RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

DECEMBER 1996

SAMPLING A DIRECHLET PROCESS MIXTURE MODEL

WORKING PAPER #9612-39

STEPHEN WALKER
IMPERIAL COLLEGE, LONDON, UK

AND

PAUL DAMIEN
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

Sampling a Dirichlet Process Mixture Model

Stephen Walker[1]
and
Paul Damien[2]

[1] Department of Mathematics, Imperial College, 180 Queen's Gate, London
SW7 2BZ

[2] Department of Statistics and Management Science, School of Business,
University of Michigan, Ann Arbor, 48109-1234, USA.

SUMMARY

We introduce a new method for sampling the so called mixture of Dirichlet process (MDP) model which has recently received a great deal of attention in the literature (Bush & MacEarchen, 1996; Mueller et al., 1996). The solution is based on the introduction of strategic latent variables to facilitate the implementation of a Gibbs sampler in which all full conditional distributions are of known types and can be sampled directly.

Keywords: Latent variables, Gibbs sampler, Dirichlet process.

1 Introduction

This paper is concerned with the following general hierarchical model involving the Dirichlet process which is commonly known as the MDP model:

$$z_i|\theta_i \sim f(.;\theta_i), \quad i = 1, ..., n,$$

$$\theta_1,\dots,\theta_n|G \sim_{\text{iid}} G$$

and

$$G \sim \mathcal{D}(\alpha G_0).$$

The notation is taken from the paper of Mueller et al. (1996). Briefly the probability model assumes z_i given θ_i $(i = 1,\dots,n)$ is an observation from the known distribution $f(.;\theta_i)$. Here $\mathcal{D}$ represents a Dirichlet process (Ferguson, 1973) and G is a discrete random distribution (taken from the Dirichlet process) conditional on which the $\{\theta_i\}$ are independent and identically distributed (iid). The parameters of the Dirichlet process are the distribution G_0 which centers the process in that $EG = G_0$ and $\alpha > 0$ which loosely represents the *strength of belief* in G_0.

There is now a substantial amount of literature involving this model (Escobar, 1994; Escobar & West, 1995; MacEarchen, 1994; West et al., 1994; MacEarchen & Mueller, 1994; Bush & MacEarchen, 1996; Mueller et al., 1996). Inference is performed using Markov chain Monte Carlo (MCMC) methods (Tierney, 1994) and in particular the Gibbs sampler (Smith and Roberts, 1993). This requires sampling sequentially from the full conditional distributions. For us (at least for the moment) there is no loss of generality in assuming that both α and G_0 are fixed.

The full conditional distribution for θ_i poses the problem for the Gibbs sampler (usually the others are of known types). This distribution is given, up to a constant of proportionality, by

$$f^*(\theta_i) \propto f(z_i|\theta_i)\Big\{\alpha G_0(\theta_i) + \sum_{j\neq i}\delta_{\theta_j}(\theta_i)\Big\},$$

where $\delta_\theta(.)$ is the measure putting mass 1 at θ. Here we will always use a $*$ to represent a conditional distribution. If $q_{i0} = \alpha \int f(z_i|\theta_i)dG_0(\theta_i)$ is analytically tractable then $f^*(.)$ can be sampled directly. That is, θ_i is taken from the distribution proportional to $f(z_i|\theta_i)G_0(\theta_i)$ with probability proportional to q_{i0} or is taken to be θ_j $(j \neq i)$ with probability proportional to $q_{ij} = f(y_i|\theta_j)$.

The only difficulties with implementing a Gibbs sampler for the MDP model are when the integral $\int f(z_i|\theta_i)dG_0(\theta_i)$ is intractable and hence q_{i0} is not available. According to MacEarchen & Mueller (1994):

> Evaluation of the integral expression for q_{i0} is non-trivial unless $G_0(\theta)$ and $f(.|\theta)$ are a conjugate pair. Current implementations

> therefore either use a conjugate model or rely on approximate computations. Overcoming this computational hurdle is important because of the wide range of current and potential applications of MDP models, and the need in most applications to leave the conjugate framework.

An excellent review of these current approximate computations is found in the MacEarchen & Mueller paper. They go on to say:

> If the base distribution $G_0(\theta)$ and the likelihood $f(.|\theta)$ are chosen as a conjugate pair, then all distributions can be efficiently generated from and no complications arise. If, however, $G_0(\theta)$ is not conjugate with $f(.|\theta)$ then resampling the configuration becomes difficult, as the integral q_{i0} may be computationally intensive.

