Modelling heterogeneity with and without the Dirichlet process

Peter J. Green*
University of Bristol, UK.

Sylvia Richardson†
INSERM, France.

25 January 1999

Summary

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. Comparative 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, which exists 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.

Some key words: Allocation, Bayesian nonparametrics, 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, nonparametric regression, autoregression, survival analysis, etc. In many applications, DP priors are not used directly because of inconvenient discreteness. Instead, DP mixtures (MDP) are introduced, tacitly exploiting the discreteness but in an appealing hierarchical framework. More precisely, since a distribution realised from a DP is almost surely discrete, a random sample drawn from that realised distribution has positive probability of ties, and so provides a flexible model for clustering of items of various kinds in hierarchical models: random effects, parameters of sampling distributions, etc, an early version of which can be found in Lo (1984). Many applied nonparametric Bayesian model building has used this clustering property in the development of statistical methods based on MDP.

*Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: P.J.Green@bristol.ac.uk.
†INSERM U.170, 16 avenue Paul Vaillant Couturier, Villejuif, France.
Email: richardson@vjf.inserm.fr

The central role of the DP/MDP models in Bayesian nonparametrics, and the various lines of research stemming from it are recounted in the review article by Walker, *et. al.*(1998). 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 generalisations of DP, including the construction of nonparametric 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.

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 exhibiting the MDP model as a special case of a simple and familiar parametric model for mixtures, we reinforce the claim that in typical settings the DP models are far from the usual sense of nonparametric. 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 comparison principally in a univariate mixture context. We show in particular that the unbalancedness of the allocation distribution which exists in the prior DP model persists *a posteriori*. We also compare the two approaches in a latent variable set-up (Section 4).

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 5 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 6, 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 nonparametric 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 parameterise 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 such settings, a realisation 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. Whether a group is regarded as existing before any items are allocated to it depends on the statistical consideration of whether it is appropriate to consider the ϕ_i as arising from a population model or not.

2.1 Dirichlet process priors

One formulation for such a random vector ϕ is that using a Dirichlet process prior. For the sake of clarity, we briefly recall some general characteristics of the Dirichlet process, following Ferguson (1973). The Dirichlet process is a model for a random distribution function: given a positive real α and a distribution G_0 on a space Ω, we say

$$G \sim \mathrm{DP}(\alpha, G_0)$$

if for all Borel sets $(A_1, A_2, \ldots, A_m)$ partitioning Ω, and for all m,

$$(G(A_1), G(A_2), \ldots, G(A_m)) \sim \mathrm{D}(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_m))$$

where $D(\ldots)$ denotes the Dirichlet distribution with the specified parameters on the unit simplex in $\mathcal{R}^m$.

The DP model for ϕ is defined in two stages:

(a) a random distribution G is drawn from the Dirichlet process: $G \sim \mathrm{DP}(\alpha, G_0)$, 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 nonempty 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 a partial 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)

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

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

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

and

$$p(d) = \frac{\alpha^d c(n, d)}{\alpha(\alpha+1)\ldots(\alpha+n-1)} \tag{3}$$

where

$$c(n, d) = \sum_g \prod_{j=1}^d (n_j - 1)!,$$

is the absolute value of a Stirling number of the first kind, denoted $S_n^{(d)}$ by Abramowitz and 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 $p(k|\alpha)$; 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 $\mathrm{D}(\delta, \ldots, \delta)$), 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. (Another possibility, not further explored here, would have been to draw z given w conditional on there being no empty components. A component j is empty if $z_i \neq j \;\forall\, i$.)

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

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

and

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

on the simplex $\{w : w_j \geq 0, \sum_{j=1}^k w_j = 1\}$, 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

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

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 $1, 2, \ldots, k$, there are $k_{(d)} = k!/(k-d)!$ allocations z giving the same partition g of the items $1, 2, \ldots, n$, 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

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

2.3 Using the DP and DMA specifications in hierarchical models

The specification of a random n-vector ϕ in terms of α, δ and G_0 in the DMA model, or α and G_0 in the DP model, will form only a part of the full hierarchical model needed for our data. 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 parameterised respectively by $(\phi_1, \phi_2, \ldots, \phi_n)$. For instance, all six of the applications listed in MacEachern and Müller (1994) include this feature.

In the DP case, this setup frequently enjoys the misleading appellation of a "mixture of Dirichlet processes" (MDP) model, the terminology of DP mixture models used by West, Müller and Escobar (1994) being clearer. See O'Hagan (1994, pp. 288ff.) for further discussion.

At the top of the graph, the parameters α, δ and G_0 could in principle be fixed or random, and if random possibly in turn 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, Erkanli and West (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 and Green (1997), α is fixed, and G_0 set to be Normal(ξ, κ)-inverse gamma(γ, β) where only β is random, with a gamma hyperprior. We return to this setup 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.

