Modelling Heterogeneity With and Without the Dirichlet Process

PETER J. GREEN
University of Bristol, UK

SYLVIA RICHARDSON
INSERM, France

ABSTRACT. We investigate the relationships between Dirichlet process (DP) based models and allocation models for a variable number of components, based on exchangeable distributions. It is shown that the DP partition distribution is a limiting case of a Dirichlet–multinomial allocation model. Comparisons of posterior performance of DP and allocation models are made in the Bayesian paradigm and illustrated in the context of univariate mixture models. It is shown in particular that the unbalancedness of the allocation distribution, present in the prior DP model, persists *a posteriori*. Exploiting the model connections, a new MCMC sampler for general DP based models is introduced, which uses split/merge moves in a reversible jump framework. Performance of this new sampler relative to that of some traditional samplers for DP processes is then explored.

Key words: allocation, Bayesian non-parametrics, entropy, finite mixture distributions, heterogeneity, Markov chain Monte Carlo, normal mixtures, partition, reversible jump algorithms, semi-parametric density estimation, sensitivity analysis, split/merge moves

1. Introduction

Models incorporating Dirichlet process (DP) priors have played an important role in recent developments in Bayesian applied statistics. The apparent flexibility of these models has found application in diverse areas: density estimation, non-parametric regression, autoregression, survival analysis, etc. In many applications, DP priors are not used directly for continuous data because of a discreteness problem: a distribution realized from a DP is almost surely discrete, so a random sample drawn from that realized distribution has positive probability of ties. Exploiting this discreteness, rather than combating it, DP mixtures (MDP) are used instead, providing a flexible model for clustering of items of various kinds in a hierarchical setting: random effects, parameters of sampling distributions, etc.; an early example of this is found in Lo (1984). Many modern applied non-parametric Bayesian methods use this clustering property.

The central role of the DP/MDP models in Bayesian non-parametrics, and the various lines of research stemming from it are recounted in the recent article by Walker *et al.* (1999). An inherent difficulty of the DP model is that a single parameter controls variability and coagulation, creating difficulties for prior specifications. This has motivated much of the recent work on generalizations of DP, including the construction of non-parametric priors based on more flexible control of the variability of the chosen partitioning of the space, as in Polya tree priors. In a hierarchical framework, a natural alternative to DP mixtures is to use mixtures based on multinomial allocations, thus increasing the flexibility of the allocation model; see Richardson (1999) and Richardson *et al.* (2000) for examples of this in the context of measurement error problems.

The purpose of this article is generally to relate the DP based models and associated clustering methods to more explicit multinomial allocation variable approaches. By relating the MDP model to a special case of a simple and familiar parametric model for mixtures, we throw

a different light on the claimed non-parametric nature of the MDP model. Distribution theory for this connection between model classes is explored in section 2. In section 3, we then investigate some of the statistical implications of DP models compared to the corresponding more flexible allocation variable approaches and illustrate these comparisons in a univariate mixture context. We show in particular that the lack of balance of the allocation distribution which exists in the prior DP model persists *a posteriori*.

Not least of the attractions of using the DP as a model component is the fact that Gibbs samplers for both prior and posterior are readily derived. We go on in section 4 to compare MCMC samplers for the two classes of models, and, motivated by this connection, introduce a new sampler for general DP based models using split and merge moves. In section 5, we compare the performance of new and old samplers for DP based univariate mixture models. Finally, in the appendix, we begin to explore wider classes of models for partition and allocation from a more axiomatic standpoint.

2. Distribution theory

Various non-parametric Bayesian hierarchical models have a structure which includes a n-vector ϕ of p-dimensional variables $(\phi_1, \ldots, \phi_n)$, with an exchangeable prior distribution giving positive probability to ties and specified, sometimes indirectly, in terms of a parameter α, and a continuous distribution G_0 on R^p. Usually, but not necessarily, the variables $(\phi_1, \ldots, \phi_n)$ are not directly observed but parametrize the distributions for observables $(y_1, y_2, \ldots, y_n)$, respectively. We give concrete motivating examples for this set-up in section 2.3.

In these settings, a realization of such a ϕ provides simultaneously a partition of the n items into groups, and a parameter value ϕ_i equal for all items in a group. Alternatively, we can view ϕ as providing a set of distinct parameter values, together with an allocation of the n items to those values. These viewpoints are not quite equivalent, since the second implies a labelling of the groups.

2.1. Dirichlet process priors

One formulation for such a random vector ϕ is that using a Dirichlet process prior (see Ferguson (1973) for the definition and properties of the DP process).

The DP model for ϕ is defined in two stages:

(a) a random distribution G is drawn from the Dirichlet process: $G \sim \mathrm{DP}(\alpha, G_0)$, where α is a positive real number and G_0 is a distribution on a space Ω, then given G,
(b) $\phi = (\phi_1, \ldots, \phi_n)$ consists of n i.i.d. draws from G.

Since in the DP, G_0 is the prior expectation of G, the ϕ_i are marginally drawn from G_0.

Let us examine the distributions of partition and allocation induced by the DP model. The pattern of ties among the entries of ϕ determines a partition of $I = \{1, 2, \ldots, n\}$, an unordered set of d disjoint non-empty subsets of I, whose union is I; the number d of subsets is the degree of the partition; here we will call the subsets groups, denoting them generically by g. If we label the groups $1, 2, \ldots, d$, we impose an ordering on them: $g_1 < g_2 < \cdots < g_d$. Then we can write $z_i = j$ if $i \in g_j$, and define θ by $\phi_i = \theta_{z_i}$, $i = 1, 2, \ldots, n$. We could use various possible rules to order the groups, for example, (i) ordering the g_j according to $\min\{i : i \in g_j\}$, or, (ii) given an order on Ω, ordering according to the values $\{\theta_j\}$. Under the DP model and using (ii), all allocations giving the same partition are equally likely.

We find (e.g. Antoniak, 1974)

$$p(g) = p(g_1, g_2, \ldots, g_d) = \frac{\alpha^d \Gamma(\alpha) \prod_{j=1}^{d} (n_j - 1)!}{\Gamma(\alpha + n)} = \frac{\alpha^d \prod_{j=1}^{d} (n_j - 1)!}{\alpha(\alpha + 1) \cdots (\alpha + n - 1)} \tag{1}$$

where $n_j = \# g_j$, $j = 1, 2, \ldots, d$. It is sometimes useful to express this conditionally on d; we have

$$p(g|d) = \frac{\prod_{j=1}^{d} (n_j - 1)!}{|S_n^{(d)}|} \tag{2}$$

and

$$p(d) = \frac{\alpha^d |S_n^{(d)}|}{\alpha(\alpha + 1) \cdots (\alpha + n - 1)}, \tag{3}$$

where

$$|S_n^{(d)}| = \sum_g \prod_{j=1}^{d} (n_j - 1)!$$

is the absolute value of a Stirling number of the first kind (see Abramowitz & Stegun, 1972, p. 824). These well-known relationships will be useful for establishing our limiting results in section 2.4.

2.2. Explicit allocation priors

A more explicit formulation that arises naturally, particularly in mixture models:

(a) draws the number of groups k from an arbitrary distribution $p(k|\xi)$; then, given k, it
(b) draws an n-vector of allocation variables z, with $z_i \in \{1, 2, \ldots, k\}$, from some distribution exchangeable over items;
(c) draws $\theta = (\theta_1, \ldots, \theta_k)$ as k i.i.d. variables from G_0; and finally
(d) sets $\phi_i = \theta_{z_i}$.

Our canonical example of step (b) in this second formulation is to first draw w from an appropriate distribution on the k-dimensional simplex and then given k and w, draw $\{z_i\}$ i.i.d. with $p(z_i = j) = w_j$. We usually take w to have the symmetric Dirichlet distribution $D(\delta, \ldots, \delta)$, so that the allocation variables are also exchangeable over groups; we then refer to this set-up as the Dirichlet/multinomial allocation (DMA) model. A default choice is to take $\delta = 1$, making the weight distribution uniform on the simplex.

Note that the multinomial distribution allows the possibility of empty components (a component j is empty if $z_i \neq j \forall i$). Another possible allocation model, not further explored here, would have been to draw z given w conditional on there being no empty components. This model has been discussed by Wasserman (2000), in the context of non-informative priors for mixture models with fixed k. It would be straightforward to implement in a fixed-k context using rejection sampling, but a little more cumbersome with variable k as some normalizing constants, depending on n and k, would need to be evaluated.

