Biometrika (2000), **87,** 2, *pp.* 371–390

Printed in Great Britain

Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models

BY HEMANT ISHWARAN

Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, Ohio 44195, U.S.A.

ishwaran@bio.ri.ccf.org

AND MAHMOUD ZAREPOUR

Department of Mathematics and Statistics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

zarepour@expresso.mathstat.uottawa.ca

SUMMARY

We present some easy-to-construct random probability measures which approximate the Dirichlet process and an extension which we will call the beta two-parameter process. The nature of these constructions makes it simple to implement Markov chain Monte Carlo algorithms for fitting nonparametric hierarchical models and mixtures of nonparametric hierarchical models. For the Dirichlet process, we consider a truncation approximation as well as a weak limit approximation based on a mixture of Dirichlet processes. The same type of truncation approximation can also be applied to the beta two-parameter process. Both methods lead to posteriors which can be fitted using Markov chain Monte Carlo algorithms that take advantage of blocked coordinate updates. These algorithms promote rapid mixing of the Markov chain and can be readily applied to normal mean mixture models and to density estimation problems. We prefer the truncation approximations, since a simple device for monitoring the adequacy of the approximation can be easily computed from the output of the Gibbs sampler. Furthermore, for the Dirichlet process, the truncation approximation offers an exponentially higher degree of accuracy over the weak limit approximation for the same computational effort. We also find that a certain beta two-parameter process may be suitable for finite mixture modelling because the distinct number of sampled values from this process tends to match closely the number of components of the underlying mixture distribution.

Some key words: Almost sure truncation; Generalised Dirichlet distribution; Mixture of Dirichlet processes; Nonparametric hierarchical model; Normal mean mixture; Random probability measure; Weak convergence in distribution.

1. INTRODUCTION

1·1. *Nonparametric hierarchical models*

We will discuss how to implement efficient Markov chain Monte Carlo algorithms for fitting the posterior distribution of a Bayesian nonparametric hierarchical model with the

following structure:

$$\begin{aligned}(X_i|Y_i) &\sim \pi(X_i|Y_i) \quad (i=1,\ldots,n),\\ (Y_i|P) &\sim P,\\ P &\sim \mathscr{P}_N.\end{aligned} \tag{1}$$

In this nonparametric setting, $X=(X_1,\ldots,X_n)$ is the observed data while $Y_1,\ldots,Y_n$ are unobserved random elements taking values in the measurable space $(\mathscr{Y},\mathscr{B})$, where $\mathscr{Y}=\Re^d$, in the examples considered here, and $\mathscr{B}$ is the corresponding Borel σ-algebra. In (1) it is assumed that the X_i are conditionally independent given the Y_i, while the Y_i conditioned on P are independently and identically distributed with distribution P. The nonparametric hierarchical model is a nonparametric method for modelling $\pi(X_i|Y_i)$, the conditional distribution of X_i given Y_i, by modelling the distribution of the Y_i through a random probability measure $\mathscr{P}_N$. We note that this model and the methods discussed in this paper can also be easily extended to the semiparametric hierarchical model formed by introducing a finite-dimensional parameter θ, where $\pi(X_i|Y_i,\theta)$ is the conditional distribution of X_i given Y_i and θ.

Our interest will focus on models (1) involving random probability measures $\mathscr{P}_N$ which closely approximate either the Dirichlet process or a generalisation which we refer to as a beta two-parameter process. These approximations are based on constructive sum-representations, and are the key to the success of our Markov chain Monte Carlo algorithms. In particular, the $\mathscr{P}_N$ in (1) are random probability measures of the form

$$\mathscr{P}_N(.)=\sum_{k=1}^{N} p_k\delta_{Z_k}(.) \quad (1\leqslant N<\infty), \tag{2}$$

where we write $\delta_Z(.)$ to denote a discrete measure concentrated at Z. Moreover, the p_k in (2) are random variables chosen to be independent of Z_k and constructed so that $0\leqslant p_k\leqslant 1$ and $p_1+\ldots+p_N=1$ with probability one, while the Z_k are independently and identically distributed random elements defined over $(\mathscr{Y},\mathscr{B})$ with distribution H.

The limit as $N\to\infty$ in (2) will correspond, in various forms of convergence, to a random probability measure $\mathscr{P}_\infty$, which in our setting will be either the Dirichlet process or more generally a beta two-parameter process; both of these have constructive sum-representations of the form

$$\mathscr{P}_\infty(.)=\sum_{k=1}^{\infty} p_k\delta_{Z_k}(.), \tag{3}$$

where

$$p_1=V_1, \quad p_k=(1-V_1)(1-V_2)\ldots(1-V_{k-1})V_k \quad (k\geqslant 2), \tag{4}$$

and where $V_1, V_2, \ldots$ are independent Be(a, b) random variables. We refer to (3) as a beta two-parameter process, writing this as $\mathscr{P}_\infty=\mathscr{B}(a, b, H)$.

One focus of the paper will be to look at different limiting approximations to the Ferguson (1973) Dirichlet process, which as mentioned has a construction of the form (3). In particular, if V_k are independent Be$(1, \alpha)$ variables, then the $\mathscr{P}_\infty$ formed by (3) and (4) is the Dirichlet process with concentration parameter $\alpha>0$ and reference distribution H (Sethuraman, 1994). We write this as $\mathscr{P}_\infty=\mathrm{DP}(\alpha H)$, or alternatively as $\mathscr{P}_\infty=\mathscr{B}(1, \alpha, H)$.

In this case, for each measurable partition $B_1, \ldots, B_d$ of $\mathscr{Y}$,

$$(\mathscr{P}_\infty(B_1), \ldots, \mathscr{P}_\infty(B_d)) \sim \mathrm{Dir}\{\alpha H(B_1), \ldots, \alpha H(B_d)\},$$

which is one way to characterise the Dirichlet process (Ferguson, 1973).

1·2. *Markov chain Monte Carlo with blocked updates*

The key to working with random probability measures like (2) is that it allows us to perform blocked updates for $p_1, \ldots, p_N$ and $Z_1, \ldots, Z_N$ in our Gibbs sampler. This will lead to a rapidly mixing Markov chain, and moreover it permits direct inference for the posterior $\mathscr{P}_N^* = (\mathscr{P}_N \mid X)$. This is in contrast to the usual method of fitting the nonparametric hierarchical model involving the Dirichlet process, where the standard practice is to integrate over P in order to exploit the Blackwell & MacQueen (1973) Pólya urn characterisation of the Dirichlet process. In particular, with a $\mathrm{DP}(\alpha H)$ prior, this method leads to a marginalised version of the nonparametric hierarchical model (1):

$$(X_i \mid Y_i) \sim \pi(X_i \mid Y_i) \quad (i = 1, \ldots, n),$$
$$(Y_1, \ldots, Y_n) \sim \pi_\infty(Y_1, \ldots, Y_n),$$

where

$$\pi_\infty(dY_1, \ldots, dY_n) = H(dY_1) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} H(dY_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{Y_j}(dY_i) \right\}$$

is the distribution for a generalised Pólya urn scheme.

This clever trick for exploiting the Pólya urn connection can be harnessed within a powerful Markov chain Monte Carlo setting, and was first discovered by Escobar (1994) with further details appearing in Escobar & West (1995). Refinements to the algorithm have been given by MacEachern (1994), West, Müller & Escobar (1994), MacEachern (1998) and MacEachern & Müller (1998). The Escobar–West–MacEachern Gibbs sampling method is a versatile Markov chain Monte Carlo approach for applying the Dirichlet process in modern Bayesian settings. However, it suffers from two limitations. First, by marginalising $\mathscr{P}_\infty$, the resulting Markov chain tends to mix slowly because of the Gibbs sampler use of one-coordinate-at-a-time updates. This can occur even in the sophisticated algorithms proposed by MacEachern (1994), which require one-at-a-time updates for cluster indicator variables. A second limitation arises from the effect of marginalising P. Although marginalising is the key underlying the Pólya urn approach, it has the undesirable side effect that it allows inference for the posterior of $\mathscr{P}_\infty$ to be based only on the values Y_i.