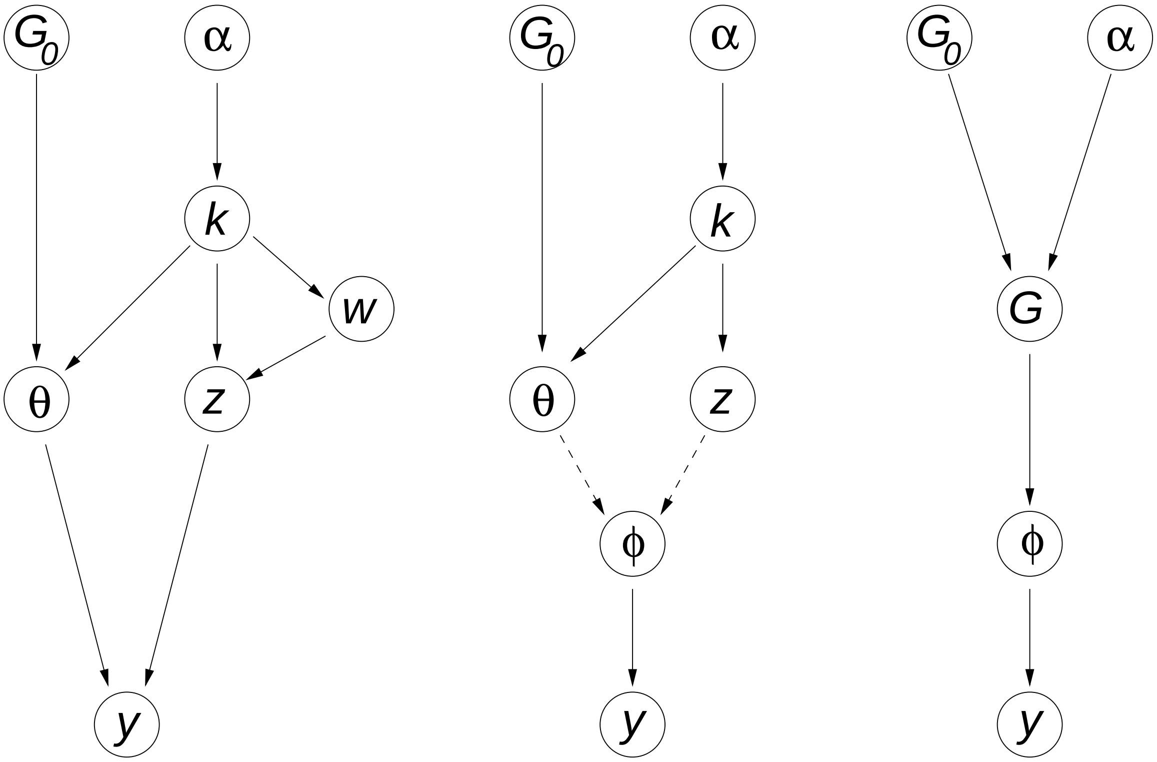

Figure 1: Directed acyclic graphs representing the DMA models (left and centre) and the DP model (right).

It may be helpful to examine the relationship between the allocation models in terms of Directed Acyclic Graphs (DAGs). Figure 1 shows, on the left, part of the DAG for a DMA allocation model. In the centre, ϕ is shown explicitly, and w integrated out, while on the right we see the DAG for the corresponding DP based model. Comparing the last two, we see that the DP and DMA models are simply providing different specifications for $p(\phi|G_0, \alpha)$.

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 equations (1) and (4). 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 (4) clearly converges to (1). Thus, so far as the occupancy of non-empty

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 $< m >$ 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		2
2	11α	(123)(4)	$< 4 >$	2α	1112	$< 2 >$	α
		(12)(34)	$< 3 >$	α	1122	$< 2 >$	$\alpha/2$
3	$6\alpha^2$	(12)(3)(4)	$< 6 >$	α^2	1123	$< 6 >$	$\alpha^2/6$
4	α^3	(1)(2)(3)(4)		α^3	1234	$< 24 >$	$\alpha^3/24$

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 clear prior statement of this connection to the DP process.

Alternatively, consider the DMA partition distribution (4) under the condition that *there are no empty components*, that is, $d = k$ or equivalently $n_j > 0 \, \forall j$. We find

$$p(g|k, \delta, d = k) = \frac{\prod_{j=1}^{d} \Gamma(\delta + n_j)}{\sum_g \prod_{j=1}^{d} \Gamma(\delta + n_j)}.$$

On letting $\delta \to 0$, this converges to (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 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 under either model to the partitions (123)(4) and (12)(34). Under the DP model, each partition of the pattern (123)(4) is 2 times 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 4 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. We start by introducing some predictive criteria that are relevant for model comparisons between DMA and DP formulations.

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 $< m >$ 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	$< k >$	a
2	$k_{(2)}(4b+3c)$	(123)(4)	$< 4 >$	$k_{(2)}b$	1112	$< k_{(2)} >$	b
		(12)(34)	$< 3 >$	$k_{(2)}c$	1122	$< k_{(2)} >$	c
3	$6k_{(3)}d$	(12)(3)(4)	$< 6 >$	$k_{(3)}d$	1123	$< k_{(3)} >$	d
4	$k_{(4)}e$	(1)(2)(3)(4)		$k_{(4)}e$	1234	$< k_{(4)} >$	e

3.1 Predictive distributions and deviances

Predictive distributions are key quantities on which to base model comparison for non-nested models, as is the case for DP and DMA. Here, we are not interested in prediction *per se*, but in using predictive density estimates in a measure of fit to be specified. Note that several 'types' of predictive densities for a new observation, corresponding to different conditionings, can be constructed. Since it is not possible to calibrate $p(k|\alpha)$ and $p(d|n,\alpha)$ *a priori* to have the same impact on the respective models, we will always condition our density estimates either on k or on the degree d. We stress that this conditioning, which is not natural if one adopts the usual DP perspective, is made here in order to uncover specific features to the performance of both models.

Let us first consider the predictive density for a new observation y^* given the data, $p(y^*|y,k)$. The extension of the DAG for the DMA mixture model to include y^* and its allocation z^* is shown in Figure 2. Using the fact that y and $y*$ are conditionally independent given θ, z and z^*, it can be shown that

$$p(y^*|y,k) = E\left[\sum_j w_j f(y^*|\theta_j) \,\middle|\, y,k\right]. \tag{5}$$

It will be convenient to write this function as $\widehat{g}_k(y^*)$, 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 (5):

$$\widehat{g}_d(y^*) = E\left[\sum_j \frac{n_j}{n} f(y^*|\theta_j) \,\middle|\, y,d\right]. \tag{6}$$

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

$$G(h) = -2\sum_{i=1}^{n} \log(h(y_i)). \tag{7}$$

(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 nonparametric setting, such a baseline would be $-\infty$.) By computing $G(\widehat{g}_k)$ and $G(\widehat{g}_d)$, we