To find the allocation distribution induced by the DMA model means marginalizing over the weights w. We have $p(z|w, k)$ specified by

$$p(z_i = j) = w_j \quad \text{independently for } j = 1, 2, \ldots, k,$$

and

$$p(w|k) = \frac{\Gamma(k\delta)}{\{\Gamma(\delta)\}^k} \prod_{j=1}^{k} w_j^{\delta-1}$$

on the simplex $\{w\colon w_j \geqslant 0, \sum_{j=1}^{k} w_j = 1\}$, where this latter expression can be interpreted as the density of any $(k-1)$ of $\{w_1, w_2, \ldots, w_k\}$ with respect to Lebesgue measure.

Integrating out w, we find

$$p(z|k, \delta) = \frac{\Gamma(k\delta)}{\{\Gamma(\delta)\}^k} \frac{\prod_{j=1}^{k} \Gamma(\delta + n_j)}{\Gamma(k\delta + n)} = \frac{\Gamma(k\delta)}{\Gamma(k\delta + n)\{\Gamma(\delta)\}^d} \prod_{j\colon n_j > 0} \Gamma(\delta + n_j)$$

where $n_j = \#\{i : z_i = j\}$.

For comparison with the DP model, it is helpful to express this as a distribution over partitions. Since the groups are labelled 1, 2, ..., k, there are $k_{(d)} = k!/(k-d)!$ allocations z giving the same partition g of the items 1, 2, ..., n, where d is the degree of the partition, $d = \#\{j : n_j > 0\}$. These allocations are equally probable under the DMA model, so we have

$$p(g|k, \delta) = \frac{k!}{(k-d)!} \frac{\Gamma(k\delta)}{\Gamma(k\delta + n)\{\Gamma(\delta)\}^d} \prod_{j\colon n_j > 0} \Gamma(\delta + n_j) \tag{4}$$

2.3. Using the DP and DMA specifications in hierarchical models

In Bayesian modelling of structured data, the specification of a random n-vector ϕ in terms of ζ, δ and G_0 in the DMA model, or α and G_0 in the DP model, will form only a part of a full hierarchical model. Other nodes will be added to the directed acyclic graph representing the model, both ancestors of α, δ and G_0 and descendants of ϕ.

As introduced in section 2, a typical data generating mechanism is that observables $(y_1, y_2, \ldots, y_n)$ are available, conditionally independent given ϕ and other parameters in the model, with distributions of known form parameterized respectively by $(\phi_1, \phi_2, \ldots, \phi_n)$. For instance, all six of the applications listed in MacEachern & Müller (1994) include this feature.

In the DP case, this set-up frequently enjoys the misleading appellation of a "mixture of Dirichlet processes" (MDP) model, the terminology of DP mixture models used by West *et al.* (1994) being clearer. See O'Hagan (1994, pp. 288ff.) for further discussion.

At the top of the hierarchy, the parameters ζ, δ and G_0 could in principle be fixed or random, and if random possibly modelled hierarchically, depending on the context. Let us consider one example, that of Bayesian density estimation using a flexible class of multivariate normal mixtures, which has recently been discussed by Müller *et al.* (1996). Here the $(y_1, y_2, \ldots, y_n)$ are observed random quantities independently drawn from an uncertain distribution, to be estimated. A hierarchical model is defined in which $y_i \sim \mathrm{N}(\mu_i, \Omega_i)$, where the pairs of parameters $\phi_i = (\mu_i, \Omega_i)$ are chosen to be dependent, but are marginally identically distributed according to a product of normal $N(a, B)$ and inverse Wishart densities $W(s, S)$. Thus $G_0(\mu, \Omega|\eta) = N(\mu; a, B)W(\Omega^{-1}; s, S)$ and above G_0 the hyperparameters $\eta = (a, B, s, S)$ are also given prior densities.

In application to univariate normal mixtures, as implemented by Richardson & Green (1997), ζ is fixed, and G_0 set to be normal $(\xi, \kappa) \times$ inverse-gamma (γ, β) where only β is random, with a gamma hyperprior. We return to this set-up in more detail later in the paper; it provides a running example, used to illustrate the calculations needed to implement MCMC methods for these models, and the basis for our experimental comparisons.

2.4. Connections between the DP and DMA models

The DP partition distribution arises from the corresponding distribution for the DMA model under two different limiting regimes, as can be seen by comparing (1) and (4); in both cases, n is fixed. For the first, suppose that in (4), $\delta \to 0$ and $k \to \infty$ in such a way that $k\delta \to \alpha > 0$. Then $k!/(k-d)! \sim (\alpha/\delta)^d$ and $\Gamma(\delta) \sim \delta^{-1}$, so the right hand side of (4) converges to that of (1). Formally, the consequence is that

$$p_{\mathrm{DMA}}(\phi, y|k, \delta, G_0) \to p_{\mathrm{DP}}(\phi, y|\alpha, G_0)$$

as $k \to \infty$ with $k\delta \to \alpha > 0$.

Thus, so far as the occupancy of non-empty components is concerned, the DP model arises as a limit in which the number of components in the DMA model goes to ∞ while the total of the Dirichlet parameters for $p(w)$ remains fixed at α. This limiting case of the DMA model was studied by Neal (1992), see also Neal (1998); in some sense this seems to have been "generally known", but we have been unable to find a prior statement of this precise connection to the DP process.

Alternatively, consider the DMA partition distribution (4) under the condition that there are no empty components; we stress that this perspective is purely for drawing comparisons between the models, and is not adopted in our proposed methodology. By analogy with the DP case, we use the term degree and the symbol d for the number of non-empty components in the DMA case. The event $\{n_j > 0 \forall j\}$ that there are no empty components can then be written $\{d = k\}$. We find

$$p_{\mathrm{DMA}}(g|k, \delta, d = k) = \frac{\prod_{j=1}^{d} \Gamma(\delta + n_j)}{\sum_{g} \prod_{j=1}^{d} \Gamma(\delta + n_j)}.$$

On letting $\delta \to 0$, this converges to the right hand side of (2).

Thus the DP also corresponds to taking the explicit allocation Dirichlet/multinomial distribution for $p(w, z|k)$, and both conditioning on $n_j > 0 \forall j$ (that is, there are no empty components) and letting $\delta \to 0$ (that is, favouring more unequal allocations). In this limiting regime we must also set the $p(d)$ distribution to be that given in (3).

In both models, the distinct ϕ_i, that is $\{\theta_j, j = 1, 2, \ldots, k\}$ are drawn i.i.d. (given α or ζ, and G_0) from $G_0 = G_0(\cdot|\eta)$.

It is instructive to see numerical values for the partition and allocation distributions for the two models, for small n. See Tables 1 and 2. For example, compare the probabilities assigned

Table 1. *Partition and allocation distribution for DP model, $n = 4$. All probabilities should be divided by $(\alpha+1)(\alpha+2)(\alpha+3)$. The notation $\langle m \rangle$ means that to save space, other cases of similar pattern and equal probability have been omitted; there are m such cases in all.*

Degree		Partition			Allocation		
d	$p(d)$	g		$p(g)$	z		$p(z)$
1	6	(1234)		6	1111		6
2	11α	(123)(4)	$\langle 4 \rangle$	2α	1112	$\langle 2 \rangle$	α
		(12)(34)	$\langle 3 \rangle$	α	1122	$\langle 2 \rangle$	$\alpha/2$
3	$6\alpha^2$	(12)(3)(4)	$\langle 6 \rangle$	α^2	1123	$\langle 6 \rangle$	$\alpha^2/6$
4	α^3	(1)(2)(3)(4)		α^3	1234	$\langle 24 \rangle$	$\alpha^3/24$

Table 2. *Partition and allocation distribution for DMA model, $n = 4$. All probabilities should be divided by $(k\delta(k\delta+1)(k\delta+2)(k\delta+3)$. The notation $\langle m \rangle$ means that to save space, other cases of similar pattern and equal probability have been omitted; there are m such cases in all. Abbreviations: $a = \delta(\delta+1)(\delta+2)(\delta+3)$, $b = \delta^2(\delta+1)(\delta+2)$, $c = \delta^2(\delta+1)^2$, $d = \delta^3(\delta+1)$, $e = \delta^4$, $k_{(r)} = k!/(k-r)!$.*