One way to combat these problems for the Dirichlet process, or more generally for the beta two-parameter process, is to use approximating random probability measures $\mathscr{P}_N$. This leads to a nonparametric hierarchical model almost indistinguishable from its limit, and a posterior $\mathscr{P}_N^*$ that yields approximate inference for the limiting posterior $\mathscr{P}_\infty^*$. The trick for making this all happen will be to find random probability measures $\mathscr{P}_N$ which provide good approximations to their limits $\mathscr{P}_\infty$, and selected so that it is possible to perform a simple multivariate update for $p_1, \ldots, p_N$ in the Gibbs sampler. Sections 3 and 4 will discuss two different approximations that satisfy these criteria by exploiting the conjugacy of the generalised Dirichlet distribution to multinomial sampling. We note for reference that other more general forms of conjugacy to the multinomial can also be explored, such as those presented in Hjort (1996). Also, for other methods which have

utilised approximations to the Dirichlet process; see Muliere & Tardella (1998) and Muliere & Secchi (1996).

We mention that the methods described in this paper can also be extended to include the two-parameter Poisson–Dirichlet process described by Pitman & Yor (1997). In this case, the relevant random probability measure $\mathscr{P}_\infty$ is also indexed by two parameters a and b, but is constructed using independent random variables $V_k \sim \text{Be}(1-a, b+ka)$, where $0 \leqslant a < 1$ and $b > -a$. As noted by Pitman & Yor (1997), the case where $a = 0$ and $b = \alpha$ corresponds to the Dirichlet process $\text{DP}(\alpha H)$, while the case $a = \alpha$ and $b = 0$ yields a process based on a stable law with index $0 < \alpha < 1$.

The layout of the paper is as follows. In § 3 we present a weak limit approximation to the Dirichlet process based on mixtures of Dirichlet processes. The same type of approximation has also been recognised by Walker & Wakefield (1998), by R. M. Neal in a University of Toronto technical report and by P. J. Green and S. Richardson in an unpublished manuscript. However, none of these papers exploits the connection to random probability measures in their Markov chain Monte Carlo algorithms. In § 4 we consider a different type of approximation based on a truncation approach which leads to an almost sure beta two-parameter process limit. We also discuss the accuracies of our different approximations; see Theorems 1 and 2. We begin our discussion in § 2 with a general method for the Gibbs sampling of nonparametric hierarchical models with random probability measures based on finite N.

2. GIBBS SAMPLING WITH RANDOM PROBABILITY MEASURES $\mathscr{P}_N$ WITH $N < \infty$

2·1. *Random variable description*

The trick for achieving efficient Markov chain Monte Carlo sampling of the nonparametric hierarchical model is to recast the model completely in terms of random variables. Let $p = (p_1, \ldots, p_N)$ and $Z = (Z_1, \ldots, Z_N)$. Model (1) can be rewritten as

$$\begin{aligned} (X_i \mid Z, K) &\sim \pi(X_i \mid Z_{K_i}), \\ (K_i \mid p) &\sim \sum_{k=1}^{N} p_k \delta_k(.), \\ (p, Z) &\sim \pi(p)\pi(Z), \end{aligned} \tag{5}$$

where $K = (K_1, \ldots, K_n)$ and the K_i are conditionally independent classification variables that identify the Z_k associated with each Y_i. Specifically, note that $Y_i = Z_{K_i}$.

By rewriting the model as (5), we can devise a Gibbs sampling scheme for exploring the posterior $\mathscr{P}_N^*$. To implement the Gibbs sampler we iteratively draw values from the following conditional distributions:

$$\pi(p, Z \mid K, X), \tag{6}$$

$$\pi(K \mid p, Z, X). \tag{7}$$

This method produces values drawn from the distribution $\pi(p, Z, K \mid X)$ and each draw $(p^{(b)}, Z^{(b)}, K^{(b)})$ produces a random probability measure $\mathscr{P}_N^{(b)}(.) = \sum_{k=1}^{N} p_k^{(b)} \delta_{Z_k^{(b)}}(.)$, which is a draw from the posterior $\mathscr{P}_N^*$. Thus, $\mathscr{P}_N^{(b)}$ can be used to estimate $\mathscr{P}_N^*$ and its functionals directly.

To determine the conditional density corresponding to (6), first observe from (5) that

$$f(p, Z \mid K, X) \propto \{f(K \mid p) f(p)\} \left\{ f(Z_K) \prod_{i=1}^{n} f(X_i \mid Z_{K_i}) \right\} f(Z^K), \tag{8}$$

where Z^K corresponds to those values in Z excluding $Z_K = (Z_{K_1^*}, \ldots, Z_{K_m^*})$, and where $K_1^*, \ldots, K_m^*$ represent the unique set of K_i values. The third term on the right-hand side of (8) is the product density for Z^K, which we can easily sample by drawing independent values from H. Therefore, we only need to work out the remaining first and second terms; the conditional densities for p and Z_K.

In §§ 3–4 we will provide two different approximations, one a weak limit approximation to the Dirichlet process and the other a truncation of the beta two-parameter process $\mathscr{B}(a, b, H)$, which will allow us to sample the conditional density of p exactly:

$$f(p \mid K) \propto f(K \mid p) f(p). \tag{9}$$

In general, to ensure a rapidly mixing Markov chain, the trick is to choose $\mathscr{P}_N$ so as to provide a good approximation to its limit $\mathscr{P}_\infty$, while leading to a simple expression for (9).

To complete the update for (6) we still need to derive the conditional distribution for Z_K. From (8), its density can be rewritten as

$$f(Z_K \mid K, X) \propto \prod_{j=1}^{m} f(Z_{K_j^*}) \prod_{\{i: K_i = K_j^*\}} f(X_i \mid Z_{K_j^*}). \tag{10}$$

In the examples to be considered in §§ 3–4, conjugacy will allow us to sample (10) exactly. In general, however, the distribution could be sampled fairly efficiently using a Metropolis–Hastings step for each $Z_{K_j^*}$. Alternatively, the Metropolis step could be replaced by a step using hybrid Monte Carlo, which could substantially improve mixing; see Neal (1996, pp. 55–63), Gustafson (1997), Daniels (1998) and Ishwaran (1999) for some recent statistical applications. It is important to note that the non-conjugacy is a much more delicate issue in the Escobar–West–MacEachern Gibbs sampler. See West et al. (1994), MacEachern & Müller (1998) or Walker & Damien (1998) for different approaches to this problem.

The last step in the Gibbs sampler involves the conditional distribution (7) for K. However, it is obvious that

$$(K_i \mid p, Z, X) \sim \sum_{k=1}^{N} p_{k,i}^* \delta_k(.)$$

are conditionally independent integers, where

$$(p_{1,i}^*, \ldots, p_{N,i}^*) \propto (p_1 f(X_i \mid Z_1), \ldots, p_N f(X_i \mid Z_N)). \tag{11}$$

2·2. *Mixtures of nonparametric hierarchical models*

Greater model flexibility can be introduced by mixing over the random probability measure $\mathscr{P}_N$, giving a mixture of nonparametric hierarchical models. The mixing is introduced through the hyperparameters γ for Z and α for p. In this case, the model (5), and hence (1), is extended by assuming that

$$(p, Z \mid \alpha, \gamma) \sim \pi(p \mid \alpha) \pi(Z \mid \gamma),$$

$$(\alpha, \gamma) \sim \pi(\alpha) \pi(\gamma).$$

The Gibbs sampler of the previous section is easily extended to this case. We present this extension through several examples in the following sections.

3. WEAK LIMIT REPRESENTATIONS FOR THE DIRICHLET PROCESS

3·1. *Dirichlet random weights*

A simple and practical approximation to the Dirichlet process can be defined using a random probability measure (2) with random probabilities p that have the Dirichlet distribution

$$(p \mid \alpha) \sim \mathrm{Dir}(\alpha/N, \ldots, \alpha/N). \tag{12}$$

Working with a Dirichlet distribution is extremely convenient because of its conjugacy to the multinomial distribution. In particular, it easily follows that the conditional distribution (9) for p is a Dirichlet distribution with an updated parameter depending only upon the number of occurrences of K_i. In particular, if $m_k = \mathrm{card}\{K_i = k\}$ is the number of K_i's which equal k, then

$$(p \mid K, \alpha) \sim \mathrm{Dir}(\alpha/N + m_1, \ldots, \alpha/N + m_N). \tag{13}$$

Therefore, this choice for $\mathscr{P}_N$ satisfies one of our conditions, namely that it leads to a simple multivariate update for the conditional distribution of p. It also satisfies our second criterion, which is that it provides a good approximation to its limit. In fact, $\mathscr{P}_N$ converges weakly in distribution to the Dirichlet process $\mathscr{P}_\infty = \mathrm{DP}(\alpha H)$. Moreover, it can be shown that

$$\mathscr{L}\{\mathscr{P}_N(g)\} \to \mathscr{L}\{\mathscr{P}_\infty(g)\}$$

for each real-valued measurable function g that is integrable with respect to H. Thus, $\mathscr{P}_N$ offers a strong approximation to the Dirichlet process and can be used to approximate integrable functionals of the process; see the unpublished report by H. Ishwaran and M. Zarepour 'Exact and approximate sum-representations for the Dirichlet process' for details.

