Controlling the reinforcement in Bayesian non-parametric mixture models

Antonio Lijoi,

Università degli Studi di Pavia, Italy

Ramsés H. Mena

Universidad Nacional Autónoma de México, Mexico City, Mexico

and Igor Prünster

Università degli Studi di Torino, Collegio Carlo Alberto and International Centre for Economic Research, Turin, Italy

Summary. The paper deals with the problem of determining the number of components in a mixture model. We take a Bayesian non-parametric approach and adopt a hierarchical model with a suitable non-parametric prior for the latent structure. A commonly used model for such a problem is the mixture of Dirichlet process model. Here, we replace the Dirichlet process with a more general non-parametric prior obtained from a generalized gamma process. The basic feature of this model is that it yields a partition structure for the latent variables which is of Gibbs type. This relates to the well-known (exchangeable) product partition models. If compared with the usual mixture of Dirichlet process model the advantage of the generalization that we are examining relies on the availability of an additional parameter σ belonging to the interval (0,1): it is shown that such a parameter greatly influences the clustering behaviour of the model. A value of σ that is close to 1 generates a large number of clusters, most of which are of small size. Then, a reinforcement mechanism which is driven by σ acts on the mass allocation by penalizing clusters of small size and favouring those few groups containing a large number of elements. These features turn out to be very useful in the context of mixture modelling. Since it is difficult to specify *a priori* the reinforcement rate, it is reasonable to specify a prior for σ. Hence, the strength of the reinforcement mechanism is controlled by the data.

Keywords: Bayesian clustering; Bayesian non-parametric inference; Dirichlet process; Mixture model; Predictive distribution; Product partition model

1. Introduction

In Bayesian hierarchical mixture models with an unknown number of components, the analysis of the distributional properties of the number of clusters in the data is a key issue. Taking a non-parametric approach implies assuming a potentially infinite number of clusters in an infinite sequence of exchangeable observations and this yields greater modelling flexibility. A recent interesting review concerning the potential of the Bayesian approach in mixture modelling is provided in Marin *et al.* (2005). The most widely used non-parametric hierarchical mixture model is the mixture of Dirichlet process (MDP) model that was introduced by Lo (1984). A random discrete probability distribution, such as the Dirichlet process, exploited as a mixing measure in

Address for correspondence: Antonio Lijoi, Dipartimento di Economia Politica e Metodi Quantitativi, Università degli Studi di Pavia, Via S. Felice 7, 27100 Pavia, Italy.
E-mail: lijoi@unipv.it

the model is an essential tool for modelling the clustering behaviour. Indeed, the occurrence of ties at the higher level of the hierarchy induces a clustering structure within the observed data. One might wonder how a specific choice of the driving random discrete distribution affects the clustering mechanism. In this respect, it is worth mentioning that various new classes of discrete priors generalizing the Dirichlet process have been introduced recently. Among them we recall species sampling models (Pitman, 1996), dependent Dirichlet processes (MacEachern, 1999), generalized stick breaking priors (Hjort, 2000; Ishwaran and James, 2001), Poisson–Kingman models (Pitman, 2003), normalized random measures with independent increments (Regazzini *et al.*, 2003) and spatial neutral to the right models (James, 2006). Hence, a natural extension of the MDP model is obtained by replacing the Dirichlet process with specific priors that are contained in the classes that were listed above. For example, Ishwaran and James (2001, 2003) investigated formal properties of mixture models based on species sampling priors and specifically examined the mixture of the two-parameter Poisson–Dirichlet process due to Pitman (1995). In Lijoi *et al.* (2005) a mixture model based on the normalization of an inverse Gaussian process was considered and a comparison with the clustering structure that is induced by the MDP was drawn. Mixtures based on spatial neutral to the right processes were dealt with in James (2007). Another extension of the MDP was also provided in De Iorio *et al.* (2004), who made use of the dependent Dirichlet process.