Degree		Partition			Allocation		
d	$p(d)$	g		$p(g)$	z		$p(z)$
1	ka	(1234)		ka	1111	$\langle k \rangle$	a
2	$k_{(2)}(4b+3c)$	(123)(4)	$\langle 4 \rangle$	$k_{(2)}b$	1112	$\langle k_{(2)} \rangle$	b
		(12)(34)	$\langle 3 \rangle$	$k_{(2)}c$	1122	$\langle k_{(2)} \rangle$	c
3	$6k_{(3)}d$	(12)(3)(4)	$\langle 6 \rangle$	$k_{(3)}d$	1123	$\langle k_{(3)} \rangle$	d
4	$k_{(4)}e$	(1)(2)(3)(4)		$k_{(4)}e$	1234	$\langle k_{(4)} \rangle$	e

under either model to the partitions (123)(4) and (12)(34). Under the DP model, each partition of the pattern (123)(4) is twice as likely as any of the pattern (12)(34), while under the DMA model the ratio of probabilities is $b/c = (\delta+2)/(\delta+1)$, or 1.5 in the uniform case $\delta = 1$. Thus, relatively, the DP model favours more unequal allocations. This is a general phenomenon, and indeed is much more dramatic numerically as n increases. For example, for 100 items partitioned into four groups, both models give astronomically more probability to each partition with $n_1 = 97$ and $n_2 = n_3 = n_4 = 1$ than to one with $n_1 = n_2 = n_3 = n_4 = 25$, but the ratio is about 4000 times greater for the DP than for the DMA with $\delta = 1$.

3. Statistical comparisons between DP and DMA models

In this section, we explore comparative posterior performance of the DP and DMA models. Our discussion is centred on marginal comparisons, whether of a global measure of fit or of the implied allocation distribution given by the two models. We focus entirely on properties following from the joint posterior of the $\{\phi_i\}$, rather than, say, from that of G, their unknown distribution in the DP model.

On the face of it, such a basis for comparison may seem unfair to the DP model, which through including G explicitly in the model, allows inference about G, and hence also uses a different basis for prediction many steps ahead. Such features clearly differentiate the DP from the DMA model in principle. However, it is precisely such aspects of DP-based inference which we find most untenable from an empirical perspective, since as already explained, G is discrete with probability 1 *a priori* and hence *a posteriori*. Further, these additional opportunities for posterior inference in the DP setting seem very seldom to be used in practice.

The ability of each model to fit a dataset is crucially dependent on the number of components allowed by the prior structure. Comparisons between the models must therefore take account of this. It would be tempting to try to calibrate $p(k|\zeta)$ and $p(d|n, \alpha)$ *a priori* to have the same impact on the respective models, but this seems not to be possible. In any case, a more general and robust basis for comparison is to condition on the actual number of components. This is aided by the fundamental conditional independence properties in each model:

(i) in the DP model, conditional on d, the hyperparameter α is independent of all other parameters and the data;
(ii) in the DMA model, conditional on k, the hyperparameter ζ is independent of all other parameters and the data. However, this assertion is not exactly true with k replaced by d.

Thus, by respectively conditioning on d and k, and exploiting these conditional independences, we have a nearly perfect basis for the elimination of the effect of hyperparameters.

The only difficulty is that in DMA, conditioning on d rather than k maintains a better parallel with DP. Thus, in each of the comparisons we make below, we have tried to exercise our best judgement about whether to use k or d in the DMA model, specific to that comparison.

3.1. Density estimates and deviances

For observables $(y_1, y_2, \ldots, y_n)$, we summarize the quality of fit of a point estimate h of their density by defining the associated deviance

$$D(h) = -2\sum_{i=1}^{n} \log(h(y_i)). \tag{5}$$

(Note that we depart from the usual sense of the term "deviance" in not subtracting from this twice-negative-log-likelihood some baseline value corresponding to a saturated model, because in this non-parametric setting, such a baseline would be $-\infty$.)

We base our comparisons between the DP and DMA models on the density estimates produced by each, and statistics derived from these, for example global goodness-of-fit measures such as D. In the Bayesian setting, the density estimate is simply the predictive distribution for the next observation and so that is what we use here; note that we are not interested in prediction *per se*, and so prediction more than one step ahead is not a relevant issue.

Note that several kinds of predictive densities for a new observation, corresponding to different conditionings, can be constructed.

Let us first consider as density estimate the unconditional predictive density for a new observation y^* given the data, $p(y^*|y, k)$. Using the event that y and y^* are conditionally independent given θ, z and z^*, it can be shown that

$$p(y^*|y, k) = E\left[\sum_j w_j f(y^*|\theta_j)\middle| y, k\right]. \tag{6}$$

It will be convenient to write this function as $\hat{g}_k(y^*)$, a quantity which can be computed on a grid of y^* values by averaging across the MCMC run, conditional on fixed values of k. For the DP model, we condition on the degree d and on the fact that the new observation does not create a group by itself to define the corresponding expression to (6):

$$\hat{g}_d(y^*) = E\left[\sum_j \frac{n_j}{n} f(y^*|\theta_j)\middle| y, d\right]. \tag{7}$$

To get a global measure of quality of fit of the density estimates $\hat{g}_k$ and $\hat{g}_d$ given by the DMA and DP models, we thus compute respectively $D(\hat{g}_k)$ and $D(\hat{g}_d)$ as defined in (5).

It is also of interest to understand the variability around these point estimates of density. Thus instead of taking expectations, we condition at the highest level at which the models are compatible, and consider the quantities $g(y^*) = p(y^*|y, z, \theta, k)$ and the associated deviance $D(g)$ as defined in (5). Note that we have integrated out $\{w_j\}$ to facilitate comparability between DMA and DP. Similarly to (6), it can be shown that for the DMA model

$$g(y^*) = \sum_j \frac{n_j + \delta}{n + k\delta} f(y^*|\theta_j). \tag{8}$$

For the DP model, we use $p(y^*|y, \theta, d)$ and the expression (8) with $\delta = 0$. It will be interesting to compare the distribution of $D(g)$ given k or d, in particular its mean and variability. Note

that, by Jensen, we always have $E(D(g)) \geqslant D(\hat{g}_k)$ and that the difference $E(D(g)) - D(\hat{g}_k)$ will be larger if g is more variable. This measure of variability is equivalent to p_D, introduced as a measure of complexity by Spiegelhalter *et al.* (1998), with a particular choice of parameterization. We intend to explore the role of p_D in mixture model determination in later work.

3.2. Model and data specification

Our comparisons will be made in the context of a distribution G_0 corresponding to univariate normal mixtures. We have used three data sets, enzyme, acidity and galaxy, described in Richardson & Green (1997), as well as four simulated data sets of 100 points. These latter correspond to a unimodal leptokurtic mixture ('lepto'): $0.67N(0, 1) + 0.33N(0.3, (0.25)^2)$, a bimodal mixture ('bimod'): $0.5N(-1, (0.5)^2) + 0.5N(1, (0.5)^2)$, an asymmetric separated mixture ('sep'): $0.1N(5, 1) + 0.4N(12, 1) + 0.3N(16, (0.5)^2) + 0.2N(20, (1.5)^2)$, and a symmetric strongly overlapping platykurtic mixture ('platy'): $0.2N(-4, 1) + 0.2N(-2, 1) + 0.2N(0, 1) + 0.2N(2, 1) + 0.2N(4, 1)$. The four synthetic mixture densities are graphed, on a standardized scale, in Fig. 1. Throughout we let R denote the interval of variation of the data and we adopt the following specification for the normal mixture model: $\theta = (\mu, \sigma^{-2})$ and $G_0 = N(\xi, \kappa^{-1}) \times \Gamma(\gamma, \beta)$ with fixed values of $\xi =$ midrange, $\kappa = 1/R^2$, $\gamma = 2$, and a random β which follows a $\Gamma(g, h)$ distribution. This hierarchical mixture model and the choice of g and h ($g = 0.2$, $h = 10/R^2$) are discussed in Richardson & Green (1997). For the DP model, we set $\alpha = 1$. For the DMA model, we let $\delta = 1$ throughout which corresponds to a uniform prior on the weights, a natural choice in the absence of real prior information. In the analysis, a prior uniform on $\{1, 2, \ldots, 30\}$ was assumed for k, although as usual this could be amended to any other prior on this support by importance sampling in the output analysis. For the DP model, the results presented correspond to runs of 100,000 sweeps (after a burn-in of 100,000 sweeps) of the reversible jump algorithm described in section 4.1; similarly for the DMA model, 100,000