It is easy to see that $\mathscr{P}_N$ is a mixture of Dirichlet processes, written in the style of Antoniak (1974) as

$$\mathscr{P}_N(.) := \int \mathrm{DP}\{\alpha\xi_N(Z, .)\}\, dH^N(Z),$$

where

$$\xi_N(Z, .) = \frac{1}{N}\sum_{k=1}^{N} \delta_{Z_k}(.)$$

is the empirical measure based on Z. Therefore, intuitively we expect that $\mathscr{P}_N \doteq \mathscr{P}_\infty$ for large N, because $\xi_N \doteq H$. In fact, the sampling behaviours of $\mathscr{P}_N$ and $\mathscr{P}_\infty$ are quite similar asymptotically. Consider the following theorem, detailed in the unpublished report by H. Ishwaran and M. Zarepour, which compares the number of distinct Y_i values.

THEOREM 1. *Let C_N and C_∞ equal the number of distinct values in Y, where $Y = (Y_1, \ldots, Y_n)$ is a sample obtained under $\mathscr{P}_N$ and $\mathscr{P}_\infty = \mathrm{DP}(\alpha H)$ respectively. If H is nonatomic,*

then

$$\frac{N!}{N^k(N-k)!} \leqslant \frac{\text{pr}\{C_N = k\}}{\text{pr}\{C_\infty = k\}} \leqslant n^{\alpha k/N} \quad (k = 1, \ldots, \min(n, N)). \tag{14}$$

Note that the two distributions agree in the limit as $N \to \infty$ because the right-hand and left-hand sides of (14) both converge to one for each value of k. Although the bound is quite crude, it does work fairly well in the range $k \leqslant \log n$, which is roughly the number of distinct values we would expect to see under the two models when $N = n$. However, when N is small and n is large, the expected number of distinct values under $\mathscr{P}_N$ will be much smaller than under the Dirichlet process model if H is nonatomic. This is because each value Y_i from $\mathscr{P}_N$ is sampled from at most N distinct values in contrast to the continuum of values available under the Dirichlet process. With a large enough sample size n this will lead to relatively few distinct values.

3·2. *Normal mean mixtures*

We first apply the weak limit approximation to normal mean mixture models. These are nonparametric hierarchical models expressible in the form (1), where

$$(X_i | Y_i) \sim \mathscr{N}(Y_i, \sigma_X),$$

and $\sigma_X > 0$ is a known variance; later we will consider the case when σ_X is unknown. To extend this model to a mixture of hierarchical models we introduce a prior for α in p and include hyperparameters for Z. If we use representation (5), the model we consider is

$$\begin{aligned}
(X_i | Z, K) &\sim \mathscr{N}(Z_{K_i}, \sigma_X), \\
(K_i | p) &\sim \sum_{k=1}^{N} p_k \delta_k(.), \\
(Z_k | \theta, \sigma_Z) &\sim \mathscr{N}(\theta, \sigma_Z), \\
(\theta | \sigma_\theta) &\sim \mathscr{N}(0, \sigma_\theta), \\
(\sigma_Z^{-1} | \tau_1, \tau_2) &\sim \text{Ga}(\tau_1, \tau_2), \\
(\alpha | v_1, v_2) &\sim \text{Ga}(v_1, v_2),
\end{aligned} \tag{15}$$

with the distribution for p specified by (12).

Model (15) uses a conjugate normal prior for θ and a conjugate inverse-gamma prior for σ_Z. To ensure that these priors are noninformative, we choose a large value for σ_θ, that is $\sigma_\theta = 1000$, and we select small values for the hyperparameters τ_1 and τ_2, that is $\tau_1 = \tau_2 = 0{\cdot}001$. Selecting an appropriate prior for α is critical to the model's performance, since the value for α is directly related to the number of distinct Y_i values. We use a $\text{Ga}(v_1, v_2)$ prior, which has been used by Escobar & West (1995) in density estimation problems involving the Dirichlet process. A gamma prior is appropriate because of its flexibility. For example, to discourage small and large values for α we choose large values for both v_1 and v_2. Selecting a large-scale parameter v_2 is especially relevant to finite mixture modelling since it encourages repetitions in the Y_i and can be used as a tool for studying the number of mixture components. In the case of the Dirichlet process, the use of a gamma prior has the added feature that it allows for exact sampling of α; see Escobar & West (1995, 1998) for details. Unfortunately, we will not be able to exploit this clever trick with the weak limit approximation, and instead we will need to resort to a Metropolis–Hastings step.

Introducing the hyperparameters in (15) is a method for extending the nonparametric hierarchical model by allowing for mixtures of random probability measures. In particular, the $\mathscr{P}_N$ defined by (12) and (15) is the mixture of Dirichlet processes

$$\mathscr{P}_N(.) := \iiiint \mathrm{DP}\{\alpha\xi_N(Z,.)\}\, dH^N(Z\,|\,\theta,\sigma_Z)\, d\pi(\theta)\, d\pi(\sigma_Z)\, d\pi(\alpha),$$

where $H^N(.\,|\,\theta,\sigma_Z)$ is the distribution for Z conditioned on θ and σ_Z.

To sample the posterior $\pi(p, Z, K, \theta, \sigma_Z, \alpha\,|\,X)$ of (15) using Gibbs sampling, we need to complete five updates in each cycle of the sampler:

$$(p, Z\,|\,K, \theta, \sigma_Z, \alpha, X), \tag{16}$$

$$(K\,|\,p, Z, X), \tag{17}$$

$$(\theta\,|\,Z, \sigma_Z), \tag{18}$$

$$(\sigma_Z\,|\,Z, \theta), \tag{19}$$

$$(\alpha\,|\,p). \tag{20}$$

As worked out in (13), we have

$$(p\,|\,K, \alpha) \sim \mathrm{Dir}(\alpha/N + m_1, \ldots, \alpha/N + m_N).$$

Therefore, to complete the update (16), we need to work out the conditional distribution for Z, which by similar reasoning to (8) has the density

$$f(Z\,|\,K, \theta, \sigma_Z, X) \propto \left\{\prod_{j=1}^{m} f(Z_{K_j^*}\,|\,\theta, \sigma_Z) \prod_{\{i:\,K_i = K_j^*\}} f(X_i\,|\,Z_{K_j^*})\right\} f(Z^K\,|\,\theta, \sigma_Z).$$

The second term involves sampling independent normal variables, but so does the first term because of conjugacy. In fact, the first term corresponds to the product of conditional normals

$$(Z_{K_j^*}\,|\,K, \theta, \sigma_Z, X) \sim \mathscr{N}(\theta_j^*, \sigma_{Z_j}^*),$$

where

$$\theta_j^* = \sigma_{Z_j}^* \left(\theta/\sigma_Z + \sum_{\{i:\,K_i = K_j^*\}} X_i/\sigma_X\right),$$

$\sigma_{Z_j}^* = (n_j/\sigma_X + 1/\sigma_Z)^{-1}$, and n_j is the number of times K_j^* occurs in K.

The conditional distribution for K_i in (17) is determined using (11), which with a normal density corresponds to

$$(K_i\,|\,p, Z, X) \sim \sum_{k=1}^{N} p_{k,i}^* \delta_k(.),$$

where

$$(p_{1,i}^*, \ldots, p_{N,i}^*) \propto \left(p_1 \exp\left\{\frac{-1}{2\sigma_X}(X_i - Z_1)^2\right\}, \ldots, p_N \exp\left\{\frac{-1}{2\sigma_X}(X_i - Z_N)^2\right\}\right).$$

The distributions for θ and σ_Z in (18) and (19) are straightforward because of conjugacy. Indeed, it easily follows that $(\theta\,|\,Z, \sigma_Z, \sigma_\theta) \sim \mathscr{N}(\theta^*, \sigma_\theta^*)$, where

$$\theta^* = \frac{\sigma_\theta^*}{\sigma_Z} \sum_{k=1}^{N} Z_k, \quad 1/\sigma_\theta^* = N/\sigma_Z + 1/\sigma_\theta.$$

Also,

$$(\sigma_Z^{-1} \mid Z, \theta) \sim \mathrm{Ga}\left\{\tau_1 + N/2, \tau_2 + \sum_{k=1}^{N} (Z_k - \theta)^2/2\right\}.$$

This completes four of the five steps. The final step (20) involving α can be implemented through the use of random walk Metropolis–Hastings. Note that the Metropolis step will make use of the conditional density for α, which is

$$f(\alpha \mid p) \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha/N)^N} p_1^{\alpha/N-1} \ldots p_N^{\alpha/N-1} f(\alpha). \tag{21}$$

To illustrate our method, we simulated $n = 45$ observations from a normal mean mixture model with $\sigma_X = 1$ and with an underlying mixing distribution with support points $\{-3, 1, 2\}$ having equal probabilities. Figures 1 and 2 contain the results from the Gibbs sampler, where we have used a Ga(2, 2) prior for α and we have selected N to equal the sample size.