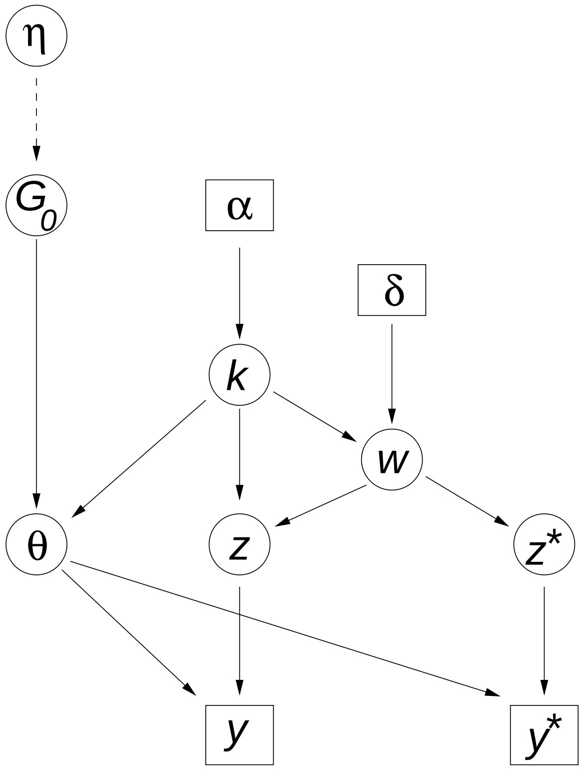

Figure 2: DAG of the DMA mixture model with new observation y^*.

thus get a global measure of quality of fit of the point estimates of the predictive densities $\widehat{g}_k$ and $\widehat{g}_d$ given by the DMA and DP models respectively.

It is also of interest to understand the variability around these point estimates. 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 $G(g)$ as defined in (7). Note that we have integrated out the w's to facilitate comparability between DMA and DP. Similarly to (5), 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 $G(g)$ given k or d, in particular its mean and variability. Note that, by Jensen, we always have $E(G(g)) \geq G(\widehat{g}_k)$ and that the difference $E(G(g)) - G(\widehat{g}_k)$ 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 parameterisation. 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 3 data sets, enzyme, acidity and galaxy, described in Richardson and Green (1997), as well as 4 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 standardised scale, in Figure 3. 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 and 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 simulation, 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 5.1; similarly for the DMA model, 100 000 sweeps of the reversible jump algorithm presented in Richardson and Green (1997) were used after a burn-in of 100 000 sweeps.

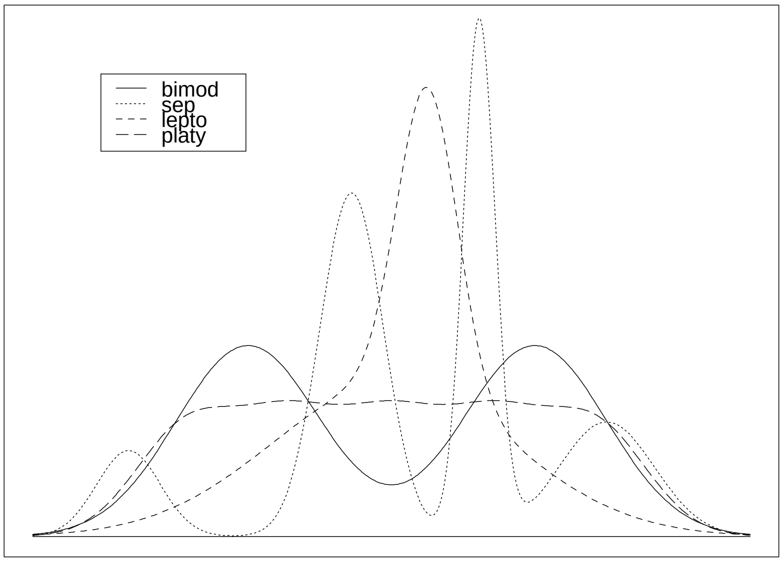

Figure 3: Plots of the bimod, sep, lepto and platy normal mixture density functions, translated and scaled to have similar ranges.

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 $p^*(k|y)=p(k|y)/p(k)$ and $p^*(d|y)=p(d|y)/p(d)$ corresponding to uniform priors for k and d in the two models.

Figure 4 plots the cumulative distribution of $p^*(k|y)$ versus $p^*(d|y)$ for the 7 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.

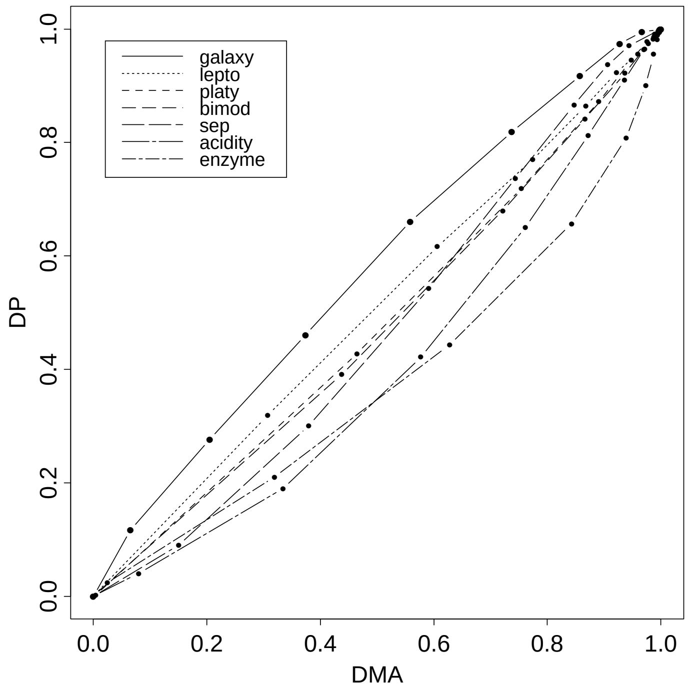

Figure 4: Modified cumulative posterior distributions for the number of components, compared for the DMA and DP models using PP plots.