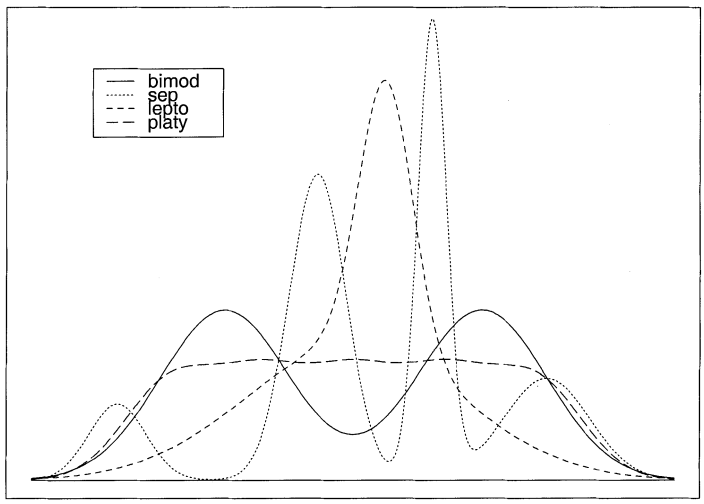

Fig. 1. Plots of the bimod, sep, lepto and platy normal mixture density functions translated and scaled to have similar ranges.

sweeps of the reversible jump algorithm presented in Richardson & Green (1997) were used after a burn-in of 100,000 sweeps.

3.3. Posterior distribution of the number of components

Recall that for the DMA model there is a free choice of the prior distribution $p(k)$ of the number of components k, which include empty components; in contrast the prior $p(d)$ on the partition degree d of the DP model is completely determined by n and α. In order partially to "factor out" the influence of the priors, we thus chose to compare modified posteriors $p^*(k|y) \propto p(k|y)/p(k)$ and $p^*(d|y) \propto p(d|y)/p(d)$ corresponding to uniform priors for k and d in the two models.

Figure 2 plots the cumulative distribution of $p^*(k|y)$ vs $p^*(d|y)$ for the seven data sets. The average number of empty components was small for most data sets, ranging from 0.07 to 0.15, except for the "sep" data set (0.45) and the galaxy data (0.65). For the "lepto" data set, the cumulative distributions are identical (diagonal line). For all the other data sets, except galaxy, the plots show small convexity, indicating that the mixture models estimated with DMA priors have fewer components—*a fortiori*, fewer non-empty components—than those corresponding to the DP priors. It is interesting to note that the single data set, galaxy, where this does not hold has small clusters of outlying observations, which is well in keeping with the DP allocation model.

3.4. Entropy and partitions

Our next concern is to investigate whether the DP model's prior emphasis on unequal allocation persists in the posterior. A similar concern was expressed in Petrone & Raftery (1997) with particular reference to change point models. We can summarize equality of allocation by the

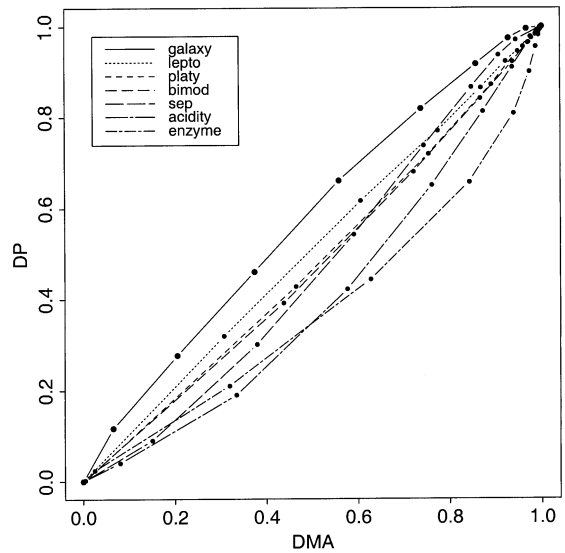

Fig. 2. Modified cumulative posterior distributions for the number of components, compared for the DMA and DP models using PP plots.

entropy, defined by $-\sum_j (n_j/n)\log(n_j/n)$, and look at the conditional posterior of entropy given degree d.

We found that the mixtures with a DP prior have systematically lower entropy, the difference being noticeable for any value of d above 3 (see Fig. 3). This difference is accentuated for larger samples drawn from the same simulated models (results not shown). The persistence of unequal allocations can also be seen when one compares mean group sizes for the two models. Figure 4 presents a typical comparison; the lack of balance is more noticeable as the degree increases. Hence, as was also noted by Petrone & Raftery (1997), in most cases the unbalancedness of the prior allocation model is still noticeable in the posterior.

It is also of interest to investigate posterior classification for the data conditional on d or k. Of course, this requires choosing an unambiguous labelling. When the components are labelled according to the order of their means, we have found that the DP model treats outlying observations differently. For example, in the classification of the acidity data into four groups, the left hand outlying observation constitutes a single component under the DP, whereas for the DMA model it is regrouped with observations belonging to a component with a large variance (results not shown).

3.5. Deviances

We computed $D(\hat{g}_k)$ and $D(\hat{g}_d)$ as defined in (6), (7) and (5) for the 7 data sets and values of k or d well-supported *a posteriori*. We found nearly identical values for simple well-separated mixtures ("bimod" and "sep"), and slightly lower values in general for $D(\hat{g}_k)$ but with few differences exceeding 1 (see Fig. 5). The only notable difference in fit concerns the enzyme data (see Table 3), a data set for which we have noticed that the induced partitions and classification (not shown) differ markedly between the DMA and the DP models.

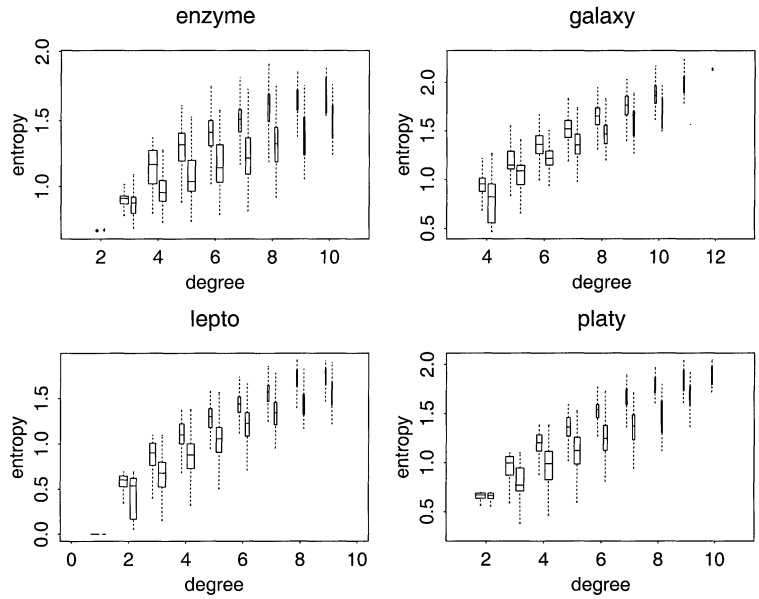

Fig. 3. Conditional distribution of entropy given degree, for DMA and DP mixture models applied to four data sets. In each pair of boxplots, DMA is on the left.

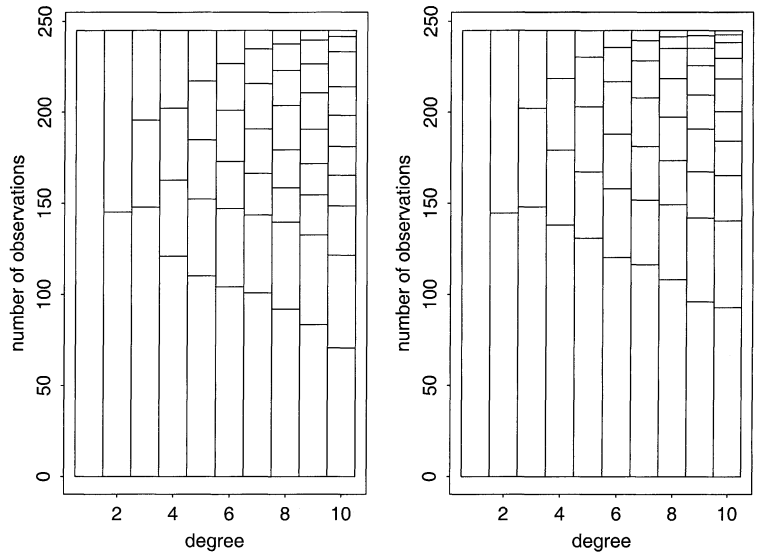

Fig. 4. Mean group sizes for DMA (left panel) and DP mixture models, enzyme data.