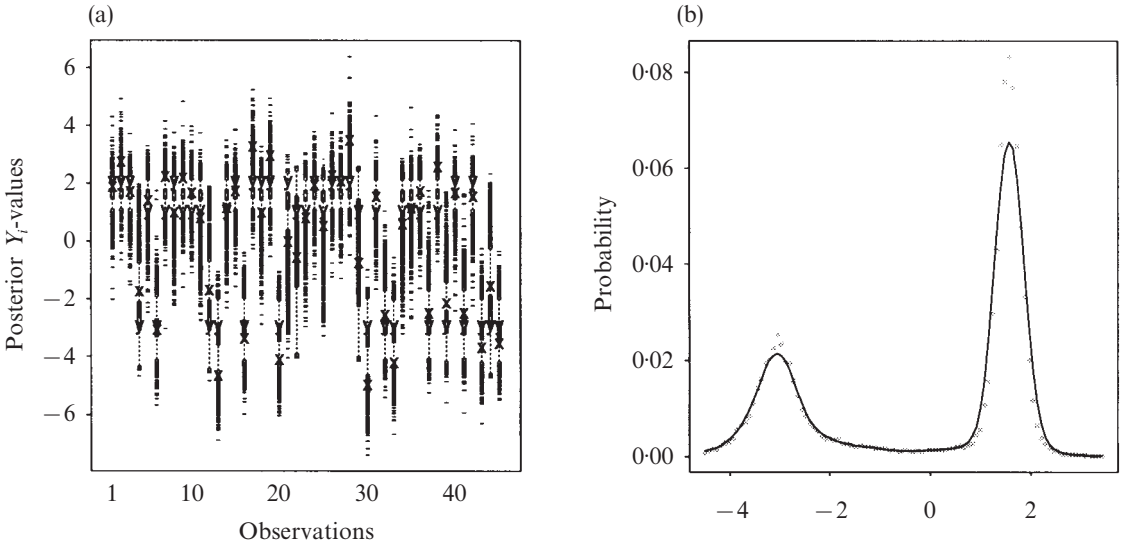

Fig. 1. (a) Posterior values for Y_i based on 45 values of X_i where each X_i was simulated from a normal mean mixture with $\sigma_X = 1$ and with a mixing distribution having support $\{-3, 1, 2\}$ with uniform probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Values for X_i and true values for Y_i are superimposed using the symbols x and y respectively. Plot is based on 4500 sampled values following an initial 2500 iteration burn-in using the Gibbs sampling scheme described in § 3·2 with $N = n = 45$ and $\alpha \sim \mathrm{Ga}(2, 2)$. The Metropolis step for α had approximately a 34% acceptance rate. (b) Averaged values for random measures $\mathscr{P}_N^{(b)}$ evaluated over a refined partition for $b = 1, \ldots, 4500$; the solid curve is a smoothed version of the individual points.

The posterior values $(p^{(b)}, Z^{(b)}, K^{(b)})$ and resulting random measures $\mathscr{P}_N^{(b)}$ obtained from the Gibbs sampler were used to study the finite-dimensional distribution of $\mathscr{P}_N^*$ evaluated over a refined partition $\{B_1, \ldots, B_d\}$ for $\mathscr{Y}$:

$$(\mathscr{P}_N^*(B_1), \ldots, \mathscr{P}_N^*(B_d)) = \left(\sum_{k=1}^{N} p_k^* \{Z_k^* \in B_1\}, \ldots, \sum_{k=1}^{N} p_k^* \{Z_k^* \in B_d\}\right). \tag{22}$$

For example, consider Fig. 1(b), which estimates (22) by averaging the different values of $\mathscr{P}_N^{(b)}$ evaluated over a refined partition. As we can see the estimate has uncovered two

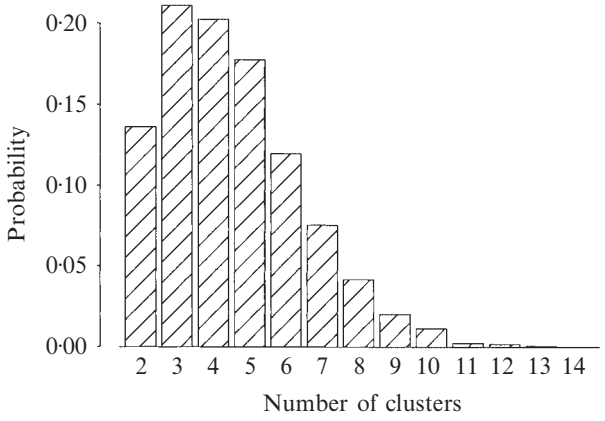

Fig. 2. Posterior distribution of the number of distinct Y_i values from the model of Fig. 1.

distinct modes with the posterior being unable to delineate between the two distinct values 1 and 2 of Y.

Another way to study the number of mixture components is to consider the posterior number of distinct Y_i values, although in general this method tends to overestimate the number of components. Looking at Fig. 2, we find that most of the posterior distribution is concentrated on anywhere from 2 to 6 clusters, thus presenting evidence for the presence of at least 2 components. This analysis agrees very closely with what we will see in § 4, when we revisit this example with a larger value for N, and hence a more accurate approximation to the Dirichlet process.

3·3. *Density estimation*

The normal mean mixture model has a simple extension to density estimation problems. By introducing σ_X as a parameter, the model can be used to estimate the unknown density for the observations X_i. In a classical setting, this would be similar to density estimation using a normal kernel with a bandwidth value $\sqrt{\sigma_X}$. In a Bayesian context, this is a semiparametric hierarchical model where $(X_i \mid Y_i, \sigma_X) \sim \mathcal{N}(Y_i, \sigma_X)$, and corresponds to the hierarchical model (1) when extended to include the finite-dimensional parameter σ_X. An especially convenient choice for a prior for σ_X is the inverse gamma,

$$(\sigma_X^{-1} \mid \gamma_1, \gamma_2) \sim \mathrm{Ga}(\gamma_1, \gamma_2),$$

where we select $\gamma_1 = \gamma_2 = 0{\cdot}001$ to yield a noninformative prior.

This model can be fitted using the same Gibbs sampler as before, although steps (16)–(20) must now include the value for σ_X. An additional step is needed for the conditional distribution for σ_X, but this is simple because of conjugacy:

$$(\sigma_X^{-1} \mid X, Z, K) \sim \mathrm{Ga}\left\{\gamma_1 + n/2, \gamma_2 + \sum_{i=1}^{n} (X_i - Z_{K_i})^2/2\right\}.$$

The output from the Gibbs sampler can be easily used to estimate a predictive density for a future observation X_{n+1}. Let $f(X_{n+1} \mid X)$ represent the predictive density for X_{n+1}

conditioned on the data X, and let Y_{n+1} be the corresponding unobserved Y value. Then

$$f(X_{n+1}|X) = \int f(X_{n+1}|Y_{n+1}, \sigma_X)\, d\pi(Y_{n+1}, \sigma_X|X)$$

$$= \iint f(X_{n+1}|Y_{n+1}, \sigma_X)\, d\pi(Y_{n+1}|P)\, d\pi(P, \sigma_X|X).$$

For a probability measure $P(.) = \sum_{k=1}^{N} p_k \delta_{Z_k}(.)$,

$$\int f(X_{n+1}|Y_{n+1}, \sigma_X)\, d\pi(Y_{n+1}|P) = \sum_{k=1}^{N} p_k f(X_{n+1}|Z_k, \sigma_X). \tag{23}$$

Consequently, $f(X_{n+1}|X)$ can be approximated by averaging the mixture of normal densities (23) over the sampled values $(p^{(b)}, Z^{(b)}, \sigma_X^{(b)})$ obtained from the Gibbs sampler. A predictive density estimate can then be derived by evaluating the averaged density over a refined partition.

