INTERNATIONAL CENTRE FOR ECONOMIC RESEARCH

ICER

WORKING PAPER SERIES

Stephen G. Walker

SAMPLING THE DIRICHLET MIXTURE MODEL WITH SLICES

Working Paper No. 16/2006

SAMPLING THE DIRICHLET MIXTURE MODEL WITH SLICES

STEPHEN G. WALKER

Institute of Mathematics, Statistics and Actuarial Science
University of Kent Canterbury, Kent, CT2 7NZ, U.K.

May 2006

Abstract: We provide a new approach to the sampling of the well known mixture of Dirichlet process model. Recent attention has focused on retention of the random distribution function in the model, but sampling algorithms have then suffered from the countably infinite representation these distributions have. The key to the algorithm detailed in this paper, which also keeps the random distribution functions, is the introduction of a latent variable which allows a finite number, which is known, of objects to be sampled within each iteration of a Gibbs sampler.

Keywords: Bayesian Nonparametrics, Density estimation, Dirichlet process, Gibbs sampler, Slice sampling.

Acknowledgements: The author is an EPSRC Advanced Research Fellow and the paper was written during a visit to the University of Turin funded by ICER.

1. Introduction. The aim of this paper is to introduce a new method for sampling the well known and widely used mixture of Dirichlet process (MDP) model. There have been a number of recent contributions to the literature on this problem, notably Ishwaran & Zarepour (2000) and Papaspiliopoulos & Roberts (2005). These papers have been concerned with sampling the MDP model while retaining the random distribution functions.

The issue and the causes of the complexities is the countably infiniteness of the discrete masses from the random distribution functions chosen from the Dirichlet process prior. Ishwaran & Zarepour (2000) circumvent this with an approximate method based on a truncation of the distributions. Motivated by the work of Ishwaran & Zarepour (2000), Papaspiliopoulos & Roberts (2005) proposed an exact algorithm based on the notion of retrospective sampling. However, the algorithm itself becomes non-trivial when applied to the MDP mocdel, and involves setting up a detailed balance criterion with connecting proposal moves (Green, 1995). On the other hand, we find a simple trick, based on the slice sampling schemes (Damien et al., 1999), which deals with the infiniteness. The introduction of latent variables makes finite the part of the random distribution function required to iterate through a Gibbs sampler. Moreover, all the conditional distributions are easy to sample and no accept/reject methods are needed.

The first sampler for the MDP model, based on a Gibbs sampler, was given in the PhD Thesis of Escobar (1988, 1994). Alternative approaches have been proposed by MacEachern (1994) and co-authors; for example, MacEachern & Müller (1998). A recent survey is given in MacEachern (1998), and other papers in the book of Dey et al. (1998), and by Müller & Quntana (2004). Richardson & Green (1997) provide a comparison with more traditional mixture models and Neal (2000) also discusses ideas for sampling the MDP model.

Recently, Ishwaran & James (2001) developed a Gibbs sampling scheme involving more general stick-breaking priors, which is a direct extension of the Escobar (1998) approach. Escobar's Gibbs sampler makes use of the Pólya-urn sampling scheme (Blackwell & MacQueen, 1973) and the idea of using the Pólya-urn scheme is connected with the procedure of integrating out of the model the random distribution function from the Dirichlet process. Recent attempts have avoided this step and retained the random distribution functions in the algorithms, notably Ishwaran & Zarepour (2001) and Papaspiliopoulos & Roberts (2005).

In Section 2 we describe the Dirichlet process mixture model and describe the latent variables of use to the sampling strategy. In Section 3 we will write down the algorithm for the Gibbs sampler and Section 4 contains a couple of illustrative examples. Finally, Section 5 concludes with a brief discussion.

2. The Dirichlet Process Model. Let $D(c, P_0)$ denote a Dirichlet process prior (Ferguson, 1973) with scale parameter $c > 0$ and prior probability measure P_0. So, for example, $E(P) = P_0$ and

$$\text{Var}(P(A)) = \frac{P_0(A)\{1 - P_0(A)\}}{c+1}$$

for all appropriate sets A. The posterior distribution of P given n independent and identically distributed samples from P is also a Dirichlet process with new parameters $c + n$ and

$$\frac{cP_0 + nP_n}{c+n},$$

where P_n is the empirical distribution function. However, we will not be needing this particular result.