The clustering structure that is induced by these classes of models has not been deeply investigated. An exception is represented by the MDP model for which the partitions generated have been studied in Petrone and Raftery (1997) and Green and Richardson (2001). Indeed, in the Dirichlet case the only free parameter which can be used to tune the distribution of the number of distinct components is the total mass of the base-line measure. Typically a prior distribution is specified for such a parameter to smooth the highly peaked distribution of the number K_n of clusters in a sample of size n. Otherwise the clustering behaviour is fixed and cannot be controlled. Here we propose an alternative mixture model with a generalized gamma prior as a mixing measure. Such a random-probability measure $\tilde{P}$ is a species sampling model obtained by normalizing the jumps of the generalized gamma process that was introduced by Brix (1999). The distribution of $\tilde{P}$ and, hence, the distribution of K_n depend on two free parameters $\beta \in (0, \infty)$ and $\sigma \in (0, 1)$. The first plays the same role as the total mass in the MDP model, whereas σ influences the grouping of the observations in distinct clusters. Since σ directly affects the way that clusters are formed, it is apparent that having gained 1 degree of freedom is of great importance. It should be mentioned that the Dirichlet and the normalized inverse Gaussian processes are special cases of this wide family of priors. However, the value of σ in these two models is fixed and cannot be tuned. In the model that we shall illustrate we can either fix σ at a suitable value that is suggested by one's prior information or assign a prior distribution to it. Clearly, owing to the availability of the additional parameter σ, there is no need to place a prior on β in our treatment.

The formal set-up can be described as follows. Let $(Y_i)_{i \geqslant 1}$ be a sequence of observable random variables with values in $\mathbb{Y}$, whereas $(X_i)_{i \geqslant 1}$ is a sequence of latent random variables with values in $\mathbb{X}$. We assume a mixture model for the observations, namely

$$\left. \begin{array}{r} Y_i | X_i \overset{\text{ind}}{\sim} f(\cdot | X_i), \\ X_i | \tilde{P} \overset{\text{IID}}{\sim} \tilde{P}, \\ \tilde{P} \sim \mathscr{P}_{\beta, \sigma} \end{array} \right\} \tag{1}$$

where $f(\cdot | X_i)$ is a density function and $\mathscr{P}_{\beta, \sigma}$ stands for the distribution of a generalized gamma process with parameters β and σ. This is the same as assuming that, given $\tilde{P}$, the observations

Y_i are independent and identically distributed with density

$$f(y) \equiv \int_{\mathbb{X}} f(y|x)\, \tilde{P}(\mathrm{d}x).$$

If $\tilde{P}$ coincides with the Dirichlet process, we have the MDP model of Lo (1984). Indeed, an early use of the MDP model is also present in Berry and Christensen (1979), who provided an alternative empirical Bayes estimator for the probability of success in a Bernoulli model.

The outline of the paper is as follows. In Section 2 we describe a general framework for studying the clustering behaviour that is generated by random-probability measures inducing Gibbs-type random partitions. These are linked to the theory of (exchangeable) product partition models and an important result which characterizes the so-called cohesion functions is reformulated in this context. Having connected exchangeable product partition models and mixture models with Gibbs-type driving measure, in Section 3 we focus attention on a special subclass which stands out for tractability and is of particular appeal for modelling. We essentially propose to use mixture models which exploit at the higher stage of the hierarchy a non-parametric prior that is derived from a generalized gamma process. We first derive an explicit expression for the prior distribution of the number of components in the mixture. Moreover, the predictive distributions, a key ingredient for simulation algorithms, are given. The asymptotic behaviour of the distribution of the number of clusters K_n in a sample of size n is studied in detail: we have that K_n increases as n^{σ}, and thus is much quicker than the well-known logarithmic rate of the Dirichlet process. The asymptotic proportion of the clusters of a given size is derived as well. The limiting behaviour is greatly influenced by the parameter σ: a value of σ that is close to 1 yields a partition structure with a large number of clusters whose size tends to be small. The analysis of the distribution of the number of clusters is completed with some qualitative study. Indeed, it is seen that the two free parameters β and σ can be used to tune the location and the flatness respectively of the prior distribution of K_n. The parameter σ turns out to control an interesting reinforcement mechanism: the sampling procedure tends to reinforce significantly, among the old clusters, those having higher frequencies. This is a very appealing feature for inferential purposes and is illustrated by means of a simulated data example. The analytical and qualitative study of the distribution of the number of clusters highlights a trade-off which must be faced: a large value of σ favours strong reinforcement but at the same time prevents us from tuning the prior expected number of clusters on a small number, which is typically the case. The best way to circumvent this problem is to elicit a prior for σ and to let the data choose the appropriate reinforcement rate. Section 3 is completed with the derivation of the exact expression for the posterior distribution of the number of clusters. Unfortunately such an expression is not of immediate use and hence suitable simulation algorithms must be exploited. In Section 4 we consider a slightly more complicated mixture model in which, owing to its importance for inferential purposes, a prior distribution is assigned to the parameter σ. The modified Markov chain Monte Carlo (MCMC) algorithm is described in detail and the numerical issues that are related to the use of a $\mathrm{GG}(\beta, \sigma)$ mixture model are discussed. The model is applied to the same simulated data set as is considered in Section 3: the performance, with an additional hierarchy on σ, turns out to be significantly better. Finally, we consider data that are generated from a complex mixture of normal distributions with different weights and variances.