To illustrate the method, we reanalysed the galaxy data in Roeder (1990), representing the relative velocities of $n = 82$ galaxies from six well-separated conic sections of space. The data have also been studied by Escobar & West (1995), using a Dirichlet process and the Escobar–West–MacEachern Gibbs sampling algorithm outlined in the Introduction. To allow our results to be more easily compared to theirs, we used their choice of a Ga(2, 4) prior for α.

Figure 3 represents the predictive density estimate (23) for the galaxy data obtained by our Gibbs sampling scheme with $N = n = 82$. Figure 3(a) shows the mean value averaged over 5 batches of 1000 sampled values, following an initial 2500 iteration burn-in. From Fig. 3(a) we see that there appear to be 5 or 6 distinct modes in our predictive density estimate. The number of distinct modes can also be studied by looking at the posterior

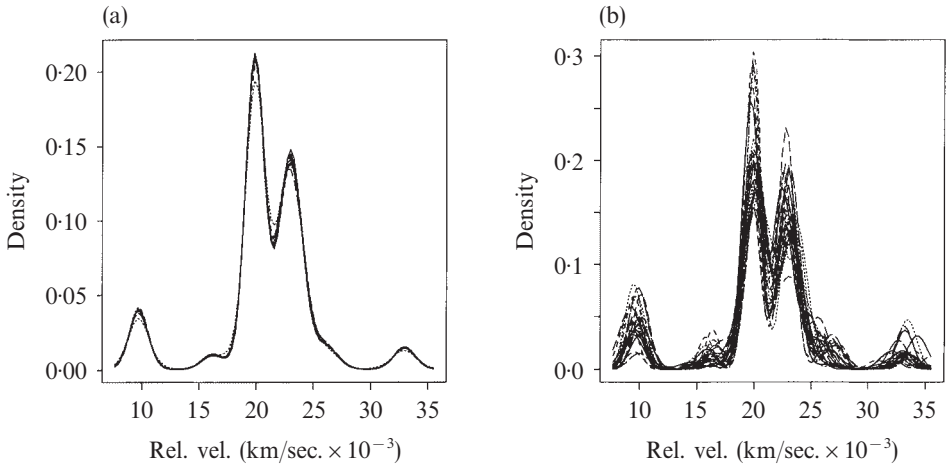

Fig. 3. Density estimation for relative velocities in thousands of kilometers/second for 82 galaxies (Roeder, 1990). (a) Posterior predictive density averaged over 5 different batches each of size 1000. (b) Twenty-five randomly selected posterior densities. All predictions are over the same refined partition. Posterior calculations are based on 5000 sampled values following an initial 2500 iteration burn-in using the Gibbs sampler outlined in §§ 3·2 and 3·3. The model used $N = n = 82$ and $\alpha \sim \text{Ga}(2, 4)$. The Metropolis step for α had a 36% acceptance rate.

distribution of C_N, the number of distinct Y_i values, although this only provides an upper bound. Looking at Table 1, we see that there are anywhere from 6 to 10 distinct clusters, with nearly half of the posterior probability on 7 or 8 clusters. These findings are similar to those seen in Escobar & West (1995, Table 5).

Table 1. *Posterior probabilities of the number of distinct Y_i values from Fig. 3 for the galaxy data*

k	$\leqslant 5$	6	7	8	9	10	11	12	>12
$\mathrm{pr}(C_N = k \mid X)$	0·01	0·12	0·24	0·24	0·18	0·11	0·06	0·02	0·02

The mixing behaviour of the Markov chain can be studied by considering the autocorrelations of the Y_i values from the Gibbs sampler. We re-ran the previous analysis using 3000 sampled values following an initial 1000 iteration burn-in. In this second run, we fixed α at 1·2, the mean value from our previous analysis, in order to facilitate a simpler comparison to the Escobar–West Gibbs sampling algorithm; we used the same Gibbs sampler described in Escobar & West (1995). The autocorrelations from both these Gibbs samplers are recorded in Fig. 4 and are based on the same priors for σ_X and for the mean θ and variance σ_Z used in the normal reference distribution H. Looking at Fig. 4, we find that our Gibbs sampler mixes well, with low autocorrelations for all parameters. This is in contrast to the Escobar–West algorithm, which contains at least one group of Y values with non-vanishing autocorrelations. This phenomenon is an inherent problem with the one-coordinate-at-a-time Pólya urn sampling method, which suppresses the ability of similar Y_i values to change as the sampler iterates. Thus, a value for Y_i can sometimes persist for many iterations. See MacEachern (1994), MacEachern & Müller (1998) and Escobar & West (1998) for more discussion and some suggested remedies.

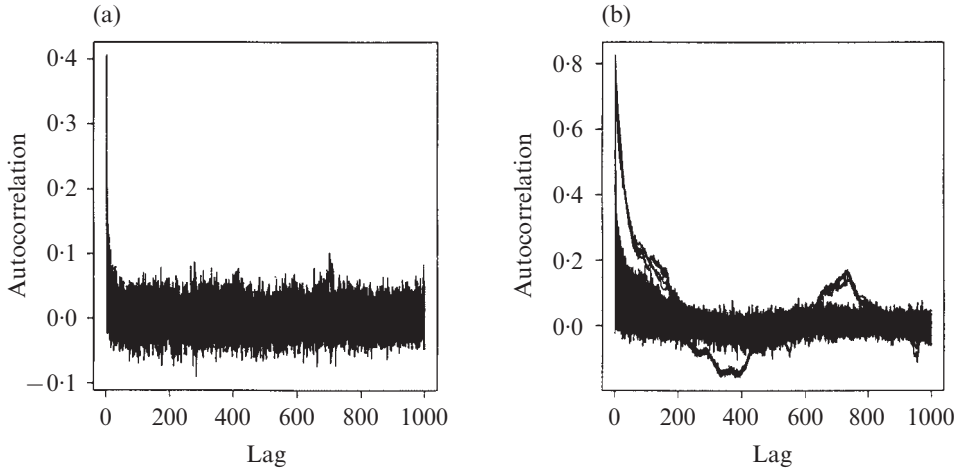

Fig. 4. Autocorrelations for each of the Y_i values from the galaxy data based on 3000 sampled values following a 1000 iteration burn-in. Models used the fixed value for $\alpha = 1{\cdot}2$. (a) Gibbs sampler of §§ 3·2 and 3·3, (b) Escobar–West Gibbs sampling algorithm.

Remark. Local bandwidth selection can also be easily accommodated by a simple extension to the normal mean mixture model. Local smoothing is introduced by modelling the mean and variance for X_i nonparametrically. In our previous notation, this would correspond to setting $Y_i = (\mu_i, \sigma_{X_i})$ and $(X_i \mid Y_i) \sim \mathcal{N}(\mu_i, \sigma_{X_i})$, with the distribution of Y_i a random

distribution over the space of distributions for the mean and variance. Our Gibbs sampler is easily modified to accommodate this setting; see Escobar & West (1995) and Müller, Erkanli & West (1996) for motivation and examples.

4. Beta two-parameter and Dirichlet process truncations

4·1. *Almost sure truncations*

An approximation with an almost sure $\mathscr{B}(a, b, H)$ limit can be obtained by truncating the higher-order terms in the sum-representation (3). This method gives an approximating random probability measure

$$\mathscr{P}_N(.) = V_1 \delta_{Z_1}(.) + \sum_{k=2}^{N} \{(1-V_1)(1-V_2)\ldots(1-V_{k-1})V_k\}\delta_{Z_k}(.), \tag{24}$$

where

$$p_1 = V_1, \quad p_k = (1-V_1)(1-V_2)\ldots(1-V_{k-1})V_k \quad (k=2,\ldots,N) \tag{25}$$

and the V_k are independent $\mathrm{Be}(a_k, b_k)$ random variables with $a_k = a$ and $b_k = b$, for $k \leqslant N-1$. We set $V_N = 1$ to ensure that $p_1 + \ldots + p_N = 1$ because

$$1 - \sum_{k=1}^{N-1} p_k = (1-V_1)\ldots(1-V_{N-1}).$$

In order to sample the posterior of the nonparametric hierarchical model associated with (24) efficiently, we need to be able to perform an efficient multivariate update for the conditional distribution of $p = (p_1, \ldots, p_n)$ as defined in (9). At the same time, however, we also need to ensure that the choice for N in the truncation leads to an adequate approximation of the beta two-parameter process.