The method of MacEarchen & Mueller to solving the nonconjugate model relies on the introduction of a latent model and is described in Section 3 of their paper. We omit the details of their algorithm.

In the next Section we introduce a latent model which runs on a Gibbs sampler in which *all* the full conditional distributions are of known types and can be sampled directly. Essentially we are introducing latent variables which mean that the q_{i0}s can be evaluated.

2 A New Latent MDP Model

Recall the problem is sampling from the distribution given by

$$f^*(\theta_i) \propto l(\theta_i)\Big\{\alpha G_0(\theta_i) + \sum_{j\neq i} \delta_{\theta_j}(\theta_i)\Big\}.$$

where we have written $f(z_i|\theta_i)$ as $l(\theta_i)$. Here we introduce the latent variable u_i and construct the joint density with θ_i given, up to a constant of proportionality, by

$$f^*(\theta_i, u_i) \propto I\left(u_i < l(\theta_i)\right)\left\{\alpha G_0(\theta_i) + \sum_{j\neq i} \delta_{\theta_j}(\theta_i)\right\}.$$

where $I(.)$ represents an indicator function. Clearly the marginal distribution for θ_i is $f^*(.)$. Note that the full posterior distribution for (θ, u) is given by

$$f^*(\theta, u) \propto \prod_i \left\{ I(u_i < l(\theta_i)) \left(\alpha G_0(\theta_i) + \sum_{j<i} \delta_{\theta_j}(\theta_i) \right) \right\}.$$

Therefore the Gibbs sampler now runs over the additional full conditionals given by

$$f^*(u_i|\theta_i) = U(0, l(\theta_i)),$$

where $U(a, b)$ is the uniform distribution on the interval (a, b), and the full conditional for θ_i is now in the more friendly form,

$$f^*(\theta_i|u_i) \propto \alpha G_0(\theta_i) I(l(\theta_i) > u_i) + \sum_{l(\theta_j) > u_i} \delta_{\theta_j}(\theta_i).$$

The full conditional for u_i is obviously trivial to sample. The second should not pose any problem either, it is merely sampling from G_0 restricted to a particular set.

Example 1

Here we consider a one-dimensional nonconjugate normal/uniform MDP model. Then $l(\theta_i) = \exp\{-0.5(z_i - \theta_i)^2/\sigma^2\}$ and $G_0 = U(\mu_L, \mu_H)$ and

$$f^*(\theta_i|u_i) \propto \alpha(\mu_H - \mu_L)^{-1} I(\max\{\mu_L, a_{ui}\} < \theta_i < \min\{\mu_H, b_{ui}\}) + \sum_{l(\theta_j) > u_i} \delta_{\theta_j}(\theta_i).$$

where $a_{ui} = z_i - \sigma\sqrt{-2\log u_i}$ and $b_{ui} = z_i + \sigma\sqrt{-2\log u_i}$ (note that $u_i < 1$). Thus θ_i is taken from the uniform distribution on the interval

$$(\max\{\mu_L, a_{ui}\}, \min\{\mu_H, b_{ui}\})$$

with probability proportional to

$$q_{i0} = \alpha(\mu_H - \mu_L)^{-1}(\min\{\mu_H, b_{ui}\} - \max\{\mu_L, a_{ui}\}),$$

or is equal to $\theta_j : l(\theta_j) > u_i$ with probability proportional to 1. Note therefore that if $\{\theta_j : l(\theta_j) > u_i\} = \emptyset$ then θ_i is automatically taken from the uniform distribution.

For the general nonconjugate model let $A_{ui} = \{\theta_i : l(\theta_i) > u_i\}$ so we can write

$$f^*(\theta_i|u_i) = \frac{\alpha G_0(A_{ui})G_0(\theta_i, A_{ui}) + \sum_{l(\theta_j)>u_i} \delta_{\theta_j}(\theta_i)}{\alpha G_0(A_{ui}) + \sum_{l(\theta_j)>u_i} 1}.$$

where $G_0(\theta_i, A_{ui})$ is the normalised $G_0(\theta_i)I(\theta_i \in A_{ui})$. Therefore we take θ_i from $G_0(\theta_i, A_{ui})$ with probability

$$\frac{\alpha G_0(A_{ui})}{\alpha G_0(A_{ui}) + \sum_{l(\theta_j)>u_i} 1}.$$

or take $\theta_i = \theta_j : l(\theta_j > u_i)$ with probability $1/(\alpha G_0(A_{ui}) + \sum_{l(\theta_j)>u_i} 1)$. Note that the 'new' $q_{i0} = \alpha G_0(A_{ui})$.