2.   A general model for clustering behaviour

An important issue that is addressed within mixture models concerns the clustering behaviour that is induced by the latent variables X_i at the higher level of the hierarchy in model (1). In what follows we suppose that the X_is take values in a complete and separable metric space $\mathbb{X}$ and

denote by $\mathscr{X}$ the corresponding Borel σ-field. The prior distribution which is employed on the space of probability distributions on $\mathbb{X}$ is the so-called *species sampling model* that is defined by

$$\tilde{P}(A) = \sum_{j=1}^{\infty} \tilde{p}_j \, \delta_{Z_j}(A) \qquad \forall A \in \mathscr{X}, \tag{2}$$

where δ_x is the point mass at x, the Z_js are independent and identically distributed (IID) with common non-atomic distribution P_0 and the random weights $\tilde{p}_j$ are independent from the Z_js. This class of random probability measures was introduced by Pitman (1996). It can easily be shown that, for this model, the prior guess at the shape of $\tilde{P}$ is P_0, i.e. $\mathbb{E}[\tilde{P}(A)] = P_0(A)$ for any A in $\mathscr{X}$. Since a species sampling model selects discrete distributions (almost surely), ties within the latent variables X_i occur with positive probability. If $k \in \{1, \ldots, n\}$ is the number of distinct values among the n variables $X_1, \ldots, X_n$, we denote such values by $X_1^*, \ldots, X_k^*$. Moreover, n_j is the number of X_is coinciding with X_j^*, thus implying that $\Sigma_{j=1}^k n_j = n$. We shall focus attention on species sampling models inducing a joint distribution of the number of ties K_n among n latent variables and the corresponding absolute frequencies of the product form

$$V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1} \qquad \sigma \in [-\infty, 1), \tag{3}$$

where the $V_{n,k}$s are a set of non-negative weights with $V_{1,1} = 1$, $n_i \geqslant 1$ for each $i = 1, \ldots, k$ and $\Sigma_{i=1}^k n_i = n$. Moreover, $(a)_n = a(a+1) \ldots (a+n-1)$ with the convention $(a)_0 = 1$. This means that the distribution of the random partition of the n latent variables $X_1, \ldots, X_n$ in model (1) factorizes in a term depending only on n and on the number of clusters k and a term which depends on the abundances n_js of the various clusters through σ. As in Pitman (2006) and Gnedin and Pitman (2005), we say that the random partition whose law is identified by expression (3) is of *Gibbs type*. Moreover, $\tilde{P}$ gives rise to the predictive distributions

$$\mathbb{P}(X_{n+1} \in A | X^{(n)}) = \frac{V_{n+1,k+1}}{V_{n,k}} \, P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \, \delta_{X_j^*}(A) \qquad \forall A \in \mathscr{X}, \tag{4}$$

where $X_1^*, \ldots, X_k^*$ are the k distinct observations in the sample $X^{(n)} = (X_1, \ldots, X_n)$. The predictive distribution results from a linear combination of P_0 and of a weighted empirical distribution which depends on the parameter σ. We shall show that the clustering behaviour of the X_is can dramatically change according to the value of σ.

It is interesting to note how we can recover the Dirichlet process as a special case. If $V_{n,k} = a^k/(a)_n$ and $\sigma = 0$, then expression (3) yields

$$\frac{a^k}{(a)_n} \prod_{i=1}^{k} (n_i - 1)!$$

which is the Ewens sampling formula induced by a Dirichlet process whose base-line measure has total mass $a > 0$. See Ewens (1972) and Antoniak (1974). This formula has found application in a variety of scientific areas ranging from Bayesian statistics to population genetics. See also Arratia *et al.* (2003).