4·2. *Exact updates for the conditional distribution of p*

In fact, for any a_k and b_k, the p defined by (25) has a generalised Dirichlet distribution, written, following the style of Walker & Muliere (1997) and Muliere & Walker (1998), as

$$p \sim \mathscr{G}(a_1, b_1, \ldots, a_{N-1}, b_{N-1}).$$

By Connor & Mosimann (1969), its density equals

$$\left\{\prod_{k=1}^{N-1} \frac{\Gamma(a_k+b_k)}{\Gamma(a_k)\Gamma(b_k)}\right\} p_1^{a_1-1}\ldots p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} \times (1-P_1)^{b_1-(a_2+b_2)}\ldots(1-P_{N-2})^{b_{N-2}-(a_{N-1}+b_{N-1})}, \tag{26}$$

where $P_k = p_1 + \ldots + p_k$. From this it easily follows that the distribution is conjugate for multinomial sampling, and consequently the conditional distribution (9) for p defined with $a_k = a$ and $b_k = b$ is $\mathscr{G}(a_1^*, b_1^*, \ldots, a_{N-1}^*, b_{N-1}^*)$, where

$$a_k^* = a + m_k, \quad b_k^* = b + \sum_{j=k+1}^{N} m_j = b + M_k \quad (k=1,\ldots,N-1)$$

and m_k is the number of K_i's which equal k, as before. Therefore, the conditional distribution for p can be sampled exactly by using the sampling scheme for the generalised Dirichlet distribution indicated by (25). This is very efficient, requiring simulation of only

$N-1$ beta random variables, and the simple computations required to determine $m_1, \ldots, m_N$. For example, if $\mathscr{P}_N$ is the truncated version of the DP(αH) process, then $a_k = 1$, $b_k = \alpha$ and the conditional distribution for p is a generalised Dirichlet distribution with parameters $a_k^* = 1 + m_k$ and $b_k^* = \alpha + M_k$.

4·3. *Selecting adequate truncation values N*

An adequate truncation level N can be determined by considering the behaviour of the higher-order p_k values in the sum-representation (3) for $\mathscr{P}_\infty$. The following theorem can be used as a guide for selecting N, as well as a diagnostic for assessing the adequacy for the truncation.

THEOREM 2. *For each positive integer $N \geqslant 1$ and each positive integer $r \geqslant 1$, let*

$$U_N(r) = \left(\sum_{k=N}^{\infty} p_k \right)^r, \quad W_N(r) = \sum_{k=N}^{\infty} p_k^r$$

for the p_k defined in (4). *Then*

$$E\{U_N(r)\} = \left\{ \frac{b^{(r)}}{(a+b)^{(r)}} \right\}^{N-1}, \tag{27}$$

$$E\{W_N(r)\} = \left\{ \frac{b^{(r)}}{(a+b)^{(r)}} \right\}^{N-1} \frac{a^{(r)}}{(a+b)^{(r)} - b^{(r)}}, \tag{28}$$

where $\gamma^{(r)} = \gamma(\gamma+1)\ldots(\gamma+r-1)$ for each $\gamma > 0$ and $\gamma^{(0)} = 1$.

Note that the expected values for $U_N(r)$ and $W_N(r)$ depend only upon the values of a, b and N, and therefore they can be evaluated from the output of our Gibbs sampler in the case when a and b are parameters. In particular, we can assess the adequacy of the truncation level N by estimating moments of the tail probability

$$U_N(1) = W_N(1) = \sum_{k=N}^{\infty} p_k.$$

One can test whether or not $U_N(1)$ is small enough from the Gibbs sampler output by evaluating its mean (27) as well as its variance

$$\text{var}\{U_N(1)\} = E\{U_N(2)\} - [E\{U_N(1)\}]^2.$$

Also note that the value for $W_1(r)$ can be used to study the sampling behaviour of $\mathscr{P}_\infty$. For example, with a nonatomic H,

$$\mathscr{P}_\infty\{Y_1 = \ldots = Y_r\} = E\{W_1(r)\}.$$

Proof of Theorem 2. First consider the case when $N = 1$. Obviously $U_1(r) = 1$, which leaves us to determine $W_1(r)$. By the method of construction (4) for the values p_k, it follows that, in distribution,

$$W_1(r) = V_1^r + (1 - V_1)^r W_1(r),$$

where, on the right-hand side, $W_1(r)$ is independent of V_1. Taking expectations and simplifying, we obtain

$$E\{W_1(r)\} = \frac{E(V_1^r)}{1 - E(1 - V_1)^r}.$$

The rth moment for $V_1 \sim \text{Be}(a, b)$ equals $a^{(r)}/(a+b)^{(r)}$. From this and the rth moment for $1 - V_1 \sim \text{Be}(b, a)$, deduce that

$$E\{W_1(r)\} = \frac{a^{(r)}}{(a+b)^{(r)} - b^{(r)}}.$$

To complete the theorem, note that, from (4),

$$U_N(r) = \{(1-V_1)\ldots(1-V_{N-1})\}^r U_1(r),$$
$$W_N(r) = \{(1-V_1)\ldots(1-V_{N-1})\}^r W_1(r),$$

where the two equalities are in distribution, and where, on the right-hand sides, both $U_1(r)$ and $W_1(r)$ are independent of $V_1, \ldots, V_{N-1}$. Take expectations to arrive at (27) and (28). □

4·4. *Truncation values for the* DP(αH) *and the* $\mathscr{B}(\alpha, 1, H)$ *processes*

Recall that the DP(αH) process is derived from the random probability measure construction (3) and (4) using independent $V_k \sim \text{Be}(1, \alpha)$ random variables. Therefore, if we use the identity $(1+\alpha)^{(r)} = \alpha^{(r)}(\alpha + r)/\alpha$, it follows from (27) that

$$E\{U_N(r)\} = \left(\frac{\alpha}{\alpha + r}\right)^{N-1}, \tag{29}$$

and from (28) that

$$E\{W_N(r)\} = \left(\frac{\alpha}{\alpha + r}\right)^{N-1} \frac{\Gamma(r)\Gamma(\alpha+1)}{\Gamma(\alpha+r)}.$$

Note that the tail moment (29) is an increasing function in α. This reflects the fact that, in the Dirichlet process, the expected number of distinct Y values is directly proportional to the concentration parameter α; see for example Korwar & Hollander (1973) or Antoniak (1974). In fact, the DP(αH) process converges weakly in distribution to H as $\alpha \to \infty$, and with a nonatomic H we are assured of a sample with all distinct values in the limit.

This is in contrast to the $\mathscr{B}(\alpha, 1, H)$ process, which is the random probability measure (3) constructed with independent $V_k \sim \text{Be}(\alpha, 1)$ variables. This reverses the role of the a and b parameters used in constructing the Dirichlet process. In particular, from (27), we have

$$E\{U_N(r)\} = \left\{\frac{1^{(r)}}{(\alpha+1)^{(r)}}\right\}^{N-1},$$

which shows that the number of distinct values in Y increases as $\alpha \to 0$ and decreases as $\alpha \to \infty$.

In applying a truncation to each of these processes, we have to be careful that the resulting random probability measure is still rich enough to model adequately the nonparametric hierarchical model. This can be tested by looking at the tail probability $U_N(1) = \sum_{k=N}^{\infty} p_k$ over a range of α values. Consider Fig. 5, which plots the mean and variance for $U_N(1)$ for various values of α and N. Figures 5(a), (b) for the Dirichlet process show that we can use values of α up to 10 and still have a negligible tail probability once N is reasonably large, such as $N = 50$. Since most applications will involve values of α in this range, we can expect an adequate truncation for reasonably large values of N. In

contrast, Figures 5(c), (d) for the $\mathscr{B}(\alpha, 1, H)$ process show that there can still be a substantial amount of tail probability for the range of values $0 < \alpha < 0{\cdot}1$, even for very large N. Therefore, if we expect a large number of distinct Y values, and hence a small α, it would appear prudent to choose a large value of N to ensure an adequate truncation.