It is well known that a random probability measure P can be chosen from $D(c, P_0)$ via the following sampling scheme, attributable to Sethuraman &

Tiwari (1982), see also Sethuraman (1994), and involving the so-called stick-breaking prior (see, for example, Freedman, 1963; Connor & Mosimann, 1969). Take $v_1, v_2, \ldots$ to be independent and identically distributed beta$(1, c)$ variables and take $\theta_1, \theta_2, \ldots$ to be independent and identically distributed from P_0, which we will assume has density g_0 with respect to the Lebesgue measure. Then define

$$P = \sum_{j=1}^{\infty} w_j \, \delta_{\theta_j},$$

where $w_1 = v_1$ and for $j > 1$,

$$w_j = v_j \prod_{l<j} (1 - v_l).$$

Here δ_θ denotes the measure with a point mass of 1 at θ. The weights are obtained via what is known as a stick-breaking procedure. Ishwaran & James (2001) consider a more general model with the $v_j \sim$ beta(a_j, b_j) and show that the sum of weights is 1 almost surely when

$$\sum_{j=1}^{\infty} \log(1 + a_j/b_j) = +\infty.$$

While we work with the v's which lead to the Dirichlet process, our algorithm for sampling the MDP model can be extended to cover other stick-breaking prior distributions in a simple way. This will be elaborated on later in the paper.

The MDP model is based on the idea of constructing absolutely continuous random distribution functions and was first considered in Lo (1984). The random distribution function chosen from a Dirichlet process is almost surely discrete (Blackwell, 1973). Consequently, consider the random density function

$$f_P(y) = \int \mathrm{N}(y|\theta) \, \mathrm{d}P(\theta).$$

Here $\mathrm{N}(y|\theta)$ denotes a conditional density function, which will typically be a normal distribution and the parameters of which are represented by θ. So in the normal case $\theta = (\mu, \sigma^2)$. Given the form for P, we can write

$$f_{w,\theta}(y) = \sum_j w_j \, \mathrm{N}(y|\theta_j).$$

The prior distributions for the w and θ have been given earlier.

Our attempt to estimate the model, via Gibbs sampling ideas, is to introduce a latent variable u such that the joint density with of (y, u) given (w, θ) is given by

$$f_{w,\theta}(y, u) = \sum_j \mathbf{1}(u < w_j) \, \mathrm{N}(y|\theta_j).$$

Clearly integrating over u with respect to the Lebesgue measure returns us the desired density $f_{w,\theta}(y)$. Hence, the joint density exists and so there will also exist a marginal density for u. Alternatively we can write

$$f_{w,\theta}(y, u) = \sum_{j=1}^{\infty} w_j \, \mathrm{U}(u|0, w_j) \, N(y|\theta_j)$$

and so with probability w_j, y and u are independent and are, respectively, normal and uniform distributed. Hence, the marginal density for u is given by

$$f_w(u) = \sum_{j=1}^{\infty} w_j \, \mathrm{U}(u|0, w_j) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j).$$

If we let

$$A_w(u) = \{j : w_j > u\}$$

then we can equally write

$$f_{w,\theta}(y, u) = \sum_{j \in A_w(u)} \mathrm{N}(y|\theta_j).$$

Note, it is quite clear that $A_w(u)$ is a finite set for all $u > 0$. The conditional density of y given u is given by

$$f_{w,\theta}(y|u) = \frac{1}{f_w(u)} \sum_{j \in A_w(u)} \mathrm{N}(y|\theta_j),$$

where $f_w(u) = \sum_j \mathbf{1}(u < w_j)$ is the marginal density for u, being defined on $(0, w^*)$ where w^* is the largest w_j.

The usefulness of the latent variable u will become clear later on. A brief comment here is that the move from an infinite sum to a finite sum, given u, is going to make a lot of difference when sampling is involved.

So, given u, we have a finite mixture model with equal weights, all equal to $1/f_w(u)$. We can now introduce a further indicator latent variable which will identify the component of the mixture from which y is to be taken. Therefore, consider the joint density

$$f_{w,\theta}(y, \delta = k, u) = \mathrm{N}(y|\theta_k)\mathbf{1}(k \in A(u)).$$

The complete data likelihood based on a sample of size n is easily seen to be

$$l_{w,\theta}\left(\{y_i, u_i, \delta_i = k_i\}_{i=1}^n\right) = \prod_{i=1}^n \mathrm{N}(y_i|\theta_{k_i})\,\mathbf{1}(u_i < w_{k_i}).$$

As has been mentioned, we already know the prior distributions for the w and θ. Though as it happens, we will use the v's rather than the w's when it comes to sampling.