There is also a close connection between Gibbs-type random partitions and the product partition models that were introduced by Hartigan (1990) and further studied, among others, by Barry and Hartigan (1993) and Quintana and Iglesias (2003). If Π_n represents a random partition of the set of integers $\{1, \ldots, n\}$, a product partition model corresponds to a probability distribution for Π_n represented as

$$\mathbb{P}[\Pi_n = \{S_1, \ldots, S_k\}] = M \prod_{i=1}^{k} c(S_i)$$

where $c(\cdot)$ is known as the *cohesion function* and

$$M = 1 \Big/ \sum_{\pi \in \mathcal{P}_n} \prod_{i=1}^{k} c(S_i),$$

$\mathcal{P}_n$ being the set of all partitions of the integers $\{1, \ldots, n\}$. The model that we are dealing with can be seen as an (exchangeable) product partition model. Indeed, if $|S| = \text{card}(S)$ and $I_j = \{i : X_i = X_j^*\}$ for each $j \in \{1, \ldots, k\}$, then we have

$$Y_i | (X_1^*, \ldots, X_k^*, \Pi_n) \overset{\text{ind}}{\sim} f(\cdot | X_j^*) \qquad i \in I_j,$$

$$X_i^* | \Pi_n \overset{\text{IID}}{\sim} P_0 \qquad i = 1, \ldots, k,$$

$$\Pi_n \sim \text{product partition distribution with } c(S) = (1-\sigma)_{|S|-1}$$

where in the above we have set k equal to the number of sets in the partition Π_n. Our choice of $c(\cdot)$ is quite general in the sense that, if we allow the cohesion function $c(S)$ to be just a function of the cardinality of S, we can reformulate an important result due to Gnedin and Pitman (2005) as follows.

Proposition 1 (Gnedin and Pitman, 2005). The exchangeable random partition Π_n has distribution of the form

$$V_{n,k} \prod_{j=1}^{k} c(n_j)$$

for any $n = 1, 2, \ldots$ and $1 \leqslant k \leqslant n$, if and only if

$$c(n_j) = (1-\sigma)_{n_j-1}$$

for some $\sigma \in [-\infty, 1]$ and also $V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1}$, with the proviso that $(1-\sigma)_{n_j-1} = 1$ when $\sigma = -\infty$ and that Π_n reduces to the singleton partition when $\sigma = 1$.

From this result it is apparent that if someone is to use a cohesion function depending on cardinalities, which seems a natural choice, then it must be of the form $c(S) = (1-\sigma)_{|S|-1}$. This provides a strong foundation for our treatment. It is worth noting that, unlike Quintana and Iglesias (2003), we do not confine ourselves to the Dirichlet process set-up: in our model the new parameter σ will have a deep influence on the clustering behaviour as will be clear from the next section.

In this setting, it is of great importance to know the distribution of the number K_n of distinct observations in the sample $X^{(n)}$ since it takes on the interpretation of a prior distribution on the number of components in the mixture model that is defined by expression (1). Within the class of Gibbs-type random partitions, this distribution is derived, e.g. in Gnedin and Pitman (2005), and is given by

$$P(K_n = k) = \frac{V_{n,k}}{\sigma^k} \mathcal{G}(n, k, \sigma), \tag{5}$$

where, for any $n \geqslant 1$ and $k = 1, \ldots, n$,

$$\mathcal{G}(n, k, \sigma) = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (-j\sigma)_n$$

are known as *generalized Stirling numbers* or *generalized factorial coefficients*. It is also the case that $\mathcal{G}(0,0,\sigma)=1$ and $\mathcal{G}(n,0,\sigma)=0$ for all $n\geqslant 1$. See Charalambides and Singh (1988) or Charalambides (2005) for a review on Stirling numbers.

To make concrete use of model (3) it is necessary to make the $V_{n,k}$s explicit by using the recursive equation that they need to satisfy. In general, this is a difficult task. Apart from the Dirichlet process prior, the other two cases in which the exact form of $V_{n,k}$ is known correspond to the two-parameter Poisson–Dirichlet process (Pitman, 1995) and the normalized inverse Gaussian process (Lijoi *et al.*, 2005). In the next section we shall consider another Gibbs exchangeable random partition for which σ lies in $(0,1)$ and the $V_{n,k}$s can be determined in closed form. It is worth noting that Gibbs partitions with negative σ give rise to finite exchangeable partitions which are mixtures of Poisson–Dirichlet distributions with appropriate parameters. See theorem 12(i) in Gnedin and Pitman (2005).