Example 2

Here we consider a binomial model with logit link, that is,

$$z_i|p_i \sim \text{binomial}(p_i, n_i) \text{ and } \text{logit} p_i = \theta_i.$$

Therefore

$$f(z_i|\theta_i) \propto \frac{\exp(\theta_i z_i)}{(1+\exp(\theta_i))^{n_i}}.$$

For this model we introduce the latent variables (u_i, v_i) so that the joint posterior with θ_i is given by

$$f^*(\theta_i, u_i, v_i|z) \propto \left\{e^{-u_i z_i - v_i(n_i - z_i)} I\left(u_i > \log\{1+e^{-\theta_i}\}, v_i > \log\{1+e^{\theta_i}\}\right) G_0(\theta_i)\right\}.$$

where $G_0(\theta_i) = N(\theta_i|\mu, \sigma^2)$ is the prior. Clearly the marginal posterior for θ_i is as required. With the inclusion of the Dirichlet prior the full conditional for θ_i becomes

$$f^*(\theta_i|u_i, v_i) \propto I(a_{ui} < \theta_i < b_{vi}) \left\{\alpha G_0(\theta_i) + \sum_{j \neq i} \delta_{\theta_j}(\theta_i)\right\}.$$

where $a_{ui} = -\log(e^{u_i} - 1)$ and $b_{vi} = \log(e^{v_i} - 1)$. Therefore we take $\theta_i = \theta_j : a_{ui} < \theta_j < b_{vi}$ or from a truncated normal distribution according to easily available probabilities noting that $q_{i0} = \int_{(a_{ui}, b_{vi})} N(\theta_i|\mu, \sigma^2) d\theta_i$.

Complications can arise when θ_i is p-dimensional $(p > 1)$ since obtaining a p-dimensional A_{ui} may be difficult. A solution to this problem is to consider $\theta_i = (\theta_{1i}, \ldots, \theta_{pi})$ where p is the dimension of each θ_i. The strategy is to replace the sampling from $f^*(\theta_i|u_i)$ by sampling from $f^*(\theta_{ki}|\theta_{(-k)i}, u_i)$, for $k = 1, \ldots, p$. Here

$$f^*(\theta_{ki}|\theta_{(-k)i}, u_i) = \frac{\alpha_{ui} G_0(\theta_i, A_{ui}) + \sum_{l(\theta_j)>u_i} \delta_{\theta_j}(\theta_i)}{\alpha_{ui} G_0(\theta_{(-k)i}, A_{ui}) + \sum_{l(\theta_j)>u_i} I(\theta_{(-k)i} = \theta_{(-k)j})},$$

where $\alpha_{ui} = \alpha G_0(A_{ui})$ and $G_0(\theta_{(-k)i}, A_{ui}) = \int G_0(\theta_i, A_{ui}) d\theta_{ki}$.

Example 3

Here we consider a probit model for which

$$y_i|\beta_i \sim \text{binomial}(\Phi(\beta_{0i} + \beta_{1i} z_i), n_i), \quad i = 1, \ldots, n,$$

where $\Phi(.)$ represents the standard normal distribution function. The full conditional distribution for β_i is given by

$$f^*(\beta_i) \propto \left\{\Phi(\beta_{0i} + \beta_{1i} z_i)^{y_i} (1 - \Phi(\beta_{0i} + \beta_{1i} z_i))^{n_i - y_i}\right\} \left\{\alpha G_0(\beta_i) + \sum_{j \neq i} \delta_{\beta_j}(\beta_i)\right\}.$$

and we assume $G_0(\beta_i) = N(\beta_i|\mu, \Sigma)$. We introduce the latent variables (u_i, v_i) such that their joint density with β_i is given, up to a constant of proportionality, by

$$f^*(\beta_i, u_i, v_i) \propto I\left\{u_i < \Phi(\beta_{0i} + \beta_{1i} z_i)^{y_i}, v_i < \{1 - \Phi(\beta_{0i} + \beta_{1i} z_i)\}^{n_i - y_i}\right\}$$

$$\times \left\{\alpha G_0(\beta_i) + \sum_{j \neq i} \delta_{\beta_j}(\beta_i)\right\}.$$