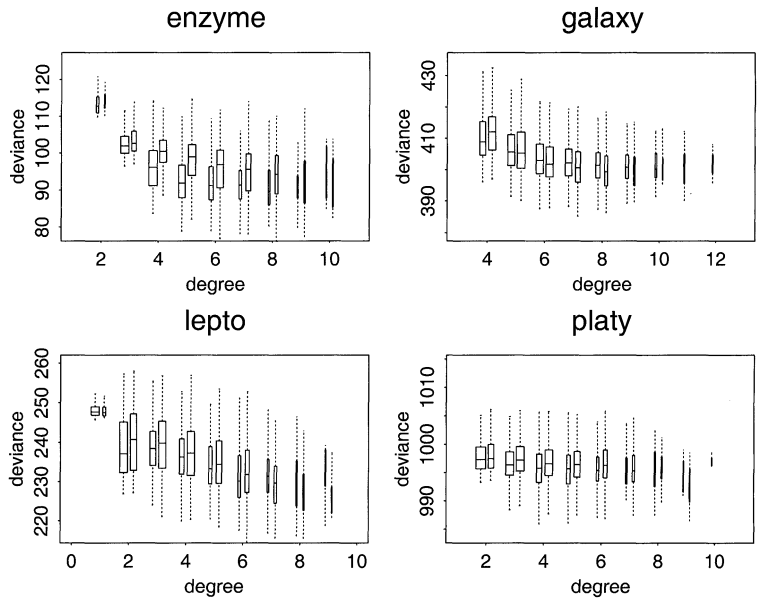

Fig. 5. Distributions of deviance $D(g)$: comparison between DMA and DP models. In each pair of boxplots, DMA is on the left.

4. MCMC methods for DP and related models

The interest in DP and MDP models for practical Bayesian analysis has generated much research into the efficient implementation of MCMC methods for computing the resulting posteriors. Significant contributions to this effort are MacEachern (1994), Escobar & West

Table 3. *Enzyme data: deviances associated with density estimates derived from DMA or DP models.*

k or d	2	3	4	5	6
$D(\hat{g}_k)$ for DMA	107.0	93.2	84.0	80.5	79.4
$D(\hat{g}_d)$ for DP	106.9	93.6	88.7	86.0	83.5

(1995), Bush & MacEachern (1996), and MacEachern & Müller (1994, 1998); these make use of the constructive incremental nature of the DP process (see appendix) leading to natural Gibbs samplers for allocation variables or parameters. Convergence rates for these Gibbs samplers have recently been investigated by Roberts *et al.* (2000). In contrast, Richardson & Green (1997) developed reversible jump Metropolis–Hastings samplers for their DMA representation of the finite mixture models.

In view of the intimate correspondence between DP and DMA models discussed above, it is interesting to examine the possibilities of using either class of MCMC methods for the other model class. We have been unsuccessful in our search for incremental Gibbs samplers for the DMA models, but it turns out to be reasonably straightforward to implement reversible jump split/merge methods for DP models.

For the necessary dimension-jumping, Richardson & Green used empty-component birth/death moves in addition to the splits and merges (and Phillips & Smith (1996) implemented general birth and death of components in their mixture methods), but we do not pursue that line here. Instead, we focus on the split/merge mechanism; this seems to be an idea with general applicability, which has been much used in implementations of reversible jump methods, although not without some problems in multivariate mixtures. In discussing this, in the next subsection, we do not need to be specific to the DP setting, but work with general allocation models.

Later, in subsection 4.2, we draw some comparisons between this new sampler and two existing methods, one of which is also suitable for non-conjugate MDP models.

4.1. Split/merge samplers for allocation models

Consider a general DP model, with a p-dimensional parameter $\theta \in \mathscr{R}^p$. A MCMC sampler set in the reversible jump framework (Green, 1995) will comprise a collection of reversible moves, some of which will be routine fixed-dimension transition kernels, but including at least one move that changes d, the degree of the partition. We follow usual practice in attempting only rather modest changes to the parameter space. A split/merge move is one that increases d by taking one group, say g_j, and its corresponding parameter θ_j and splits it into two non-empty groups g_{j-} and g_{j+} with corresponding θ_{j-} and θ_{j+}; the reverse merge move merges the groups, and produces a single parameter θ_j. As always, we use intuition to specify the details of these mechanisms and ensure that detailed balance is obtained with respect to the required target (posterior) distribution by correctly calculating the Metropolis acceptance ratio, which deals with the split and merge as a pair.

In terms of counting parameters, note that we are jumping between $(K + p)$ and $(K + 2p)$-dimensional parameter spaces, where K denotes the number of other parameters of the model, not altered by this move. How can this be accomplished?

We need to generate $\theta_{j-,l}$ and $\theta_{j+,l}$, $l = 1, 2, \ldots, p$. Intuitively, proposed values will be well-supported in the posterior if they provide similar explanatory power as $\{\theta_{j,l}\}$. We follow the pattern of the applications in Green (1995) and Richardson & Green (1997) by aiming to conserve p conditions of the form

$$m_l(\theta_j) = w_- m_l(\theta_{j-}) + w_+ m_l(\theta_{j+}), \tag{9}$$

for suitably chosen "mock weights" w_-, w_+ summing to 1, choice of which is to be discussed shortly. We assume the vector function $m : \mathscr{R}^p \to \mathscr{R}^p$ is invertible. (For example in mixture density estimation, $m_l(\theta)$ might be the lth moment of the density specified by θ.) Then in merging, (9) defines θ_j. In splitting, we have considerable freedom, but it may be useful here to sketch out some generic methods. Which is most suitable will depend on the detail of the model, and the form of the matching functions $\{m_l(\cdot)\}$.

The general pattern is to draw a p-vector of auxiliary random numbers $u = (u_1, u_2, \ldots, u_p)$, and set up a bijection between (θ_j, u) and $(\theta_{j-}, \theta_{j+})$ using the $\{m_l(\cdot)\}$.

For example, if the $\{m_l(\cdot)\}$ vary freely over $\mathscr{R}$, we might use (9) together with

$$u_l = w_+ m_l(\theta_{j+}) - w_- m_l(\theta_{j-}).$$

This provides an invertible transformation between (θ_j, u) and $(\theta_{j-}, \theta_{j+})$ whose Jacobian can be simplified to the form

$$\left| \frac{\partial(\theta_{j-}, \theta_{j+})}{\partial(\theta_j, u)} \right| = \frac{|\nabla m(\theta_j)|}{|\nabla m(\theta_{j-})||\nabla m(\theta_{j+})|(2w_- w_+)^p},$$

where ∇ denotes the gradient operator.

Alternatively, if the $\{m_l(\cdot)\}$ are positive but free of any other constraints, then we might draw $u_l \sim U(0, 1)$ independently (or indeed use any other continuous distribution on $[0, 1]^p$) and use

$$m_l(\theta_{j-}) = \frac{u_l m_l(\theta_j)}{w_-} \quad \text{and} \quad m_l(\theta_{j+}) = \frac{(1 - u_l) m_l(\theta_j)}{w_+},$$

and this time the Jacobian reduces to

$$\left| \frac{\partial(\theta_{j-}, \theta_{j+})}{\partial(\theta_j, u)} \right| = \frac{|\nabla m(\theta_j)| \Pi_l |m_l(\theta_j)|}{|\nabla m(\theta_{j-})||\nabla m(\theta_{j+})|(w_- w_+)^p}$$

In fact for the normal mixture application Richardson & Green (1997) use neither of these, as their matching functions are the mean and mean square of the corresponding components, and of course, the mean square must exceed the square of the mean.