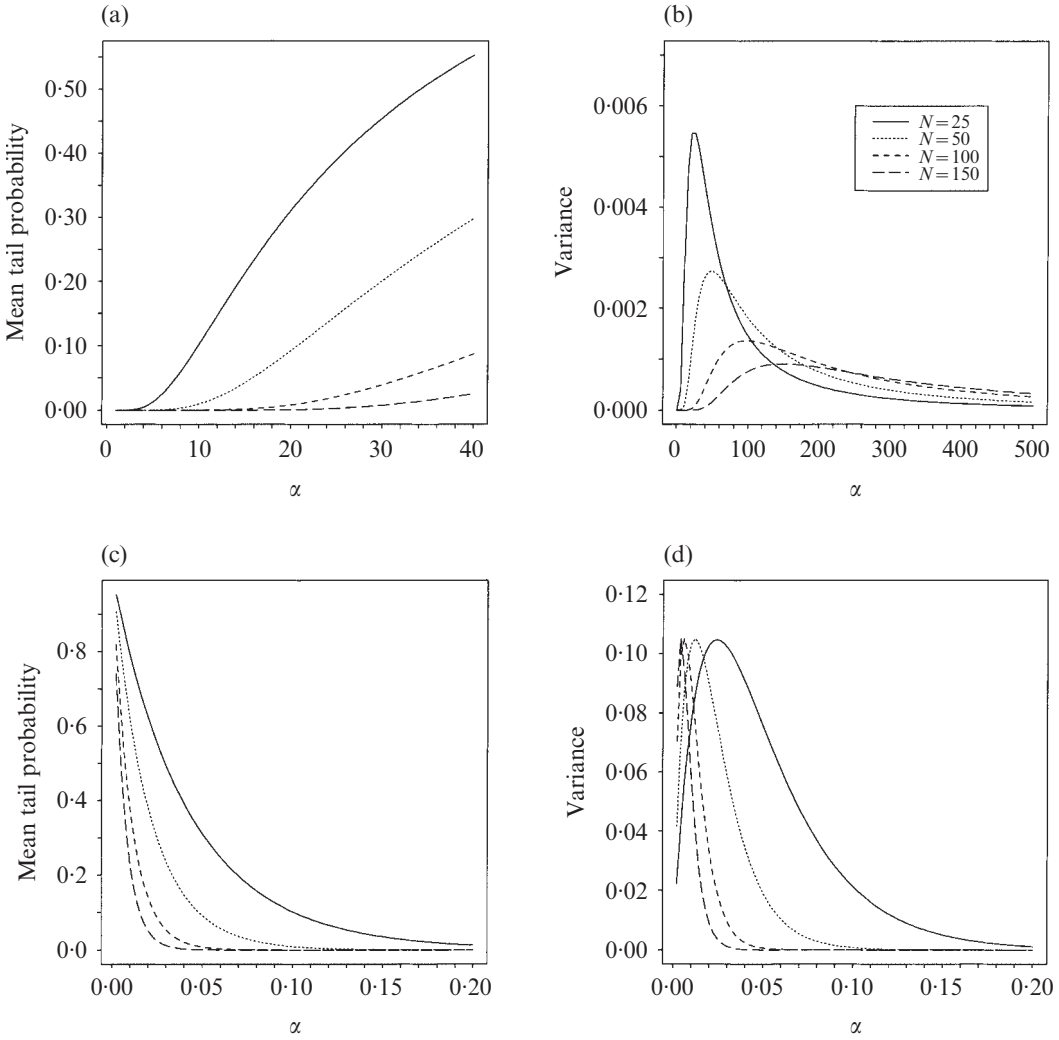

Fig. 5. Comparison between the tail probability $U_N(1) = \sum_{k=N}^{\infty} p_k$ for the Dirichlet DP(αH) process, shown in (a) and (b), and for the beta two-parameter $\mathscr{B}(\alpha, 1, H)$ process, shown in (c) and (d). (a) and (c) show the mean tail probability $E\{U_N(1)\}$ over various values of α and N. (b) and (d) show the variance, $\mathrm{var}\{U_N(1)\}$.

Figure 5 also illustrates a key difference between the two processes. Figures 5(b), (d) show that there is substantially higher variability in the tail probabilities of the beta two-parameter process. In fact, for small values of α, the variance of $U_N(1)$ can be at least 20 times higher than that for the Dirichlet process, at any α value, and this difference increases rapidly as N becomes large. This implies that the $\mathscr{B}(\alpha, 1, H)$ process tends to spread its p_k values less evenly than the Dirichlet process, and consequently it will tend to produce fewer distinct values than the Dirichlet process. This can be an advantage in finite mixture modelling, as we now describe.

4·5. *Normal mean mixture models for the* DP(αH) *and* $\mathscr{B}(\alpha, 1, H)$ *processes*

Using the truncation just discussed, we compare the $\mathscr{B}(\alpha, 1, H)$ process to the DP(αH) for the normal mean mixture data considered in § 3·2. Fitting these data proceeds by using the same Markov chain Monte Carlo algorithm as before, although we now can take advantage of exact sampling for α. A nice feature of the truncation approximation is that we no longer need to run a Metropolis step for α. In particular, for the DP(αH) truncation, it follows from (26) that the conditional density for α is

$$f(\alpha \mid p) \propto \alpha^{N-1} p_N^{\alpha-1} f(\alpha),$$

and with a $\mathrm{Ga}(v_1, v_2)$ prior for α it follows that

$$(\alpha \mid p) \sim \mathrm{Ga}(N + v_1 - 1, v_2 - \log p_N). \tag{30}$$

Exact sampling is also possible in the $\mathscr{B}(\alpha, 1, H)$ truncation, because

$$f(\alpha \mid p) \propto \alpha^{N-1} p_1^{\alpha-1} \ldots p_{N-1}^{\alpha-1} (1 - P_1)^{-\alpha} \ldots (1 - P_{N-2})^{-\alpha} f(\alpha)$$

and consequently

$$(\alpha \mid p) \sim \mathrm{Ga}\left\{N + v_1 - 1, v_2 - \sum_{k=1}^{N-1} \log p_k + \sum_{k=1}^{N-2} \log(1 - P_k)\right\}. \tag{31}$$

We compared the two processes using a truncation level of $N = 250$, and we also ran the mixture of Dirichlet processes approximation used in § 3 with this new larger value for N. The results are depicted in Fig. 6 and are based on the same priors and Gibbs sampling strategy that we used earlier. As Fig. 6 shows, the posterior number of distinct cluster values is roughly similar in all three models, although the beta two-parameter posterior tends to put less mass on two clusters. With such a large value for N, it is not surprising that the two Dirichlet process approximations yield similar results, but it is surprising that these results are close to those observed for the beta two-parameter approximation. One possible explanation is that in all three models we found the posterior for α to have a mean of approximately 1, which is the value at which their limits are identical. Another explanation is that the $\mathscr{B}(\alpha, 1, H)$ process, like the Dirichlet process, may be very good at uncovering the underlying components in a finite mixture model. As we noted earlier, the manner in which it distributes its p_k values ensures that it generates few distinct Y values, making it a practical tool in problems like this.

The truncation level of $N = 250$ that we have used appears to be more than adequate. We kept track of the mean and variance for the tail probability $U_N(1)$ based on the sampled value for α. In the Dirichlet truncation these quantities had average values smaller than 10^{-26}, and in the beta two-parameter truncation they were smaller than 10^{-6}. Unfortunately, there is no simple method for testing the adequacy of the mixture of Dirichlet processes approximation. However, given the similarity of the results to the truncation approximation, it would appear that N is sufficiently large. Interestingly, the posterior observed earlier for $N = n = 45$ is still quite similar to the results seen here. Even with such a small value for N, the approximation appears to be good.

