with density $g(x^*|x)$ updates the state x as follows:

> Draw a candidate state, x^*, according to the density $g(x^*|x)$. Compute the acceptance probability
>
> $$a(x^*, x) \;=\; \min\left[1,\ \frac{g(x|x^*)}{g(x^*|x)}\,\frac{\pi(x^*)}{\pi(x)}\right] \tag{13}$$
>
> 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 π 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 π.

This approach can be applied to model (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 ϕ_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

$$P(c_i = c \mid c_{-i}) \quad = \quad \frac{n_{-i,c} + \alpha/K}{n-1+\alpha} \tag{14}$$

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 (4) 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 (13), leaving

$$a(c_i^*, c_i) \quad = \quad \min\left[1,\ \frac{F(y_i, \phi_{c_i^*})}{F(y_i, \phi_{c_i})}\right] \tag{15}$$

This approach continues to work as we let $K \to \infty$ in order to produce an algorithm for a Dirichlet process mixture model. The conditional prior for c_i becomes

$$\begin{aligned} \text{If } c = c_j \text{ for some } j \neq i\text{:} \quad P(c_i = c \mid c_{-i}) \quad &= \quad \frac{n_{-i,c}}{n-1+\alpha} \\ P(c_i \neq c_j \text{ for all } j \neq i \mid c_{-i}) \quad &= \quad \frac{\alpha}{n-1+\alpha} \end{aligned} \tag{16}$$

If we use this as the proposal distribution for an update to c_i, we will need to draw an associated value for ϕ from G_0 if the candidate, c_i^*, is not in $\{c_1, \ldots, c_n\}$. 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 — ie, 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 behaviour obtained by taking the $K \to \infty$ 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 $c_1, \ldots, c_n$ and $\phi = (\phi_c \ : \ c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$, repeat the following update of c_i R times: Draw a candidate, c_i^*, from the conditional prior for c_i given by equation (16). If a c_i^* not in $\{c_1, \ldots, c_n\}$ is proposed, chose a value for $\phi_{c_i^*}$ from G_0. Compute the acceptance probability, $a(c_i^*, c_i)$, as in equation (15), 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 $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to ϕ_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 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 ϕ_c in the last step of Algorithm 5 are omitted, the result is equivalent to the following:

> **Algorithm 6:** Let the state of the Markov chain consist of $\theta_1, \ldots, \theta_n$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$, repeat the following update of θ_i R times: Draw a candidate, θ_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 θ_i to θ_i^* with this probability; otherwise let the new value of θ_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 θ_i^* is not continuous introduces conceptual, or at least notational, difficulties. Note that this algorithm suffers from the same problem of not being able to change several θ_i simultaneously as was discussed for Algorithm 1.

The behaviour 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 α. (Of course, the probability of actually making such a change depends on α for all methods; here the issue is whether such a change is even considered.)

It is difficult to say which behaviour is better. Algorithm 5 does appear to perform adequately in practice, but since small values of α (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 ϕ drawn from G_0. In order to allow the reverse change, 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, but such a chain would often sample inefficiently, since it can change an observation from one existing component to another only by passing though 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 observations not changing. No difficult integrations are required for this partial Gibbs sampling operation.

Combining the modified Metropolis-Hasting updates, the partial Gibbs sampling updates, and the usual updates to ϕ_c for $c \in \{c_1, \ldots, c_n\}$ produces the following algorithm:

> **Algorithm 7:** Let the state of the Markov chain consist of $c_1, \ldots, c_n$ and $\phi = (\phi_c \,:\, c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$, update c_i as follows: If c_i is a not a singleton (ie, $c_i = c_j$ 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 $i = 1, \ldots, n$: If c_i is a singleton (ie, $c_i \neq c_j$ for all $j \neq i$), do nothing. Otherwise, choose a new value for c_i from $\{c_1, \ldots, c_n\}$ using the following probabilities:
>
> $$P(c_i = c \mid c_{-i},\, y_i,\, \phi,\, c_i \in \{c_1, \ldots, c_n\}) \;=\; b\, \frac{n_{-i,c}}{n-1}\, F(y_i, \phi_c)$$
>
> where b is the appropriate normalizing constant.
>
> - For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to ϕ_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.