Now, we must discuss allocating items into the groups g_{j-} and g_{j+}. Having chosen to split g_j, and given the new parameter values θ_{j-} and θ_{j+}, we suppose we distribute $i \in g_j$ between g_{j-} and g_{j+} according to the natural conditional probabilities

$$P(i \to g_{j-}) = \frac{w_- p(y|z_i = j-)}{w_- p(y|z_i = j-) + w_+ p(y|z_i = j+)}.$$

It remains only to define the mock weights w_-, w_+. Their purpose is to allow uneven splitting, and adjust for unequal n_j in merging. On splitting, we propose to generate $w_- \sim U(0, 1)$; on merging, $w_- \sim Be(n_{j-} + \omega, n_{j+} + \omega)$ for a simulation parameter ω, in our experiments taken to have the value 5.

There is no additional contribution to the Jacobian from either these weights, or the remaining K unchanged parameters.

For definiteness, let us now complete the specification of the move probabilities by saying that when we split we choose each group with equal probability, and that when we merge we choose each pair of groups with equal probability. This can easily be modified. The move is now fully specified. For the sake of comparison with eq. (11) of Richardson & Green (1997), we give the complete acceptance probability, in the context of the univariate normal mixture problem using split and merge moves defined by Richardson & Green.

The probability for the split move is min(1, A), where A is

$$\text{(likelihood ratio)} \times \frac{\alpha B(n_-,\, n_+)}{(k+1)}$$
$$\times (k+1)\sqrt{\frac{\kappa}{2\pi}} \exp\left[-\frac{1}{2}\kappa\{(\mu_{j-}-\xi)^2+(\mu_{j+}-\xi)^2-(\mu_j-\xi)^2\}\right]$$
$$\times \frac{\beta^\gamma}{\Gamma(\gamma)}\left(\frac{\sigma_{j-}^2\sigma_{j+}^2}{\sigma_j^2}\right)^{-\gamma-1} \exp(-\beta(\sigma_{j-}^{-2}+\sigma_{j+}^{-2}-\sigma_j^{-2}))$$
$$\times \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \frac{g_{\omega+n_-,\omega+n_+}(w_-)}{g_{1,1}(w_-)g_{2,2}(u_1)g_{1,1}(u_2)}$$
$$\times \frac{\sigma_j(w_-\sigma_{j-}^2+w_+\sigma_{j+}^2)}{(w_-w_+)^{3/2}}$$

Note that by comparison to eq. (11) of Richardson & Green, the new terms are $\alpha B\,(n_-,\, n_+)/(k+1)$, the prior ratio for $(k+1)$ vs k, the additional factor $g_{\omega+n_-,\omega+n_+}(w_-)$ in the proposal ratio, and that the Jacobian has been expressed differently in the final line in the expression. The notation $g_{a,b}(\cdot)$ refers to the Beta(a, b) density. As usual, the acceptance probability for the merge move, providing the reverse of this split, is min(1, A^{-1}).

4.2. Comparison of samplers

The split/merge procedure defined above differs quite fundamentally from the approaches customarily used for computing MDP models, so it is of interest to draw comparisons. The methods we choose to compare are the "incremental" sampler of Bush & MacEachern (1996), and what we call the "augmentation" sampler, which is a variant on the proposals of MacEachern & Müller (1994, 1998) and which avoids the need for integration.

There is much current interest in comparing the performance of different variants of the augmentation sampler (Neal, 1998). Here we do not aim to be comprehensive in our comparisons and, building on the work of Neal, deliberately focus on comparing our split/merge reversible sampler to one version the augmentation sampler, chosen to represent the family of samplers introduced by MacEachern & Müller. Empirical comparisons are made in section 5, but some general points can be made here.

Each of the three methods: split/merge, incremental and augmentation are examples of hybrid MCMC methods, in which a portfolio of reversible moves, each maintaining detailed balance with respect to the target (posterior) distribution, is available, and these are used in cyclic fashion to form a sampler that is irreducible. In each case, one of the moves involves updating θ by sampling from its full conditional, thus conditioning in particular on k and z; this Gibbs update being available because of conjugacy.

The methods differ in their approaches to updating k and z, and in their amenability for use in a hierarchical setting, in which hyperparameters governing the prior for θ need to be updated.

In the incremental sampler, in place of the split/merge and allocation moves, k and z are updated implicitly, by drawing each ϕ_i in turn from its full conditional. Since this step may lead to either or both of a component being created or destroyed, irreducibility is attained. Also, θ is gradually updated during this process. Note that no separate move updating θ would be necessary for irreducibility, but that such a move is included in the portfolio to improve performance.

However, the full conditional for ϕ_i involves an off-line integration (see West *et al.* 1994), sometimes approximated by a Monte Carlo estimate; this integral (over the prior for a single θ_j) of course depends on values of hyperparameters for θ. Unless, therefore, conjugate hyperpriors

are used, the incremental method is cumbersome to use in the context of variable hyperparameters.

This difficulty is circumvented in the approach of MacEachern & Müller (1994, 1998). The idea is to draw one or more potential additional values of θ_j first, and only then to compute the probability that an observation is reassigned to such a new component—this probability does not involve any integral. In the "no-gaps" variant of their algorithm, a single additional component is created, while the "complete" variant uses a full set of n potential components. They do not propose simulating the additional $\{\theta_j\}$ anew for each observation considered.

We propose another variant on this idea, aimed at correctly simulating from the posterior distribution conditional on $d \leqslant d_{\max}$, where $d_{\max}$ is a fixed sufficiently large integer (we used $d_{\max} = 30$). We augment the θ vector once each sweep by generating $(d_{\max} - d)$ additional θ_j independently from G_0. The probabilities of assigning observation i to component j are analogous to eq. (9) of MacEachern & Müller (1994), but with n replaced by $d_{\max}$.

Neal (1998) suggests yet another variant in a similar spirit. His uses a fixed number m of additional components θ_j, which are re-simulated for every observation considered. There are complex trade-offs between the costs of generating extra variables, or introducing more serial dependence, which we will not pursue here.

Both the incremental and augmentation methods have the apparent disadvantage that new components are formed by moving one observation at a time, in contrast to the split/merge approach, in which a large but heterogeneous component can be split into two more homogeneous parts in one go. The augmentation method appears to carry an overhead, through the state space being extended to include $\{\theta_j\}$ not currently in the model. But it is difficult to quantify these factors in the abstract, and we therefore conduct comparative numerical experiments on the three samplers in the next section.

Finally, we observe that since all the moves mentioned maintain detailed balance, there is the potential for new methods to be devised that pool the best features of each of the current ones.

5. Comparative performance of the MCMC samplers for the DP model

We compare the performance of the MCMC samplers described in the previous section in the case of univariate normal mixtures. What summaries from a multidimensional posterior distribution are most useful is a matter for debate. We have chosen to concentrate our discussion on the output of two functionals: the degree of the partition and the deviance,

$$D(g) = -2 \sum_{i=1}^{n} \log \sum_{j} \frac{n_j}{n} f(y_i|\theta_j).$$

Monitoring the change in the degree against the number of sweeps is clearly an important characteristic of the samplers, while the deviance is used as a meaningful global function of all the parameters. Visual assessment of the burn-in period is helped by plotting the ergodic averages of the cumulative frequencies of degree of partition. The efficiency of the samplers in their stationary regime is characterized by computing, for each monitored functional, an estimate of the integrated autocorrelation time $\tau = \sum_{l=-\infty}^{\infty} \rho_l$, where ρ_l is the lag-l autocorrelation of the realized values of the functional. For the results below, we have used an adaptive window estimate of τ due to Sokal (see Besag & Green, 1993) which was calculated on the last 25,000 sweeps of long runs thinned by subsampling at the rate 1/20. For our comparisons, we have used three data sets, enzyme, acidity and galaxy, as well as the simulated data sets.

5.1. Comparison of the three samplers in the case of fixed hyperparameters

As commented by several authors, one of the shortcomings of the incremental sampler is the necessity of computing an integral of $f(\cdot|\theta)$ with respect to $G_0(\theta)$, which restricts its use mostly to fixed hyperparameter cases. For our comparison of the three samplers, we thus consider the following specification for the normal mixture model: $\alpha = 1$, $\theta = (\mu, \sigma^{-2})$ and $G_0 = N(\xi, \kappa^{-1}) \times \Gamma(\gamma, \beta)$ with fixed values of $\xi =$ midrange, $\kappa = 1/R^2$, $\gamma = 2$, $\beta = 0.02R^2$. We computed the required integral by adaptive 15-point Gauss–Kronrod quadrature.