The full conditional for β_{0i} is given by

$$f^*(\beta_{0i}|\beta_{1i}, u_i, v_i) \propto I(a_i < \beta_{0i} < b_i) \left\{\alpha N(\beta_i|\mu, \sigma) + \sum_{\beta_{1j} = \beta_{1i}} \delta_{\beta_{0j}}(\beta_{0i})\right\}.$$

and the full conditional for β_{1i} is given by

$$f^*(\beta_{1i}|\beta_{0i}, u_i, v_i) \propto I(c_i < \beta_{1i} < d_i) \left\{\alpha N(\beta_i|\mu, \sigma) + \sum_{\beta_{0j} = \beta_{0i}} \delta_{\beta_{1j}}(\beta_{1i})\right\},$$

where $a_i = \Phi^{-1}(\tau_i) - \beta_{1i}z_i$, $b_i = (\Phi^{-1}(\tau_i) - \beta_{0i})/z_i$, $c_i = \Phi^{-1}(\lambda_i) - \beta_{1i}z_i$, and $d_i = (\Phi^{-1}(\lambda_i) - \beta_{0i})/z_i$, with $\tau_i = u_i^{1/y_i}$ and $\lambda_i = 1 - v_i^{1/(n_i-y_i)}$. Only minor modifications are required if $y_i = 0$, $y_i = n_i$ or $z_i = 0$.

Provided we can implement the algorithm for the parametric model $z_i|\theta_i \sim f(.\,\theta_i)$ with $\theta_i \sim_{\text{iid}} G_0$ using the latent variable idea then we can also implement it for the corresponding MDP model. Damien and Walker (1996) show that such an algorithm is applicable for a large class of nonconjugate models. This is detailed in the following Theorem whose proof is omitted since it is similar to the theorem appearing in Damien and Walker (1996) from which further details and examples can be found.

THEOREM. *If*

$$l(\theta) = \prod_{k=1}^{K} l_k(\theta),$$

where the $l_k(\theta)$ are nonnegative invertible functions, that is, if $l_k(\theta) > u$ then it is possible to obtain the set $A_{ku} = \{\theta : l_k(\theta) > u\}$, then it is possible to implement a Gibbs sampler for generating random variates from the MDP model in which all the full conditionals are of known types.

Note that the 'new' $q_{i0} = \alpha G_0(\cap_{k=1}^{K} A_{ku_i})$.

3 Numerical Example

We use the same example as that used by MacEarchen at al. (1996) who demonstrate their "next generation" sequential importance sampling technique in that paper. The data comes from Beckett and Diaconis (1994) and consists of 320 binomial experiments where 320 thumbtacks were flicked nine times each. The data for each experiment was the number of times the tack landed point up. That is, for $i = 1, \ldots, 320$, $y_i|p_i \sim \text{binomial}(p_i, 9)$. Therefore taking $G_0 = U(0,1)$ and introducing latent variables (u_i, v_i) we have

$$f^*(p_i, u_i, v_i) \propto I\left(u_i < p_i^{y_i}, v_i < (1-p_i)^{9-y_i}\right)\left\{\alpha I(0 < p_i < 1) + \sum_{j\neq i}\delta_{p_j}(p_i)\right\}.$$

Therefore

$$f^*(p_i|u_i, v_i) \propto \alpha I(a_i < p_i < b_i) + \sum_{a_i < p_j < b_i} \delta_{p_j}(p_i).$$

where

$$a_i = \begin{cases} u_i^{1/y_i} & \text{if } y_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

and

$$b_i = \begin{cases} 1 - v_i^{1/(9-y_i)} & \text{if } y_i < 9 \\ 1 & \text{otherwise.} \end{cases}$$

Also $f^*(u_i) = U(0, p_i^{y_i})$ and $f^*(v_i) = U(0, (1-p_i)^{9-y_i})$. These full conditionals are sampled straightforwardly. We implement the Gibbs sampler with α fixed at four different values (0.1, 1, 5, 10). Within each iteration we generate a sample from the distribution of $p_{n+1}|$data to compare with the analysis of MacEarchen at al. (1996). We collect 5000 samples for each of the four analyses which takes several seconds. The estimates for the predicitve density of $p_{n+1}|$data for the four values of α are shown in Figure 1. These compare very well with the results of MacEarchen at al. (1996, Figure 2).