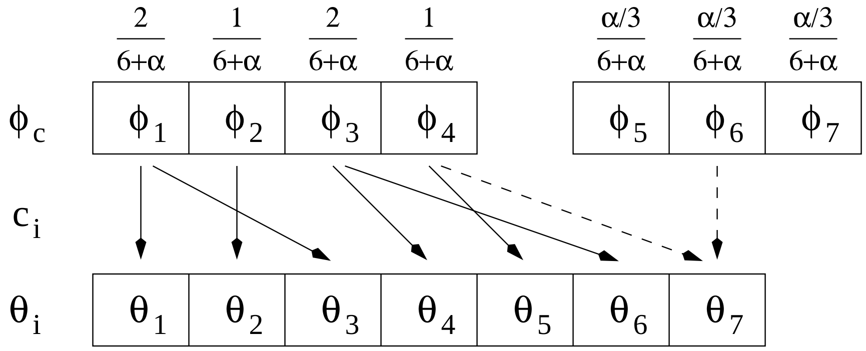

Figure 1: Representing the conditional prior distribution for a new observation using auxiliary parameters. The component for the new observation is chosen from among the four components associated with other observations plus three possible new components, with parameters, ϕ_5, ϕ_6, ϕ_7, drawn independently 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.

The basic idea of auxiliary variable methods is that we can sample from a distribution π_x for x by sampling from some distribution π_{xy} for (x, y), with respect to which the marginal distribution of x is π_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 π_{xy}.
2) Perform some update of (x, y) that leaves π_{xy} invariant.
3) Discard y, leaving only the value of x.

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

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 ϕ_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 $\{1, \ldots, k^-\}$, 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 $\{1, \ldots, k^-\}$ 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 $j \neq i$ (ie, 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 ϕ must of course be equal to the existing ϕ_{c_i}. The ϕ 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 ϕ 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 $c_1, \ldots, c_n$ and $\phi = (\phi_c \,:\, c \in \{c_1, \ldots, c_n\})$. Repeatedly sample as follows:
>
> - For $i = 1, \ldots, n$: 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 $\{1, \ldots, k^-\}$. If $c_i = c_j$ for some $j \neq i$, draw values independently from G_0 for those ϕ_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 ϕ_c for which $k^- + 1 < c \leq h$. Draw a new value for c_i from $\{1, \ldots, h\}$ using the following probabilities:
>
> $$P(c_i = c \mid c_{-i}, y_i, \phi_1, \ldots, \phi_h) \;=\; \begin{cases} b \, \dfrac{n_{-i,c}}{n-1+\alpha} \, F(y_i, \phi_c) & \text{for } 1 \leq c \leq k^- \\[2ex] b \, \dfrac{\alpha/m}{n-1+\alpha} \, F(y_i, \phi_c) & \text{for } k^- < c \leq h \end{cases}$$
>
>   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 ϕ_c that are now associated with one or more observations.
>
> - For all $c \in \{c_1, \ldots, c_n\}$: Draw a new value from $\phi_c \mid y_i$ s.t. $c_i = c$, or perform some other update to ϕ_c that leaves this distribution invariant.

Note that the relabellings 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 α is small, this seems to be a clear benefit, since the probabilities for other changes are affected only slightly.

In the other extreme, as $m \to \infty$, Algorithm 8 approaches the behaviour of Algorithm 2, since the m (or $m-1$) values for ϕ_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 α 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 α given the other parameters depends only on k, the number of distinct c_i. It can be updated by some Metropolis-Hastings method, or by methods discussed by Escobar and West (1995).