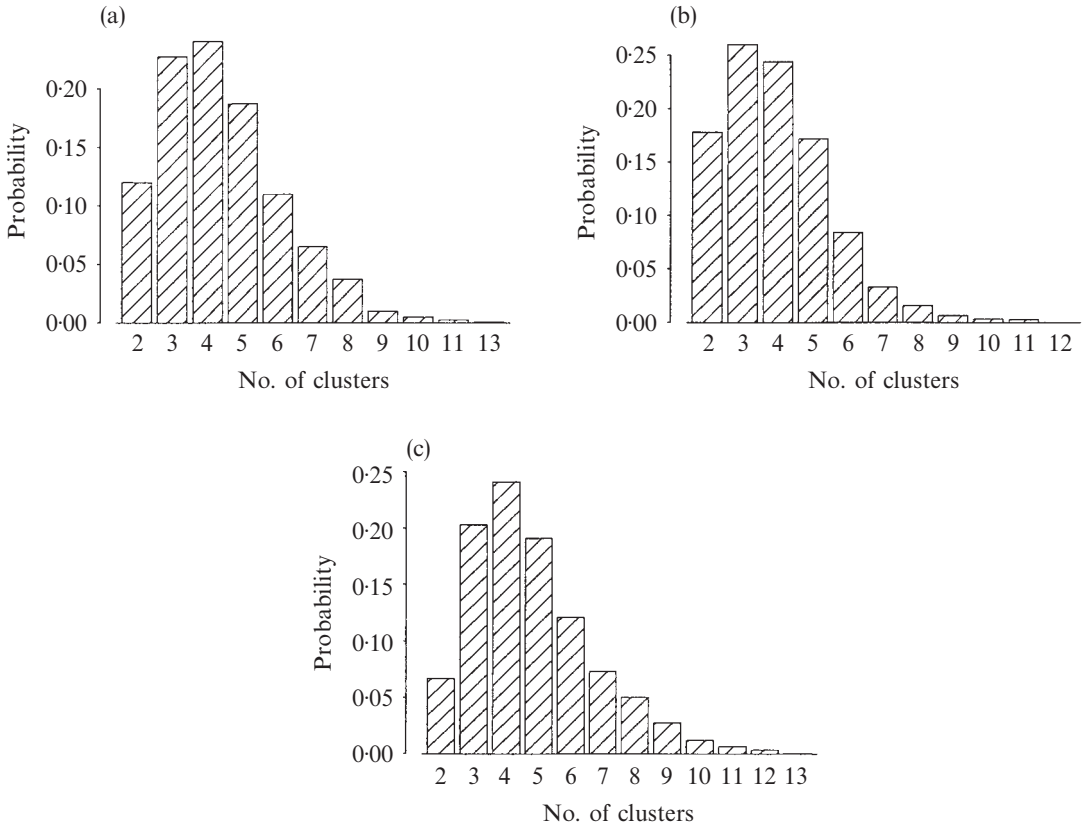

Fig. 6. Posterior distribution of the number of distinct Y_i values for the normal mean mixture data used in Fig. 1. Models are based on (a) the weak limit Dirichlet process approximation of § 3, (b) truncation of the Dirichlet process, and (c) truncation of the $\mathscr{B}(\alpha, 1, H)$ process. In each case, $N = 250$ with the same priors and sampling scheme employed as in the analysis for Fig. 1.

5. DISCUSSION

For the Dirichlet process, both the truncation approximation and the mixture of Dirichlet processes approximation involve Markov chain Monte Carlo algorithms that require roughly the same amount of computation. However, the truncation-based approximation offers the advantage of exact sampling for α, which becomes very relevant as N increases. With a large value of N, it can sometimes happen that a few p_k values become extremely small, and this can lead to serious numerical problems when running the Metropolis step; see equation (21).

Small p_k values are still an issue in evaluating the conditional distributions (30) and (31) for α in the $\mathrm{DP}(\alpha H)$ and $\mathscr{B}(\alpha, 1, H)$ processes. However, there is a simple way to increase the numerical stability if we remember that the current value for p is sampled from a generalised Dirichlet distribution. Since the value of p is constructed in terms of V_k beta random variables in (25), it means that we can re-express the conditional distribution for α in terms of V_k, as long as we remember to keep track of these terms each time we update p. For the $\mathrm{DP}(\alpha H)$ process, the conditional distribution (30) can be rewritten as

$$(\alpha \mid p) \sim \mathrm{Ga}\left\{N + v_1 - 1, v_2 - \sum_{k=1}^{N-1} \log(1 - V_k)\right\},$$

while, for the $\mathscr{B}(\alpha, 1, H)$ process, a nice cancellation occurs in the two sums in (31), giving

$$(\alpha \mid p) \sim \mathrm{Ga}\left(N + v_1 - 1, v_2 - \sum_{k=1}^{N-1} \log V_k\right).$$

Another attribute of the truncated Dirichlet process is that its accuracy increases exponentially in terms of N. To see why, recall from § 4 that the Dirichlet process tail probability has moments which decrease exponentially fast in N:

$$E\{U_N(r)\} = E\left(\sum_{k=N}^{\infty} p_k\right)^r = \left(\frac{\alpha}{\alpha + r}\right)^{N-1}.$$

Therefore, if $Y_1, \ldots, Y_n$ is a sample from the Dirichlet process $\mathscr{P}_\infty$, then

$$\mathscr{P}_\infty\{Y_1 = Z_{K_1}, \ldots, Y_n = Z_{K_n}, \text{ where } 1 \leqslant K_1, \ldots, K_n < N\} = E\{1 - U_N(1)\}^n,$$

which can be bounded by

$$E[\exp\{-nU_N(1)\}] = 1 + O\left\{n\left(\frac{\alpha}{\alpha + 1}\right)^{N-1}\right\}.$$

Note that the sample size n makes almost no impact on the exponential decrease in the mean of $U_N(1)$, and consequently the sampling behaviour under $\mathscr{P}_N$ and $\mathscr{P}_\infty$ will be almost identical. This is much better than the accuracy of the weak limit approximation of § 3, which depends on the accuracy of the empirical measure ξ_N. One method for gauging the weak limit accuracy is through Theorem 1. Using these bounds we find that the difference in the distinct number of values under $\mathscr{P}_N$ and the Dirichlet process is order $O(\log n/N)$. Thus, for the same value of N we anticipate exponentially higher accuracy using a truncation.

Another benefit in using a truncation-based approximation is that there is a simple diagnostic based on Theorem 2 for assessing its adequacy. As demonstrated, this diagnostic can easily be evaluated from the output of the Gibbs sampler and applies to the Dirichlet process, as well as to the beta two-parameter process. Unfortunately, there appears to be no simple method for assessing the adequacy of the weak limit approximation discussed in § 3.

Acknowledgement

The authors are very grateful to the reviewers for their invaluable advice and helpful comments.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–74.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Polya urn schemes. *Ann. Statist.* **1**, 353–5.

Connor, R. J. & Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Statist. Assoc.* **64**, 194–206.

Daniels, M. J. (1998). Computing posterior distributions for covariance matrices. In *Computing Science and Statistics*, **30**, Proceedings of the 30th Symposium on the Interface, Ed. S. Weisberg, pp. 192–6. Minneapolis, MN: Computing Science and Statistics.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Am. Statist. Assoc.* **89**, 268–77.

ESCOBAR, M. D. & WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–88.

ESCOBAR, M. D. & WEST, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 1–22. New York: Springer.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.

GUSTAFSON, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics* **53**, 230–42.

HJORT, N. L. (1996). Bayesian approaches to non- and semiparametric density estimation. In *Bayesian Statistics 5*, Ed. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, pp. 223–53. Oxford: Oxford University Press.

ISHWARAN, H. (1999). Applications of hybrid Monte Carlo to Bayesian generalized linear models: quasi-complete separation and neural networks. *J. Comp. Graph. Statist.* **8**, 779–99.

KORWAR, R. & HOLLANDER, M. (1973). Contributions to the theory of Dirichlet processes. *Ann. Prob.* **1**, 705–11.

MACEACHERN, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Commun. Statist.* B **23**, 727–41.

MACEACHERN, S. N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 23–43. New York: Springer.

MACEACHERN, S. N. & MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comp. Graph. Statist.* **7**, 223–38.

MULIERE, P. & SECCHI, P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Ann. Inst. Statist. Math.* **48**, 663–73.

MULIERE, P. & TARDELLA, L. (1998). Approximating distributions of random functionals of Ferguson–Dirichlet priors. *Can. J. Statist.* **26**, 283–97.

MULIERE, P. & WALKER, S. (1998). Extending the family of Bayesian bootstraps and exchangeable urn schemes. *J. R. Statist. Soc.* B **60**, 175–82.

MÜLLER, P., ERKANLI, A. & WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. New York: Springer-Verlag.

PITMAN, J. & YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable sub-ordinator. *Ann. Prob.* **25**, 855–900.

ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Am. Statist. Assoc.* **85**, 617–24.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639–50.

WALKER, S. & DAMIEN, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, Ed. D. Dey, P. Müller and D. Sinha, pp. 243–54. New York: Springer.

WALKER, S. & MULIERE, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25**, 1762–80.

WALKER, S. & WAKEFIELD, J. (1998). Population models with a nonparametric random coefficient distribution. *Sankhyā* B **60**, 196–214.

WEST, M., MÜLLER, P. & ESCOBAR, M. D. (1994). Hierarchical priors and mixture models, with applications in regression and density estimation. In *Aspects of Uncertainty: A Tribute to D. V. Lindley*, Ed. A. F. M. Smith and P. R. Freeman, pp. 363–86. London: John Wiley.

[*Received February* 1999. *Revised December* 1999]