3. The Sampling Algorithm. In order to implement a Gibbs sampler we require the set of full conditional density functions. For the infinite collection of variables v and θ, it would seem that we would need to sample the entire set. But this is not required. We only need to sample a finite set of them at each stage in order to progress to the next iteration. All un-sampled v_j's and θ_j's will be independent samples from the priors; that is beta$(1, c)$ and g_0, respectively. Let us proceed to consider the full conditional densities; listed **A** to **E**.

A. We will start with the u_i's. These are easy to find and are the uniform distributions on the interval

$$(0, w_{k_i}).$$

B. Next we have θ_j, and this is easily seen to be the density function given up to a constant of proportionality by

$$f(\theta_j|\cdots) \propto g_0(\theta_j) \prod_{k_i=j} \mathrm{N}(y_i|\theta_j).$$

If there are no k_i equal to j then $f(\theta_j|\cdots) = g_0(\theta_j)$.

C. Slightly harder, but quite do-able, is the sampling of the v_j's. For the joint full conditional density we have

$$f(v|\cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}(w_{k_i} > u_i),$$

where $\pi(v)$ denotes the collection of independent beta variables, and we have already given the relation between the w_j's and the v_j's. Hence,

$$f(v|\cdots) \propto \pi(v) \prod_{i=1}^{n} \mathbf{1}\left(v_{k_i} \prod_{l<k_i} (1-v_l) > u_i\right).$$

It is quite evident from this that only the v_j's for $j \le k^*$, where k^* is the maximum of $\{k_1, \ldots, k_n\}$, will be affected; that is, for $j > k^*$, we have $f(v_j|\cdots) = \mathrm{beta}(1, c)$. For $j \le k^*$ we have

$$f(v_j|v_{-j}, \cdots) \propto \mathrm{beta}(v_j|1, c)\mathbf{1}\left(\alpha_j < v_j < \beta_j\right),$$

where

$$\alpha_j = \max_{k_i=j} \left\{ \frac{u_i}{\prod_{l<j}(1-v_l)} \right\}.$$

and

$$\beta_j = 1 - \max_{k_i > j} \left\{ \frac{u_i}{v_{k_i} \prod_{l<k_i, l \neq j} (1 - v_l)} \right\}.$$

Then the distribution function, on $\alpha_j < v_j < \beta_j$, is given by

$$F(v_j) = \frac{(1-\alpha_j)^c - (1-v_j)^c}{(1-\alpha_j)^c - (1-\beta_j)^c}$$

and so a sample can be taken via the inverse cdf technique. Clearly, it is now evident that this approach covers more general stick-breaking models; it is no more difficult to sample a truncated beta variable when we have $v_j \sim \text{beta}(a_j, b_j)$ as the priors.

D. We now discuss the sampling of the indicator variables. We clearly have

$$\text{pr}(\delta_i = k | \cdots) \propto \mathbf{1}(k \in A_w(u_i))\, \text{N}(y_i|\theta_k).$$

Clearly $A_w(u_i)$ is not empty; at least $k_i \in A_w(u_i)$.

Before providing details on how to sample this, we mention that without the latent variables u_i, the possible choices of δ_i would be infinite and problems then arise with the normalising constant. Papaspiliopoulos & Roberts (2005) attempted to circumvent the problem via retrospective sampling and the use of a detailed-balance criterion, which is non-trivial. Our approach is quite easy to implement. The choice of δ_i is from a finite set, which is $\{k : w_k > u_i\}$. So we sample as many of the w_k's until we are sure that we have all the $w_k > u_i$. How do we know this? We are sure there can be no further $k > k^i$ for which $w_k > u_i$ when we have k^i such that

$$\sum_{j=1}^{k^i} w_j > 1 - u_i.$$

So, to cover all the i's, we find the smallest k^* such that

$$\sum_{j=1}^{k^*} w_j > 1 - u^*,$$

where $u^* = \min\{u_1, \ldots, u_n\}$. Hence, we now know how many of the w_k's we need to sample in order for the chain to proceed; it is $\{w_1, \ldots, w_{k^*}\}$. It is that k^* will be necessary to find to implement the algorithm. One needs to know how many of the w_j are larger than u and it is only at k^* that one knows for sure that all have been found. Hence, k^* is not a loose approximation; it is an exact piece of information.

For the prior model it is that,

$$k^* \sim 1 + \text{Poisson}(-c \log u^*).$$

See Muliere & Tardella (1998).