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 and Raftery (1997) with particular reference to change point models. We can summarise equality of allocation by the *entropy* $-\sum_j (n_j/n)\log(n_j/n)$ and look at the conditional posterior of entropy given degree d, interpreted for DMA models as the number of non-empty components.

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 Figure 5). 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 6 presents a typical comparison; the lack of balance is more noticeable as the degree increases. Hence, as was also noted by Petrone and Raftery, in most cases the unbalancedness of the prior allocation model is still noticeable in the posterior.

To illustrate further the difference of the induced partitions, it is also of interest to investigate posterior classification for the data conditional on d and a declared unambiguous labelling, for example by ordering of the means. This can be done either within-sample or predictively. Let y^* be a future observation with allocation variable z^*. As detailed in Richardson and Green

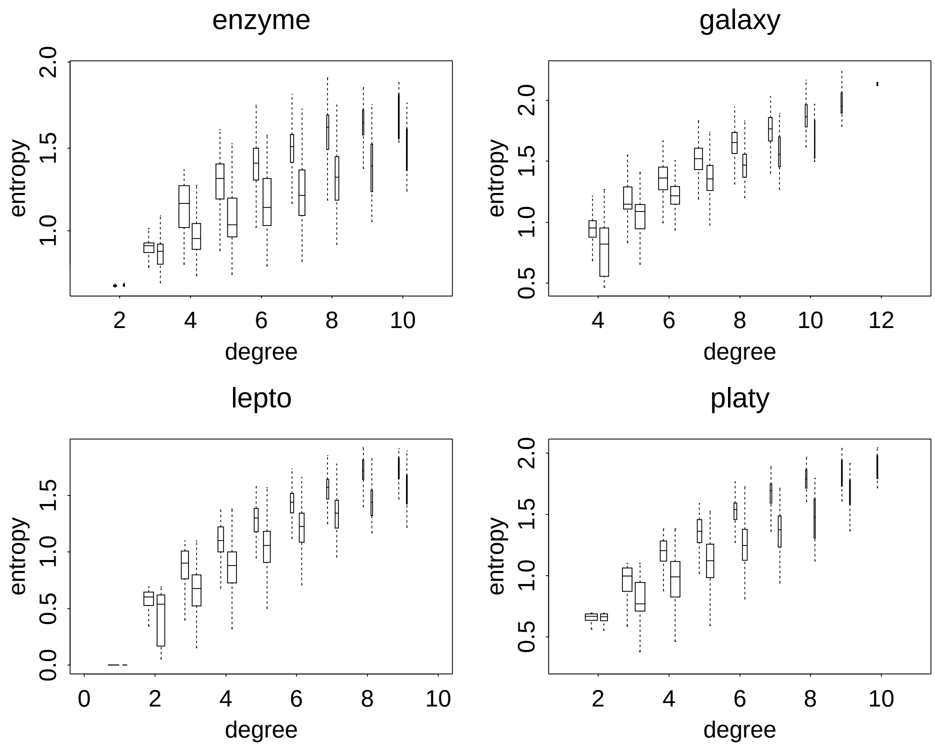

Figure 5: Conditional distribution of entropy given degree, for DMA and DP mixture models applied to four data sets.

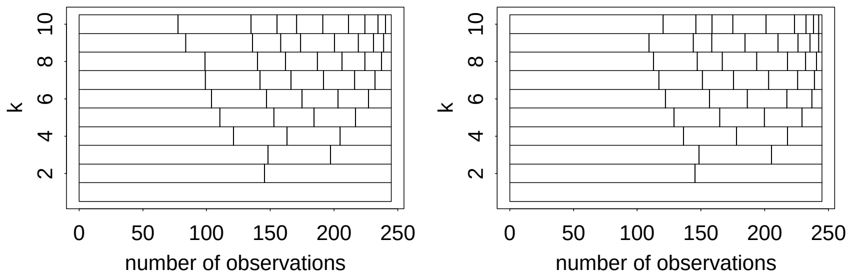

Figure 6: Mean group sizes for DMA and DP mixture models, enzyme data.

Table 3: Enzyme data: deviances associated with predictive densities derived from DMA or DP models.

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

(1997), $p(z^* = j|y, y^*, k)$ can be estimated from the MCMC run for the DMA model by averaging $w_j \phi(y^*; \mu_j, \sigma_j) / \sum_{j=1}^{k} w_j \phi(y^*; \mu_j, \sigma_j)$ where $\phi(\cdot\,; \mu, \sigma)$ is the normal density. For the DP model, the same expression is used with w_j replaced by n_j/n. Figure 7 exhibits some of the most striking differences we found on our data sets. For the enzyme data set, note that the DMA model assesses a noticeable lower limit (below 1.5) for the the sub-group classified as high enzymatic metabolisers than the DP model. When classifying the acidity data into 4 groups, it is interesting to see that the left hand outlying observation is treated quite differently under the two models, the DP model placing this single observation in one component.