3. The generalized gamma process prior

The generalized gamma process has been introduced in Brix (1999) for constructing shot noise Cox processes. It is obtained from a Poisson random process on $\mathbb{R}^+$ with mean intensity given by

$$\nu(\mathrm{d}s)=\Gamma(1-\sigma)^{-1}\exp(-\tau s)s^{-(1+\sigma)}\,\mathrm{d}s \qquad s\in\mathbb{R}^+, \tag{6}$$

where $\sigma\in(0,1)$ and $\tau\geqslant 0$. In a Bayesian framework, such random measures have been exploited in, for example, Epifani *et al.* (2003), James (2002) and James *et al.* (2005). Here, we define a *generalized gamma prior* as species sampling model (2) with

$$\tilde{p}_i=J_i\bigg/\sum_{k=1}^{\infty}J_k \tag{7}$$

where the J_is are the points of a generalized gamma process. This means that, if $N(A)=\mathrm{card}(\{J_i: i=1,2,\ldots\}\cap A)$ and $A\in\mathscr{B}(\mathbb{R}^+)$, the latter being the Borel σ-field on $\mathbb{R}^+$, is such that $\nu(A)<\infty$, then $N(A)$ is a Poisson random variable with

$$\mathbb{E}[N(A)]=\nu(A) \qquad \forall A\in\mathbb{R}^+.$$

The class that we are considering contains some noteworthy priors as particular cases. For example, if $\tau=1$ and $\sigma\to 0$ we have the Dirichlet process. However, if $\tau=0$, then $\tilde{P}$ coincides with the normalized stable process that was first considered in Kingman (1975). Finally, if $\sigma=\frac{1}{2}$, we obtain the normalized inverse Gaussian process. It is known that these three particular cases induce Gibbs-type random partitions.

The following proposition establishes that, for any $\sigma\in(0,1)$, a generalized gamma prior induces a Gibbs-type partition and provides the distribution of the number of distinct components K_n. Note that, once the Gibbs structure has been proved, the determination of the distribution of K_n is achieved by deriving an explicit expression for $V_{n,k}$ to be inserted in formula (5).

Proposition 2. A generalized gamma prior induces a Gibbs-type partition and the corresponding distribution of the number of distinct components K_n is given by

$$\mathbb{P}(K_n=k)=\frac{\exp(\beta)\,\mathcal{G}(n,k,\sigma)}{\sigma\,\Gamma(n)}\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\beta^{i/\sigma}\,\Gamma\!\left(k-\frac{i}{\sigma};\beta\right), \tag{8}$$

where $\beta=\tau^\sigma/\sigma$, $k\in\{1,\ldots,n\}$ and

$$\Gamma(a,x)=\int_x^\infty s^{a-1}\exp(-s)\,ds$$

is the incomplete gamma function.

$V_{n,k}$ and the prior for K_n in equation (8) depend on two parameters $\beta\in\mathbb{R}^+$ and $\sigma\in(0,1)$. For this reason, we refer to the generalized gamma process prior that is defined via equation (7) as $\mathrm{GG}(\beta,\sigma)$. With this new parameterization, the normalized σ-stable process is obtained by setting $\beta=0$, whereas the normalized inverse Gaussian process arises when fixing $\sigma=\frac{1}{2}$.

As a by-product of proposition 2 we can easily determine from expression (4) the predictive distributions that are associated with a generalized gamma prior. Indeed, we have

$$\mathbb{P}(X_{n+1}\in B|X^{(n)})=w_0^{(n)}P_0(B)+w_1^{(n)}\sum_{j=1}^k(n_j-\sigma)\delta_{X_j^*}(B),\qquad(9)$$

with

$$w_0^{(n)}=\frac{\sigma}{n}\frac{\sum_{i=0}^n\binom{n}{i}(-1)^i\beta^{i/\sigma}\Gamma(k+1-i/\sigma;\beta)}{\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\beta^{i/\sigma}\Gamma(k-i/\sigma;\beta)},\qquad(10)$$

$$w_1^{(n)}=\frac{\sum_{i=0}^n\binom{n}{i}(-1)^i\beta^{i/\sigma}\Gamma(k-i/\sigma;\beta)}{n\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\beta^{i/\sigma}\Gamma(k-i/\sigma;\beta)}.\qquad(11)$$

Knowledge of the predictive distributions is fundamental for the application of appropriate simulation algorithms for sampling from a mixture of generalized gamma processes as will become clear in Section 4.