If F depends on hyperparameters γ, the conditional density for γ given the current θ_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 η, the conditional density for η given the current c_i and ϕ_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 $\{c_1, \ldots, c_n\}$. 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 γ and η 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 ϕ parameters are used. If ϕ values not associated with any observation are retained in the state, the conditional distribution for η given the rest of the state will include factors of $G_0(\phi, \eta)$ for these ϕ as well as for the ϕ values associated with observations. Since this will tend to slow convergence, it is desirable to discard all unused ϕ values, regenerating them from G_0 as needed, as is done for the algorithms in this paper.

8 A demonstration

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

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

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.

	Time per iteration in microseconds	*Autocorrelation time for k*	*Autocorrelation time for θ_1*
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 ϕ 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

Table 1: Performance of the algorithms tested.

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 20000 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 θ_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 distribution (see Ripley 1987, Section 6.3). It was estimated using autocorrelation estimates from the 20000 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 $m = 2$ and $m = 3$. 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 behaviour 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 6 performs about as well as Algorithm 8 with $m = 1$. As expected, performance is much worse when updates for the ϕ_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. Of course, the relative performance of the methods may be different from what is seen here when they are applied to more realistic problems, in which the number of observations is often greater, and updates are also done for various hyperparameters. The methods tested here are implemented in my software for flexible Bayesian modeling (version of 1998-09-01), available from my web page, which can be used to experiment with some more complex models.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

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", *Annals of Statistics*, vol. 2, pp. 1152-1174.

Blackwell, D. and MacQueen, J. B. (1973) "Ferguson distributions via Pólya urn schemes", *Annals of Statistics*, vol. 1, pp. 353-355.

Bush, C. A. and MacEachern, S. N. (1996) "A semiparametric Bayesian model for randomised block designs", *Biometrika*, vol. 83, pp. 275-285.

Escobar, M. D. (1994) "Estimating normal means with a Dirichlet process prior", *Journal of the American Statistical Association*, vol. 89, pp. 268-277.

Escobar, M. D. and West, M. (1995) "Bayesian density estimation and inference using mixtures", *Journal of the American Statistical Association*, vol. 90, pp.577-588.

Ferguson, T. S. (1973) "A Bayesian analysis of some nonparametric problems", *Annals of Statistics*, vol. 1, pp. 209-230.

Ferguson, T. S. (1983) "Bayesian density estimation by mixtures of normal distributions", in H. Rizvi and J. Rustagi (editors) *Recent Advances in Statistics*, pp. 287-303, New York: Academic Press.

Green, P. J. and Richardson, S. (1998) "Modelling heterogeneity with and without the Dirichlet process", draft manuscript.

Hastings, W. K. (1970) "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika*, vol. 57, pp. 97-109.

MacEachern, S. N. (1994) "Estimating normal means with a conjugate style Dirichlet process prior", *Communications in Statistics: Simulation and Computation*, vol. 23, pp. 727-741.

MacEachern, S. N. and Müller, P. (1998) "Estimating mixture of Dirichlet process models", *Journal of Computational and Graphical Statistics*, vol. 7, pp. 223-238.

Neal, R. M. (1992) "Bayesian mixture modeling", in C. R. Smith, G. J. Erickson, and P. O. Neudorfer (editors) *Maximum Entropy and Bayesian Methods: Proceedings of* *the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Seattle, 1991*, p. 197-211, Dordrecht: Kluwer Academic Publishers.

Ripley, B. D. (1987) *Stochastic Simulation*, New York: Wiley.

Walker, S. and Damien, P. (1998) "Sampling methods for Bayesian nonparametric inference involving stochastic processes", draft manuscript.

West, M., Müller, P., and Escobar, M. D. (1994) "Hierarchical priors and mixture models, with application in regression and density estimation", in P. R. Freeman and A. F. M. Smith (editors) *Aspects of Uncertainty*, pp. 363-386, John Wiley.