E. We can incorporate a prior on c, say $\pi(c)$. We will sample $f(c, w, \theta | y, u, \delta)$ as a block, and will sample this in two stages; first by sampling from $f(c|y, u, \delta)$ and then $f(w, \theta | c, y, u, \delta)$. We have already described how to sample from the latter of these. For the former, it is equivalent to the full conditional density that would arise from the marginal model, that is the one in which the random distribution functions are removed from the model. Therefore, as is well known, it is only the δ and the sample size that provides information about c. To elaborate on this, the conditional distribution of c depends only on the number of clusters; that is, the number of distinct k_i's, call this d, and that

$$f(c|d, n) \propto c^d \Gamma(c)\, \pi(c) / \Gamma(c+n),$$

where $\Gamma(\cdot)$ denotes the usual gamma function. A nice way to sample from this is given in Escobar & West (1995) when $\pi(c)$ is a gamma distribution.

Hence, all the conditional densities are easy to sample and the Markov chain we have constructed is automatic. It requires no tuning nor retrospective steps.

For density estimation we would like to sample from the predictive distribution of

$$f(y_{n+1}|y_1,\ldots,y_n).$$

At each iteration we have (w_j,θ_j) and we sample a θ_j using the weights. The idea is to sample a uniform random variable r from the unit interval and to take that θ_j for which $w_{j-1}<r<w_j$, with $w_0=0$. If more weights are required than currently exist then it is straightforward to sample more as we know the additional v_j's for $j>k^*$ are independent and identically distributed from beta$(1,c)$ and the additional θ_j's are independent and identically distributed from g_0. Having taken θ_j, we draw y_{n+1} from $\mathrm{N}(\cdot|\theta_j)$.

4. Illustration. Here we present a normal example in which $\theta=(\mu,\sigma^2)$ and we will take $\lambda=\sigma^{-2}$. The prior for the μ_j's will be independent $\mathrm{N}(0,1/s)$ and the prior for the λ_j's will be independent $\mathrm{Ga}(\epsilon,\epsilon)$. To complement Section 3 we now provide the conditional distributions for μ_j and λ_j. We have

$$f(\mu_j|\cdots)=\mathrm{N}\left(\frac{\xi_j\lambda_j}{m_j\lambda_j+s},\frac{1}{m_j\lambda_j+s}\right),$$

where

$$\xi_j=\sum_{k_i=j}y_i$$

and

$$m_j=\sum_{k_i=j}1.$$

We also have

$$f(\lambda_j|\cdots)=\mathrm{Ga}(\epsilon+m_j/2,\epsilon+d_j/2),$$

where

$$d_j=\sum_{k_i=j}(y_i-\theta_j)^2.$$

In the simulated data set example that follows, the code was written using scilab, which is freely downloadable from the internet.

We sampled 50 random variables independently from the mixture of normal distributions given by

$$f(y)=\frac{1}{3}\mathrm{N}(y|-4,1)+\frac{1}{3}\mathrm{N}(y|0,1)+\frac{1}{3}\mathrm{N}(y|8,1).$$

Choosing non-informative specifications, we took $\epsilon=0.5$, $s=0.1$ and the gamma prior for c to be $\mathrm{Ga}(0.1,0.1)$, the Gibbs sampler was run for 20,000 iterations and at each iteration from 10,000 onwards a predictive sample y_{n+1} was taken. A histogram of the 50 data points with the density estimator based on the 10,000 samples of y_{n+1} is provided in Figure 1. The density estimator was obtained using the R density routine with bandwidth set to 0.3.

Figure 2 presents the running average for the number of clusters sampled at each iteration. So it is clear that 10,000 samples is good enough for the chain to reach stationarity and hence the samples from 10,000 onwards can be taken as coming from the predictive distribution.

5. Discussion. We have provided a simple and fast way to sample the MDP model. The key is the introduction of the latent variables which truncate the weights of the random Dirichlet distributions. It is a highly simple piece of code to write and is direct in the sense that no accept/reject sampling nor retrospective sampling is required. It is also remarkably quick to run. It improves on current approaches in the following way: we know exactly how many of the w_j's and θ_j's we need to sample at each iteration - it is k^*. This fundamental result eludes the alternative approaches.

Retaining the random distribution function is useful as it removes the dependence between the θ_{k_i}'s which exist in the Pólya-urn model. However, retaining the random distributions leads to problems with the countably infinite representation. In this paper we deal with it by introducing a latent variable which makes the representation finite for the purposes of proceeding with the sampling and allowing sampling from the predictive distribution. The sampling of the latent variable given the other variables is a uniform distribution.