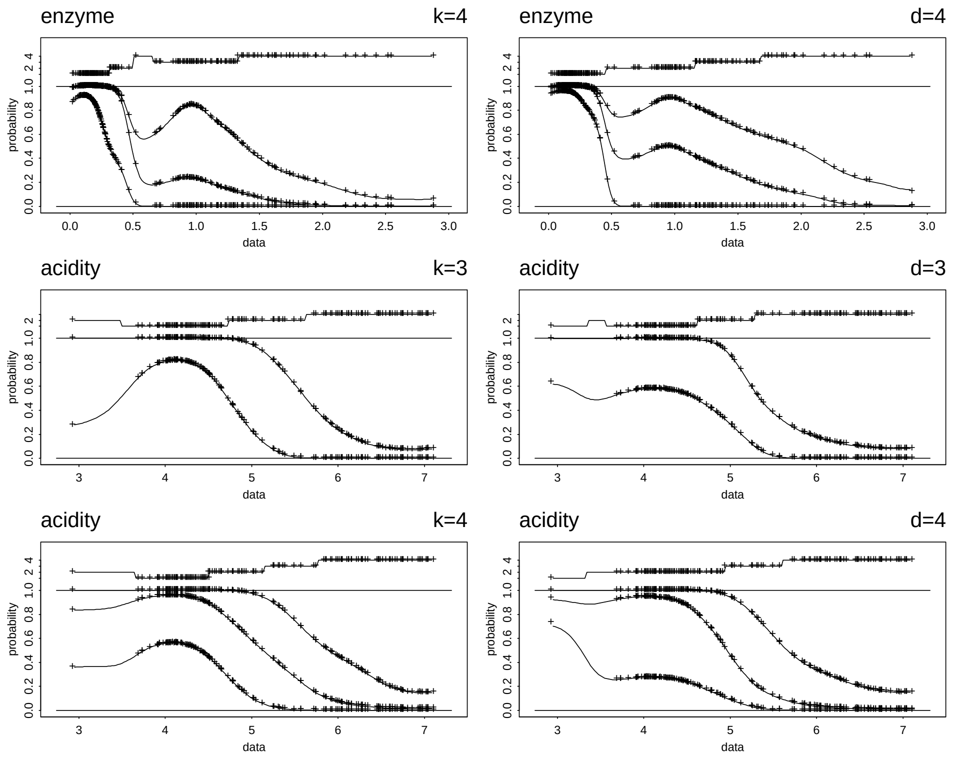

Figure 7: Posterior classification for DMA and DP models on enzyme and acidity data.

3.5 Deviances

We computed $G(\widehat{g}_k)$ and $G(\widehat{g}_d)$ as defined in (5), (6) and (7) 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 $G(\widehat{g}_k)$ but with few differences exceeding 1. The only notable difference in fit concerns the enzyme data (see Table 3), a data set for which we had already noticed that the induced partitions and classification differ markedly between the DMA and the DP models.

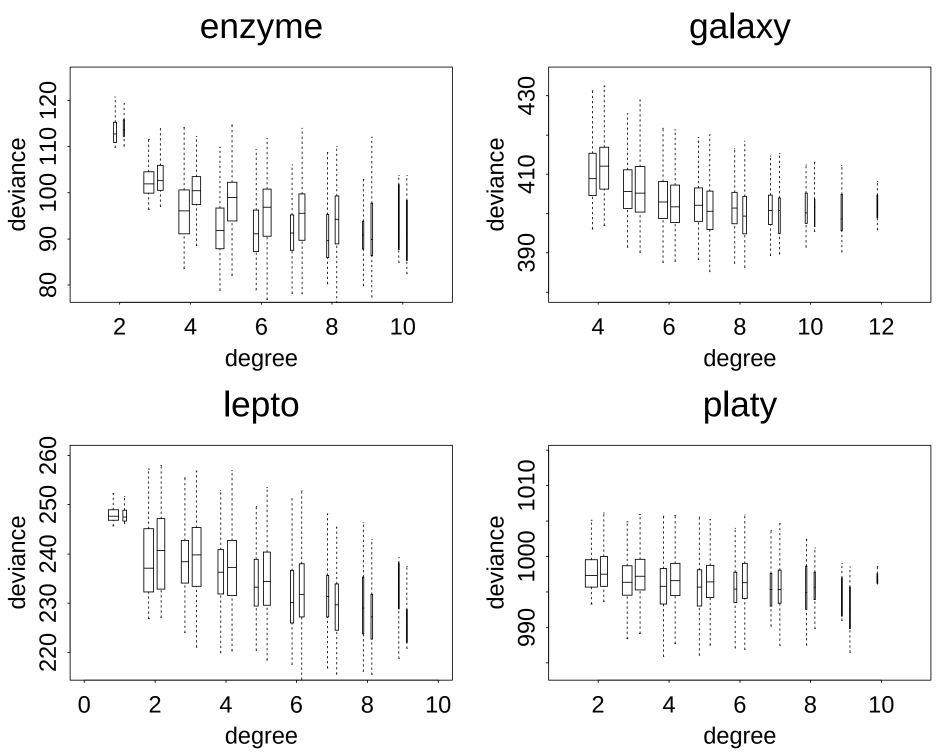

Figure 8: Distributions of deviance $G(g)$: comparison between DMA and DP models.

4 DP and DMA models as prior distributions in measurement error problems

A common use for DP models is as a class of nonparametric priors in hierarchical models or in models with latent variables. Here, we present a brief comparison of DMA and DP models in the context of measurement error analysis. Measurement error problems are concerned with inference on regression coefficients for an outcome y in terms of covariates x, in cases where x is not measured accurately on all subjects, but information on x is available through the recording of an imperfect surrogate u.

Bayesian analysis of measurement error problems is based on *structural* and *functional specifications*. Structural specifications entail the formulation of three submodels: an *outcome model* relating x and y, a *measurement model* relating u and x and a *prior model* $\pi(\cdot)$ for the distribution of x. These three submodels are combined through a DAG using the fundamental assumption that y is conditionally independent of u given x, an assumption which is also referred to as *non-differential measurement error*. Information on the measurement error process can be built into the graphical model in a flexible way. Typical designs involve validation groups, that is, subgroups of subjects, usually small, where both x and u are recorded. At a second stage, functional forms for the distributions involved in the submodels are chosen to reflect knowledge of the underlying outcome and measurement process. Fully Bayesian implementation of such models has been discussed by Richardson and Gilks (1993) and illustrated in the context of epidemiological studies.

An important element of the analysis is the modelling of $\pi(x)$. In observational studies, heterogeneity in the distribution of x among the population studied is common and it would be somewhat restrictive to assume a particular parametric shape for $\pi(x)$. In epidemiology, the heterogeneity of the distribution of the latent exposure variable x follows, for example, from the different genetic, socio-demographic or cultural composition of the population studied. Recent approaches aiming at flexible specification of $\pi(x)$ involve maximum likelihood with modelling of $\pi(x)$ using nonparametric mixture distributions (Roeder, Carroll and Lindsay, 1996) or a Bayesian framework with a Dirichlet process prior model for the unknown joint distribution of x and u (Müller and Roeder, 1997).