Figure 6 shows a typical output for the cumulative frequencies of partition degrees of the galaxy data for the three samplers: incremental, augmentation and reversible-jump. Stability to the same posterior levels is achieved quickly for the three samplers, the incremental sampler having the shortest burn-in. In terms of running times, the incremental sampler – which does not separately update the allocations – is the fastest. The other two samplers update the allocations; unsurprisingly, we found the augmentation sampler to be approximately 4 times slower than the reversible-jump sampler. Our display plots correspond to approximately equivalent running times for 250,000, 50,000, and 200,000 sweeps of the incremental, augmentation and reversible-jump samplers respectively. On the three data sets (enzyme, acidity and galaxy), we found similar integrated autocorrelation times for the three samplers (between 0.9 and 1.7) on the deviance output. For the partition degree, we found a somewhat higher value of τ for the reversible-jump sampler (3 to 4) than for the other two samplers (1 to 1.7).

5.2. Comparisons between augmentation and reversible-jump samplers for a hierarchical DP model

It is of interest to compare the performance of our proposed reversible-jump sampler with that of the augmentation sampler in a situation with random hyperparameters. We thus modify the setting defined above to assume a random β which follows a $\Gamma(g, h)$ distribution with $g = 0.2$, $h = 10/R^2$ as before.

These two samplers are constructed on radically different principles. The augmentation sampler proposes new components containing only single observations, these will be accepted conditional on all other allocations if there is support from that data point and a prior which is not too tight. This construction suggests a mixing behaviour which could be influenced by the value of α (small αs correspond to tighter priors on low partition degree), in interaction with the shape of the mixture (well-separated or not). On the other hand, the proposal of the reversible-jump sampler is not influenced by α, its performance should not deteriorate for small αs, but with a high number of components containing fewer observations, the random splits might be less effective.

Figure 7 displays values of the integrated autocorrelation time for the "lepto" data set and values of α ranging from 0.1 to 2. We see clearly the difficulties encountered by the

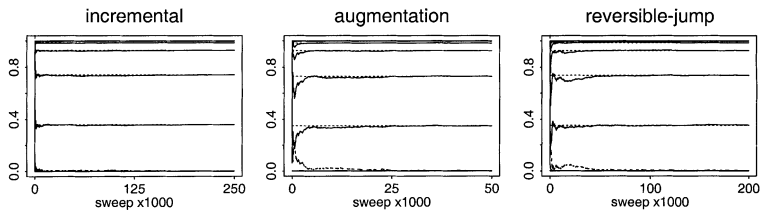

Fig. 6. Cumulative frequencies of partition degrees of the galaxy data for the three samples.

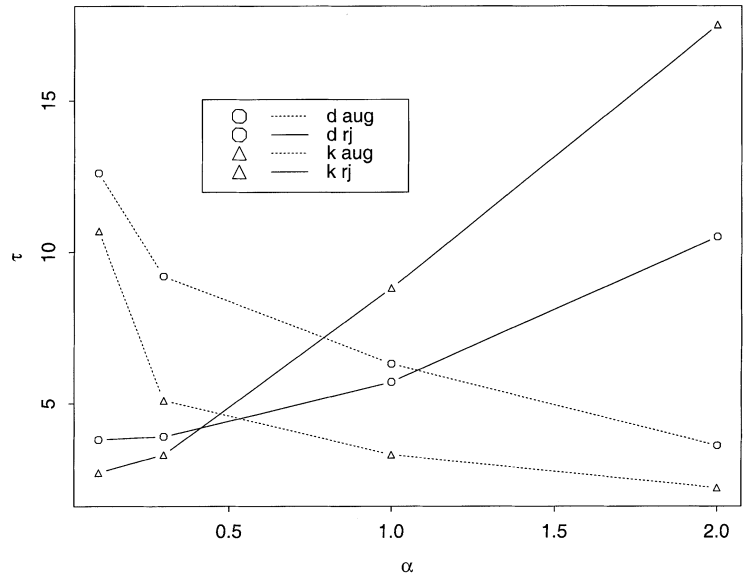

Fig. 7. Integrated autocorrelation time for "lepto" data set.

augmentation sampler when α is small and the data are not well-separated, which leads to high τs. (See also Figure 8 for an illustration of the differences in burn-in induced by small values of α). On the other hand, the reversible-jump sampler has somewhat opposite behaviour with higher τs for larger αs. Even then, the loss of efficiency is compensated by the faster running time of the reversible-jump sampler. Thus the reversible-jump sampler is competitive, not

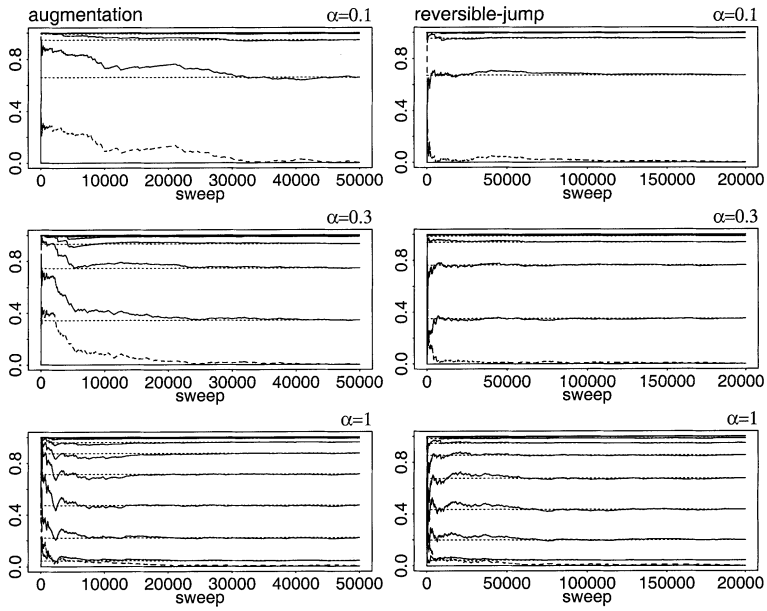

Fig. 8. Cumulative frequencies of partition degrees for "lepto" data set (number of sweeps adjusted for equivalent run time).

markedly superior overall, but mixes well in particular situations where it is known that DP type samplers become trapped.

As a final point, we recall that there is some freedom in designing the reversible-jump sampler which can be usefully exploited for tuning its performance. In particular, the Beta parameter ω of the "mock weights" can be adapted. All the results reported correspond to a default value of $\omega = 5$. For a model with a large number of components, this value might be too large. For example, with $\omega = 1$, we found for $\alpha = 2$ and the "lepto" data set that the values of τ were more than halved.

Efficient sampling of DP processes is an active area of research. Effective recursive approximations have been proposed by Newton *et al.* (1998) which are particularly useful in high dimensional space. The development of sequential importance sampling (MacEachern *et al.*, 1999) in connection with incremental type samplers is another way forward. Our simulations support our belief that the reversible-jump sampler, which derives from a different principle than incremental-like samplers, has the potential to be a useful addition to the menu of samplers for DP. It could be used in conjunction with other moves and/or importance sampling ideas.

Acknowledgements

We wish to thank Jim Berger, Guido Consonni, Peter Donnelly, Steve MacEachern, Peter Müller, Agostino Nobile, Tony O'Hagan, Sonia Petrone, and Mike West for stimulating discussions about this work. We are grateful to the referees and associate editor for their constructive comments, which helped to improve the paper. We acknowledge the financial support of the EPSRC Complex Stochastic Systems Initiative (PJG), INSERM and an EPSRC visiting fellowship (SR), and the ESF programme on Highly Structured Stochastic Systems. Sylvia Richardson is now Professor of Biostatistics at the Imperial College School of Medicine, London.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of mathematical functions*, 9th edn. Dover, New York.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *Ann. Statist.* **2**, 1152–1174.

Besag, J., Green, P. J., Higdon, D. & Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10**, 3–66.

Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika* **83**, 275–285.

Consonni, G. & Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *J. Amer. Statist. Assoc.* **90**, 935–944.

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

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

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

Lo, A. Y. (1984). On a class of Bayesian non-parametric estimates: (I) Density estimates. *Ann. Statist.* **12**, 351–357.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727–741.