We now provide a deeper analysis of the behaviour of the distribution of K_n. This turns out to be important since it provides some hints for the prior specification of the parameters appearing in equation (5). In particular, the parameter σ plays a remarkable role in determining the clustering behaviour of the X_is. We first provide an asymptotic result which establishes the rate of growth of K_n as n increases. Such a statement can also be derived from proposition 13 in Pitman (2003) by suitably rewriting the $\mathrm{GG}(\beta,\sigma)$ process as a Poisson–Kingman model.

Proposition 3. Let K_n be the number of clusters that are induced by a $\mathrm{GG}(\beta,\sigma)$ prior. Then

$$K_n/n^\sigma\to S_\sigma\qquad(12)$$

almost surely. The random variable S_σ is strictly positive and its density is given by

$$g_{\beta,\sigma}(s)=\exp\left\{\beta-\left(\frac{\beta}{s}\right)^{1/\sigma}\right\}\frac{f_\sigma(s^{-1/\sigma})}{\sigma s^{1+1/\sigma}}\qquad(13)$$

where f_σ is the density function of a positive stable random variable with parameter σ.

In particular, for the normalized inverse Gaussian process ($\mathrm{GG}(\beta,\frac{1}{2})$), the asymptotic rate of growth for K_n is $\sqrt{n}$ and the density of the limiting random variable $S_{1/2}$ coincides with

$$g_{\beta,1/2}(s)=\frac{1}{\sqrt{\pi}}\exp\left(\beta-\frac{\beta^2}{s^2}-\frac{s^2}{4}\right).\qquad(14)$$

Result (12) yields important information for mixture modelling. Indeed, it establishes that the asymptotic growth of the number of distinct components in a sample of size n for a GG(β, σ) prior is of the type n^σ. This clearly suggests that σ is responsible for the number of distinct latent variables that are generated by the generalized gamma prior: the bigger σ, the larger the number of clusters that will be generated among the latent X_is.

Proposition 3 can be compared with analogous results for the Dirichlet and the two-parameter Poisson–Dirichlet process. In Korwar and Hollander (1973) it was shown that the number K_n of clusters that are induced by the Dirichlet process is such that

$$K_n / \log(n) \to a$$

where $a > 0$ is the total mass of the base-line measure α. As for the Poisson–Dirichlet process with parameters (a, σ), in Pitman (2006) it was shown that

$$K_n / n^\sigma \to Y_\sigma$$

where Y_σ has density given by

$$\frac{\Gamma(a+1) s^{a/\sigma} f_\sigma(s^{-1/\sigma})}{\Gamma(a/\sigma + 1) \sigma s^{1+1/\sigma}}.$$

Recall that the normalized stable process can be seen as a particular case of both the GG(β, σ) family and of the two-parameter Poisson–Dirichlet process. In the latter case, the normalized stable process is recovered by setting $a = 0$ and, obviously, Y_σ coincides in distribution with S_σ in proposition 3 with the choice $\beta = 0$.

Another important issue is associated with the clustering mechanism of a GG(β, σ) prior to be considered. It concerns the asymptotic behaviour of the number of clusters with a fixed size. If we let $K_{n,j}$ denote the number of clusters, among $X_1, \ldots, X_n$, of size j, a combination of proposition 3 with lemma 3.11 in Pitman (2006) leads to the following corollary.

Corollary 1. If K_n is the number of clusters that are generated by a GG(β, σ) prior, then

$$\frac{K_{n,j}}{K_n} \to p_{\sigma, j} = \frac{\sigma (1 - \sigma)_{j-1}}{j!}$$

almost surely, for any $j = 1, 2, \ldots$, as $n \to \infty$.