4 Conclusions

The MDP model is very useful in a variety of applications. With the increasing use of the Gibbs sampler in Bayesian analysis it is necessary to have easy and fast ways of generating random variates from awkward conditional distributions. For the MDP model several researchers referred to in this paper have pointed out a serious computational hurdle in implementing the Gibbs sampler. In this paper, we have provided a general solution to the problem which bypasses the computational difficulty.

References

Beckett, L. & Diaconis, P. (1994). Spectral analysis for discrete longitudinal data. *Adv. in Math.* **103**, 107-128.

Bush, C.A. & MacEarchen, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275-286.

Damien,P. & Walker,S.G. (1996). Sampling probability densities via uniform random variables and a Gibbs sampler. *Submitted for publication.*

Escobar,M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Assoc.* **89**, 268-277.

Escobar,M.D. & West,M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577-588.

Ferguson,T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-230.

MacEarchen,S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist. Simul. and Comp.* **23**, 727-741.

MacEarchen,S.N. & Mueller,P. (1994). Estimating mixtures of Dirichlet process models.

MacEarchen, S.N.,Clyde,M. & Liu,J.S. (1996). Sequential importance sampling for nonparametric Bayes models: The next generation.

Mueller,P.,Erkanli,A. & West,M. (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.

Smith,A.F.M. & Roberts,G.O. (1993). Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55**, 3-23.

Tierney,L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701-1762.

West,M.,Mueller,P. & Escobar,M.D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of uncertainty: A tribute to D.V.Lindley*. eds A.F.M.Smith and P.Freeman. 363-368.

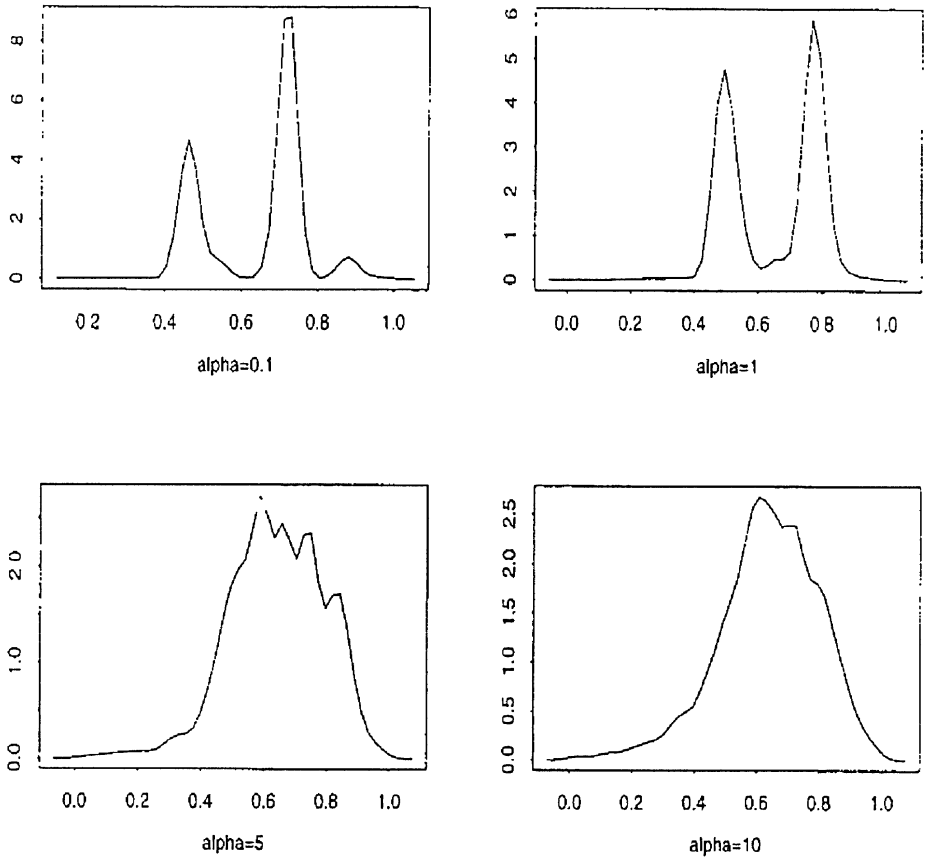

Figure 1: Estimates of the densities for p_{n+1} given the data for four vlaues of α.