MacEachern, S. N. & Müller, P. (1994). Efficient estimation of mixture of Dirichlet process models. Discussion paper 94-38, Institute of Statistics and Decision Sciences, Duke University.

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223–238.

MacEachern, S. N., Clyde, M. & Liu, J. S. (1999). Sequential importance sampling for non-parametric Bayes models: the next generation. *Canad. J. Statist.* **27**, 251–267.

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

Neal, R. M. (1992). Bayesian mixture modelling. In *Maximum entropy and Bayesian methods: Proceedings of the 11th international workshop on maximum entropy and Bayesian methods of statistical analysis*, (eds C. R. Smith, G. J. Erickson & P. O. Neudorfer), Seattle, 1991, pp. 197–211, Kluwer Academic, Dordrecht.

Neal, R. M. (1998) Markov chain sampling methods for Dirichlet process mixture models. http://www.cs.utoronto.ca/∼radford/

Newton, M. A., Quintana, F. A. & Zhang, Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical nonparametric and semiparametric Bayesian statistics* (eds D. Dey, P. Müller & D. Sinha), Lecture Notes in Statistics **133**, 45–61. Springer Verlag, New York.

O'Hagan, A. (1994). *Bayesian inference (Kendall's advanced theory of statistics, **2 B**)*, Wiley, New York.

Petrone, S. & Raftery, A. E. (1997). A note on the Dirichlet process prior in Bayesian non-parametric inference with partial exchangeability. *Statist. Probab. Lett.* **36**, 69–83.

Phillips, D. B. & Smith, A. F. M. (1996). Bayesian model comparison via jump diffusions. *In Practical Markov chain Monte Carlo* (eds W. R. Gilks, S. Richardson & D. J. Spiegelhalter), ch. 13, 215–239. Chapman & Hall, London.

Richardson, S. (1999). Contribution to the Discussion of paper by Walker *et al*. *J. Roy. Statist. Soc. Ser. B* **61**, 513–516.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion) *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.

Richardson, S., Leblond, L., Jaussent, I. & Green, P. J. (2000). Mixture models in measurement error problems, with reference to epidemiological studies. Technical report, INSERM.

Roberts, G. O., Petrone, S. & Rosenthal, J. S. (2000). Rates of convergence for markov chains associated with Dirichlet processes. *Far East J. Theoret. Statist*. To appear.

Spiegelhalter, D. J., Best, N. G. & Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Unpublished manuscript.

Walker, S. G., Damien, P., Laud, P. W. & Smith, A. F. M. (1999). Bayesian non-parametric inference for random distributions and related functions (with discussion). *J. Roy. Statist. Soc. Ser. B* **61**, 485–527. Technical report, Imperial College London.

Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. Ser. B* **62**, 159–180.

West, M., Müller, P. & Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of uncertainty: a tribute to Lindley* (eds A. F. M. Smith & P. Freeman), Wiley, New York.

Received February 1999; in final form July 2000

P. J. Green, Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

Appendix

General aspects of partition and allocation models

Here we attempt a more axiomatic approach to the specification of partition and allocation distributions.

There are two key properties possessed by both the DP and DMA models for partitions, that should also hold for any alternative model, namely:

(a) *exchangeability*: the probability of any partition should be invariant to any relabelling of the items;
(b) *heritability*: the model should remain self-consistent as the number of items increases, the probability assigned to a partition of a set of items being the same whether these are all the available items, or just a subset.

The first of these properties seems essential; the second is less vital, but desirable, the more so for some applications (eg. mixture modelling) than for others (eg. clustering).

The simplest way to ensure exchangeability is to work with a notation in which it is automatic. The maximal invariant of a partition under relabelling of the items is the set of group sizes, called here the signature. The signature can be written in a standard order, say listing the sizes in decreasing numerical order, to avoid double-counting. Thus the partitions $\{\{1, 2\}, \{3\}, \{4\}\}$ and $\{\{1\}, \{2, 4\}, \{3\}\}$ both have signature (2, 1, 1) and are assigned the same probability $q(2, 1, 1)$. Note that the signature $(n_1, n_2, \ldots, n_d)$ determines both the degree of the partition d and the number of items $n = \sum_j n_j$. Thus any exchangeable model for partitions is equivalent to a specification of the probabilities $q(n_1, n_2, \ldots, n_d)$ of any *single* partition with this signature. The only consistency condition required is that the probabilities sum to 1 for each number of items n. Let $m(n_1, n_2, \ldots, n_d)$ denote the number of partitions of $n = \sum_j n_j$ items with signature $(n_1, n_2, \ldots, n_d)$, then for each n we need

$$\sum m(n_1, n_2, \ldots, n_d) q(n_1, n_2, \ldots, n_d) = 1, \tag{10}$$

where the sum is over all sets of positive integers $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_d$ summing to n.

There is obviously great flexibility in choosing such models. The only constraint, (10), is easily imposed, especially as there is an explicit formula for the counts $m(n_1, n_2, \ldots, n_d)$, namely

$$m(n_1, n_2, \ldots, n_d) = \frac{n!}{n_1! n_2! \ldots n_d!} \frac{1}{\prod_r (\#j\colon n_j = r)!}.$$

For an exchangeable partition distribution, the necessary and sufficient condition on the qs for heritability is that for all signatures $(n_1, n_2, \ldots, n_d)$, the effect of adding one item maintains consistency, that is

$$\sum_{j=1}^{d} q(n_1 + \delta_{j1}, n_2 + \delta_{j2}, \ldots, n_d + \delta_{jd}) + q(n_1, n_2, \ldots, n_d, 1) = q(n_1, n_2, \ldots, n_d) \tag{11}$$

where δ is the Kronecker symbol (and note that addition of these may have disrupted the standard order in the signature). Any set of non-negative numbers $q(n_1, n_2, \ldots, n_d)$ satisfying (11) and the initial condition $q(1) = 1$ automatically satisfies (10): nothing else is needed to guarantee a proper, exchangeable, partition distribution.

We see that the heritability condition is much more demanding than exchangeability, as it imposes much more stringent constraints on the qs.

The DP and DMA models form familiar examples of allocation models that are both exchangeable and heritable. A third class possessing both properties is that of the partition models of Consonni & Veronese (1995), in which the degree d is drawn from a distribution of convenience (in fact, they use the form $p(d) \propto d^{-1}$ for $d = 1, 2, \ldots, n$), and then partitions drawn uniformly given d: $p(g|d) = \text{constant}$.

Recursive construction of partition distributions, and incremental samplers

A heritable exchangeable partition distribution can be constructed recursively, by considering the placement of the $(n+1)$th item conditional on the partition of the first n items, for $n = 1, 2, \ldots$. The recursion is started trivially with $q(1) = 1$. Given the partition g with signature $(n_1, n_2, \ldots, n_d)$ for the first n items, the additional item may join one of the existing

groups $j = 1, 2, \ldots, d$, or form a new group by itself. The probabilities $a_1, a_2, \ldots, a_d, b$, say, of these options are given by the corresponding terms on the left hand side of (11), divided by the sum.

For the DMA model we have

$$a_j = \frac{\delta + n_j}{k\delta + n}, \quad b = \frac{(k-d)\delta}{k\delta + n}, \tag{12}$$

on substituting from (4) into (11). For the DP model, we have

$$a_j = \frac{n_j}{\alpha + n}, \quad b = \frac{\alpha}{\alpha + n},$$

which can either be obtained explicitly from (1), or from (12) by letting $\delta \to 0$, $k \to \infty$ and $k\delta \to \alpha > 0$.

Another simple model is obtained by letting $\delta \to \infty$ in (12); we obtain

$$a_j = \frac{1}{k}, \quad b = 1 - \frac{d}{k}.$$

This corresponds to the symmetric multinomial model in which the z_i are drawn i.i.d. from the uniform distribution on $\{1, 2, \ldots, k\}$, that is, the items are allocated independently, equally likely to each of the groups.

Because of exchangeability, these recursive probabilities are equally appropriate for conditional distributions such as $p(z_i = j | z_{i'}, i' \neq i)$, which are needed in MCMC sampling item-by-item, such as in the "incremental" method, described in section 3.2. Unfortunately, we have not been able to derive incremental methods using these recursive probabilities for *posterior* simulation, except for the DP case.