We have implemented a measurement error model with validation group where $\pi(x)$ is modelled either by a DMA or a DP mixture. Specifically, we consider the following setting: a logistic outcome model for a dichotomous y and a Gaussian measurement error model:

$$\text{logit}\{P(y=1|x)\} = \beta_0 + \beta_1 x, \qquad u \sim N(x, \theta^{-1}).$$

For the study design, we suppose that x, u and y are observed on a subset of n_0 individuals (the validation group), while only u and y are observed in a larger group of size n_1 (the main study). Detailed formulation of the graphical model, the prior settings and the algorithms will be presented in a forthcoming paper.

Table 4 summarises the results of a small simulation study comparing the performances of using either DP or DMA models for π, on the estimation of the measurement error precision θ and the regression slope β_1. 50 replicate data sets were used. The simulation model for generating the unknown values of x was that of a mixture of 3 normals: $0.6N(0.19, (0.08)^2) + 0.2N(1.05, (0.2)^2) + 0.2N(1.63, (0.48)^2)$, chosen to have an asymetric bimodal shape with a long tail, to mimic a common feature of many exposure distribution found in epidemiological studies . We chose $n_0 = 30$ and $n_1 = 300$ and a logistic model with $\beta_0 = -0.8$, and $\beta_1 = 0.4$. From Table 4, we see that θ and β_1 are both well-estimated, with similar results for DMA and DP in terms of mean square error for the regression coefficient. We also compared DP and DMA posterior estimation of the latent variable. We found that the posterior means were nearly identical, but that in some cases the DP model led to higher variability of the latent variables for some extreme observations (Figure 9). This is well in keeping with our previous observations on the treatment of extreme observations by the DP.

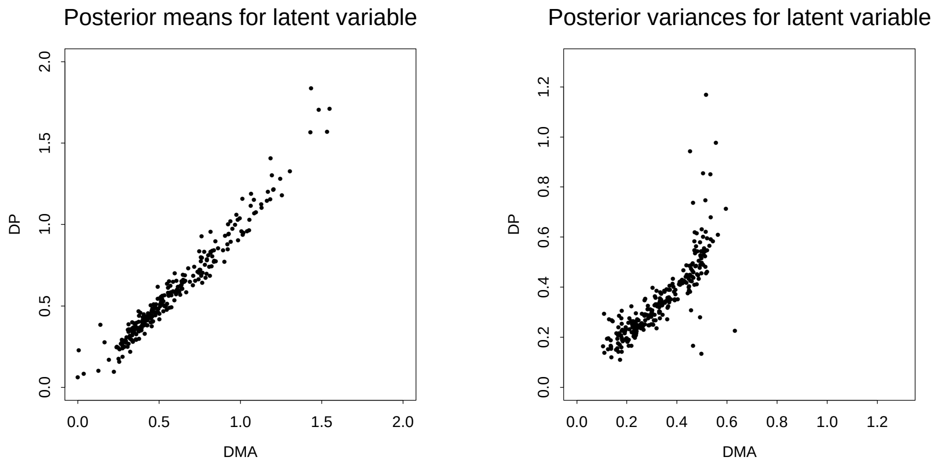

Figure 9: Posterior mean and variance for the latent variable x in one of the data sets of Table 4: comparison of DP and DMA estimation.

Table 4: Performance of DMA and DP models as priors in a measurement error model, from a simulation study with 50 replicates.

		Analysis	
True θ		DMA	DP
0.9	$\overline{\theta}$	0.92 (0.11)	0.93 (0.11)
	$\overline{\beta_1}$	0.47 (0.33)	0.46 (0.33)
	mse(β_1)	0.08	0.08
0.5	$\overline{\theta}$	0.51 (0.05)	0.51 (0.05)
	$\overline{\beta_1}$	0.40 (0.40)	0.41 (0.39)
	mse(β_1)	0.07	0.07

5 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 and West (1995), and MacEachern and 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 Petrone *et al.* (1998). In contrast, Richardson and 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 and Green used empty-component birth/death moves in addition to the splits and merges (and Phillips and Smith (1994) implemented general birth and deaths 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 very general applicability, and 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 5.2, we draw some comparisons between this new sampler and two existing methods, one of which is also suitable for non-conjugate MDP models.

5.1 Split/merge samplers for allocation models

Consider a general DP model, with a p-dimensional parameter $\theta \in \mathcal{R}^p$. 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, based on an intuitive notion of probable acceptance, 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 and 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 $m : \mathcal{R}^p \to \mathcal{R}^p$ is invertible. (For example in mixture density estimation, $m_l(\theta)$ might be the l^{th} 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 $\mathcal{R}$, we might use (9) together with

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

which evidently satisfies (9). 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

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

and this time the Jacobian reduces to

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

In fact for the normal mixture application Richardson and 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. So they use

$$\begin{aligned}
\mu_{j-} &= \mu_j - u_1\sigma_j\sqrt{\frac{w_+}{w_-}} \quad , \quad \mu_{j+} = \mu_j + u_1\sigma_j\sqrt{\frac{w_-}{w_+}} \\
\sigma^2_{j-} &= u_2(1-u_1^2)\sigma_j^2/w_- \quad , \quad \sigma^2_{j+} = (1-u_2)(1-u_1^2)\sigma_j^2/w_+
\end{aligned}$$

where $\theta=(\mu,\sigma^2)$, $m_1(\theta)=\mu$ and $m_2(\theta)=\mu^2+\sigma^2$. For this transformation, the Jacobian is

$$\frac{\sigma_j(w_-\sigma^2_{j-} + w_+\sigma^2_{j+})}{(w_-w_+)^{3/2}}$$

They draw $u_1 \sim Be(2,2)$ and $u_2 \sim U(0,1)$.

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, and it may be of interest, for comparison with the corresponding expression (equation(11)) of Richardson and Green (1997), to give the complete acceptance probability, in the context of the univariate normal mixture problem.