In the non-conjugate case, that is when $N(y|\theta)$ and $g_0(\theta)$ form a non-conjugate pair and perhaps difficult to sample, then a possible useful solution is again provided by the latent variable ideas presented in Damien et al. (1999, Sections 4 & 5).

References

Blackwell, D. (1973). The discreteness of Ferguson selections. *Annals of Statistics* **1**, 356–358.

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

Connor, R.J. & Mosimann, J.E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association* **64**, 194–206.

Damien, P., Wakefield, J.C. & Walker, S.G. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Series B* **61**, 331–344.

Dey, D., Sinha, D. & Müller, P. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Lecture Notes in Statistics. Springer, New York.

Escobar, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

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

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

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

Freedman, D.A. (1963). On the asymptotic behaviour of Bayes estimates in the discrete case I. *Annals of Mathematical Statistics* **34**, 1386–1403.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Ishwaran, H. & Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two parameter process hierarchical models. *Biometrika* **87**, 371–390.

Ishwaran, H. & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics* **12**, 351–357.

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

MacEachern, S.N. (1998). Computational methods for Mixture of Dirichlet Process Models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D.Dey, P.Müller, D.Sinha eds.) 23–43. Springer, New York.

MacEachern, S.N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

Muliere, P. & Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics* **26**, 283–297.

Müller, P. & Quintana, F.A. (2004). Nonparametric Bayesian Data Analysis.*Statistical Science* **19**, 95–110.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Richardson, S. & Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

Papaspiliopoulos, O. & Roberts, G.O. (2005). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Submitted.

Sethuraman, J. & Tiwari, R. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Proceedings of the third Purdue symposium on statistical decision theory and related topics.* Gupta, S.S. and Berger, J.O. (Eds.) Academic press, New York.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

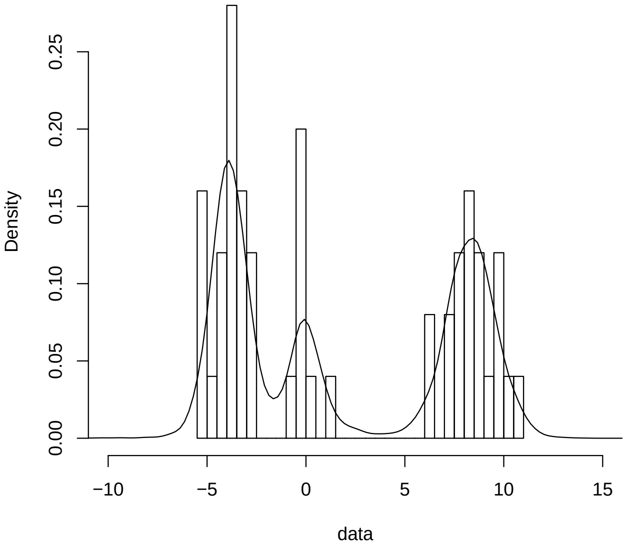

Figure 1: Histogram of data and density estimate of predictive density for 1/3N(-4,1)+1/3N(0,1)+1/3N(8,1)

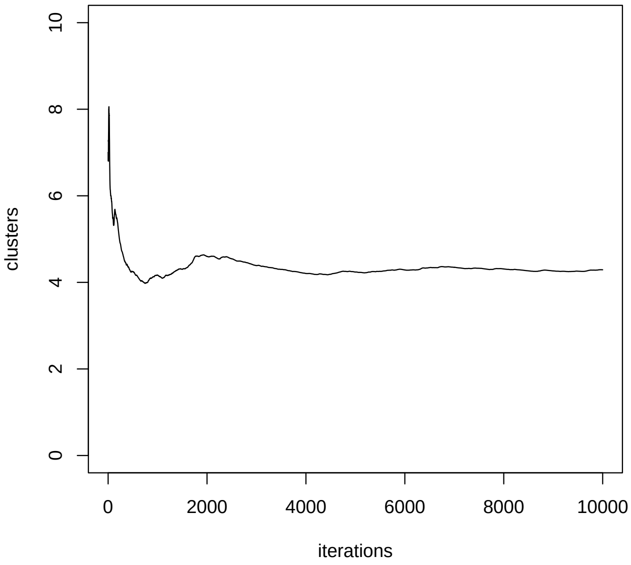

Figure 2: Running average for the number of clusters up to iteration 10000