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

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

In this expression, the second and third lines, being the prior ratios for parameters given k, are unchanged from the DMA version as given in equation(11) of Richardson and Green (1997). The new terms are $\alpha B(n_-, n_+)/(k+1)$, the prior ratio for $(k+1)$ versus k, the additional factor $g_{\omega+n_-,\omega+n_+}(w_-)$ in the proposal ratio, and the new Jacobian of transformation (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 $\min(1, A^{-1})$.

5.2 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 West, Müller and Escobar (1994), and what we call the "augmentation" sampler, which is a variant on the proposals of MacEachern and Müller (1994, 1998). Empirical comparisons are made in Section 6, 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, Müller and Escobar (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 and 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. It is not fully clear from their description whether they 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 $d \leq d_{\max}$, 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 equation (9) of MacEachern and 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.

6 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 debatable question. We have chosen to concentrate our discussion on the output of two functionals: the degrees of the partition and the deviance,

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

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 degrees of partition. The efficiency of the samplers in their stationary regime is characterised 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 realised values of the functional. For the results below, we have used an adaptive window estimate of τ due to Sokal (see Besag and Green, 1993) which was calculated on the last 25000 sweeps of long runs thinned by subsampling at the rate 1/20. For our comparisons, we have used 3 data sets, enzyme, acidity and galaxy, as well as the simulated data sets.

6.1 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 $f(.|\theta)$ with respect to $G_0(\theta)$, which restricts its use mostly to fixed hyperparameter cases. For our comparison of the 3 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.

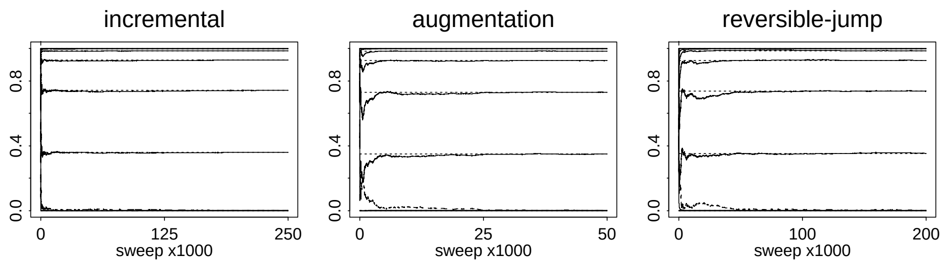

Figure 10: Cumulative frequencies of partition degrees of the galaxy data for the three samplers.

Figure 10 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).

6.2 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 α's 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 α's, but with a high number of components containing fewer observations, the random splits might be less effective.

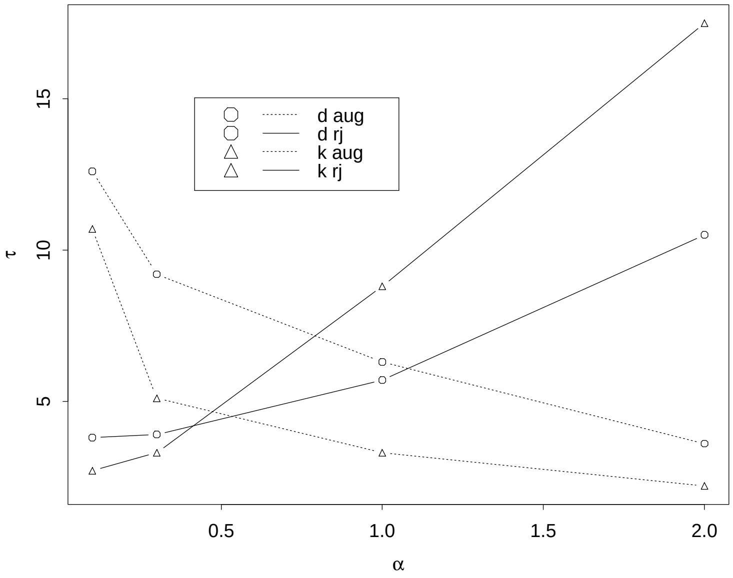

Figure 11: Integrated autocorrelation time for "lepto" data set.

Figures 11 and 12 display values of the integrated autocorrelation time for two deliberately contrasting data sets ("lepto" and galaxy) and values of α ranging from 0.1 to 2. On the "lepto" data set, we see clearly the difficulties encountered by the augmentation sampler when α is small and the data is not well-separated, which leads to high τ's. (See also Figure 13 for an illustration of the differences in burn-in induced by small values of α). On the galaxy data set, the performance of the augmentation sampler is less influenced by α, but again worse when α is small. On the other hand, the reversible-jump sampler has somewhat opposite behaviour with higher τ's for larger α's. Altogether Figures 11 and 12 show satisfactory performance for the reversible-jump sampler. The integrated autocorrelation times are smaller than or equal to those of the augmentation sampler in about half the cases. When they are larger, they are below 4 times those of the augmentation sampler, except in one case, so the loss of efficiency is compensated by the faster running time of the reversible-jump sampler. Thus the reversible-jump sampler is competitive, not markedly superior overall, but mixes well in particular situations where it is known that DP type samplers

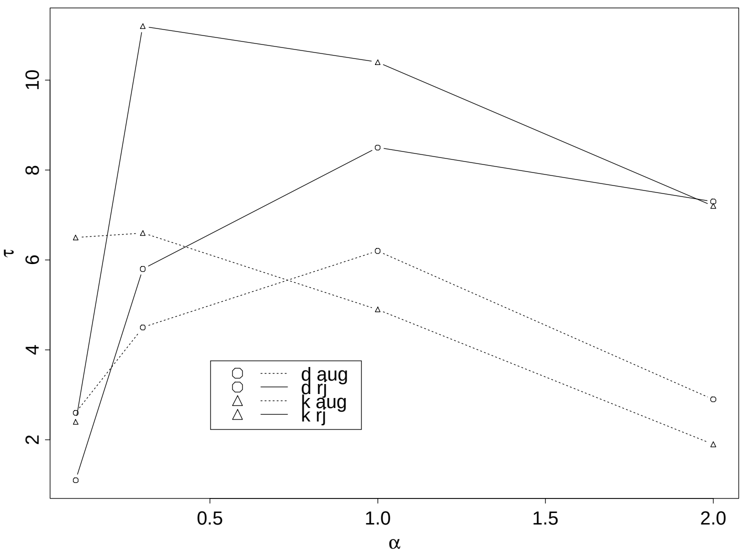

Figure 12: Integrated autocorrelation time for galaxy data set.

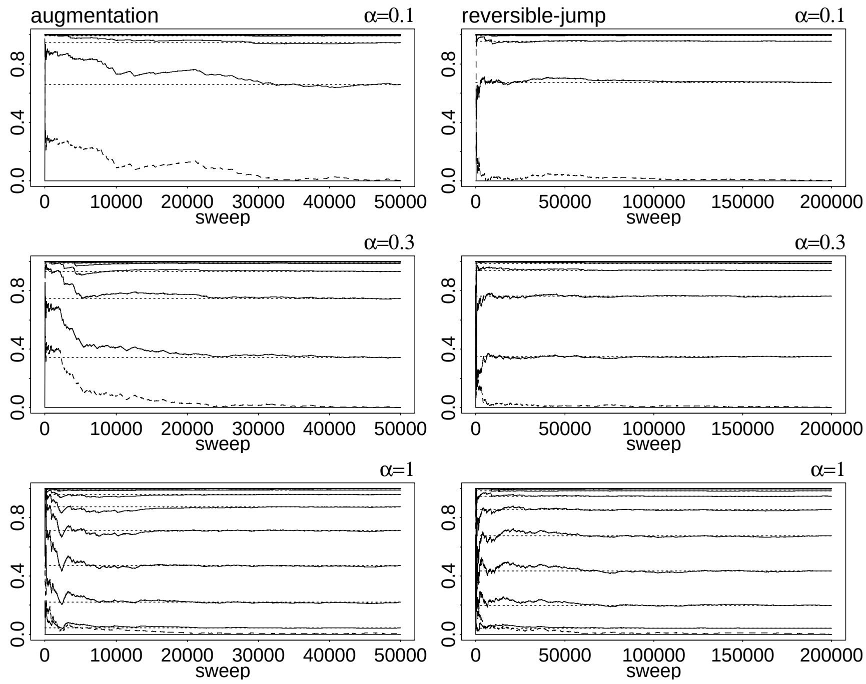

Figure 13: Cumulative frequencies of partition degrees for "lepto" data set (number of sweeps adjusted for equivalent run time).

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.

In this section, we have focussed on comparisons between the augmentation sampler, chosen to represent the family of samplers introduced by MacEachern and Müller, and our split/merge reversible-jump sampler, both of which avoid the need for integration. There is much current interest in comparing the performance of different variants of the augmentation sampler (Neal, 1998), as well as developing other approaches to efficient sampling of DP processes, see for example Newton, *et. al.*(1998), Liu, *et. al.*(1996). As was shown in our simulations, we believe that the reversible-jump sampler, which derives from a different principle, has the potential to be a useful addition to the menu of samplers for DP.

Acknowledgements

We wish to thank Jim Berger, Guido Consonni, Peter Donnelly, Steve MacEachern, Peter Müller, Agostino Nobile, Tony O'Hagan and Mike West for stimulating discussions about this work, and Isabelle Jaussent and Laurent Leblond for assistance with the measurement error study. 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.

References

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

Antoniak, C. E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**, 1152–1174.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society*, B, **55**, 25–37.

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

Consonni, G. and Veronese, P. (1995) A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, **90**, 935–944.

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

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

Green, P. J. (1994) Contribution to the discussion of paper by Grenander and Miller. *Journal of the Royal Statistical Society*, B, **56**, 589–590.

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

Lindsay, B. G. (1995) *Mixture models: theory, geometry, and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics: Hayward, California.

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

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

MacEachern, S. N. and 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., Clyde, M. and Liu, J.S. (1996) *Sequential importance sampling for nonparametric Bayes models: the next generation.* Discussion paper 96-09, Institute of Statistics and Decision Sciences, Duke University.

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

Mengersen, K. and Robert, C. (1996) Testing for mixtures: a Bayesian entropy approach. In *Bayesian Statistics 5*, J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds., Oxford: Oxford University Press.

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

Müller, P. and Roeder, K. (1997) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, 523–537.

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

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. and Zhang, Y. (1998) Nonparametric Bayes methods using predictive updating. In D. Dey, P. Müller, and D. Sinha (editors) *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lecture Notes in Statistics 133, pp. 45-61, New-York: Springer.

Nobile, A. (1994) *Bayesian analysis of finite mixture distributions.* Ph. D. thesis, Carnegie Mellon University.

O'Hagan, A. (1994) *Bayesian Inference (Kendall's Advanced theory of Statistics,* **2 B***)*, Wiley, New York.

Petrone, S. and Raftery, A. E. (1997) A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statistics and Probability Letters,* **36**, 69–83.

Petrone, S., Roberts, G. O. and Rosenthal, J. S. (1998) *A note on convergence rates of Gibbs sampling for nonparametric mixtures*, Unpublished manuscript.

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

Richardson, S. and Gilks, W. R. (1993) Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, **12**, 1703–1722.

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

Robert, C. (1996) Mixtures of distributions: inference and estimation. Chapter 24 (pp. 441–464) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in variables. *Journal of the American Statistical Association*, **91**, 722–732.

Roeder, K. and Wasserman, L. (1995) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

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

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1998) *Bayesian nonparametric inference for random distributions and related functions.* Technical report, Imperial College London.

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

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 (for example, mixture modelling) than for others (for example, 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 $n_1 \geq n_2 \geq \ldots \geq n_d$ 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

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

For an exchangeable partition distribution, the necessary and sufficient condition on the q's 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 q's.

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 and 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 $(n+1)^{\text{th}}$ 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

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

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

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

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

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

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.