Hence, the asymptotic proportion of clusters of size j equals $p_{\sigma, j}$ and $\Sigma_{j=1}^\infty p_{\sigma, j} = 1$. Note that the asymptotic proportion of clusters of size 1 coincides with σ. As for the tail behaviour of $p_{\sigma, j}$, we have that, for j sufficiently large,

$$p_{\sigma, j} \sim \frac{\sigma}{\Gamma(1 - \sigma)} j^{-\sigma - 1},$$

thus suggesting a power law decay of index $\sigma + 1$.

The parameter σ again appears to be the most influential for the clustering structure of the latent variables. It should be remarked that a value of σ that is close to 1 yields a partition structure with a large number of clusters whose size tends to be small. The considerations that have been developed so far suggest some qualitative analysis of the distribution of K_n to provide some intuition on the prior specification of a GG(β, σ) model. This task is fulfilled in the next subsections.

3.1. The prior distribution of the number of clusters

Keeping in mind the results that were given in the previous section, it is worth investigating the qualitative behaviour of the distribution of K_n in equation (5) as the parameters β and σ vary. This seems of some importance since it should provide guidance on the prior specification of β and σ. In Fig. 1 we plot the graphs of the distribution of K_n for various values of n and of (β, σ). The probability points are connected by straight lines only for visual simplification.

It is evident that β can be used to control the location: the bigger β the larger the expected number of components tends to be. In contrast, σ allows the tuning of flatness of the distribution of K_n. Indeed, the bigger σ, the flatter is the distribution of K_n, suggesting that a large value of σ yields a non-informative prior for K_n. Such a finding is in accordance with proposition 3 and corollary 1, from which it can be deduced that, for a given n, a bigger σ tends to favour a large number of clusters and, among these, most of them have small size.

A reasonable strategy for the prior specification of (β, σ) would be to fix $\mathbb{E}_{\beta,\sigma}[K_n]$ equal to the prior opinion on the number of clusters. If we proceed in this way, we discover that there are some constraints on the possible choices. In particular, if we would like to tune $\mathbb{E}_{\beta,\sigma}[K_n]$ on a small value, large values of σ are not allowed whatever the choice of β. To make this argument more precise, one can numerically check that $\mathbb{E}_{0,\sigma}[K_n] \leqslant \mathbb{E}_{\beta,\sigma}[K_n]$ for any fixed σ and n. Table 1 displays the values of $\mathbb{E}_{0,\sigma}[K_n]$ for various sample sizes n and various choices of σ. These represent lower bounds for the possible values of $\mathbb{E}_{\beta,\sigma}[K_n]$.

3.2. The reinforcement mechanism induced by σ

Prior distributions inducing a Gibbs-type random partition exhibit a mechanism for allocating the mass that can be split into two stages, as suggested by the predictive distribution in equation (9). Given a sample $X_1, \ldots, X_n$, with k distinct values $X_1^*, \ldots, X_k^*$, the first step consists in allocating the mass between a newly observed value X_{k+1}^* sampled from P_0 and the set of observed values $\{X_1^*, \ldots, X_k^*\}$. This first step depends on n and k only. The second step consists in spreading the mass of $\{X_1^*, \ldots, X_k^*\}$ to each X_i^*. This allocation is determined by the size n_i of each cluster and by σ. At this stage, a reinforcement mechanism which is driven by σ takes place. Indeed, we can see that the ratio of the probabilities that are assigned to any pair of (X_i^*, X_j^*) is given by $(n_i - \sigma)/(n_j - \sigma)$. As $\sigma \to 0$, the previous quantity reduces to the ratio of the sizes of the two clusters, which characterizes the Dirichlet case. If $n_i > n_j$, the ratio is an increasing function of σ. Hence, as σ increases the mass is reallocated from X_j^* to X_i^*. This means that

Table 1. Lower bounds on $\mathbb{E}_{\beta,\sigma}[K_n]$ for various choices of σ and n

σ	*Results for the following values of n:*				
	n = 100	*n = 250*	*n = 500*	*n = 750*	*n = 1000*
0.125	1.89	2.12	2.31	2.43	2.52
0.25	3.49	4.39	5.22	5.77	6.20
0.375	6.32	8.92	11.57	13.47	15.00
0.5	11.27	17.83	25.23	30.90	35.68
0.625	19.81	35.15	54.22	69.86	83.63
0.75	34.38	68.38	115.03	155.92	193.47
0.875	58.95	131.46	241.13	343.84	442.263
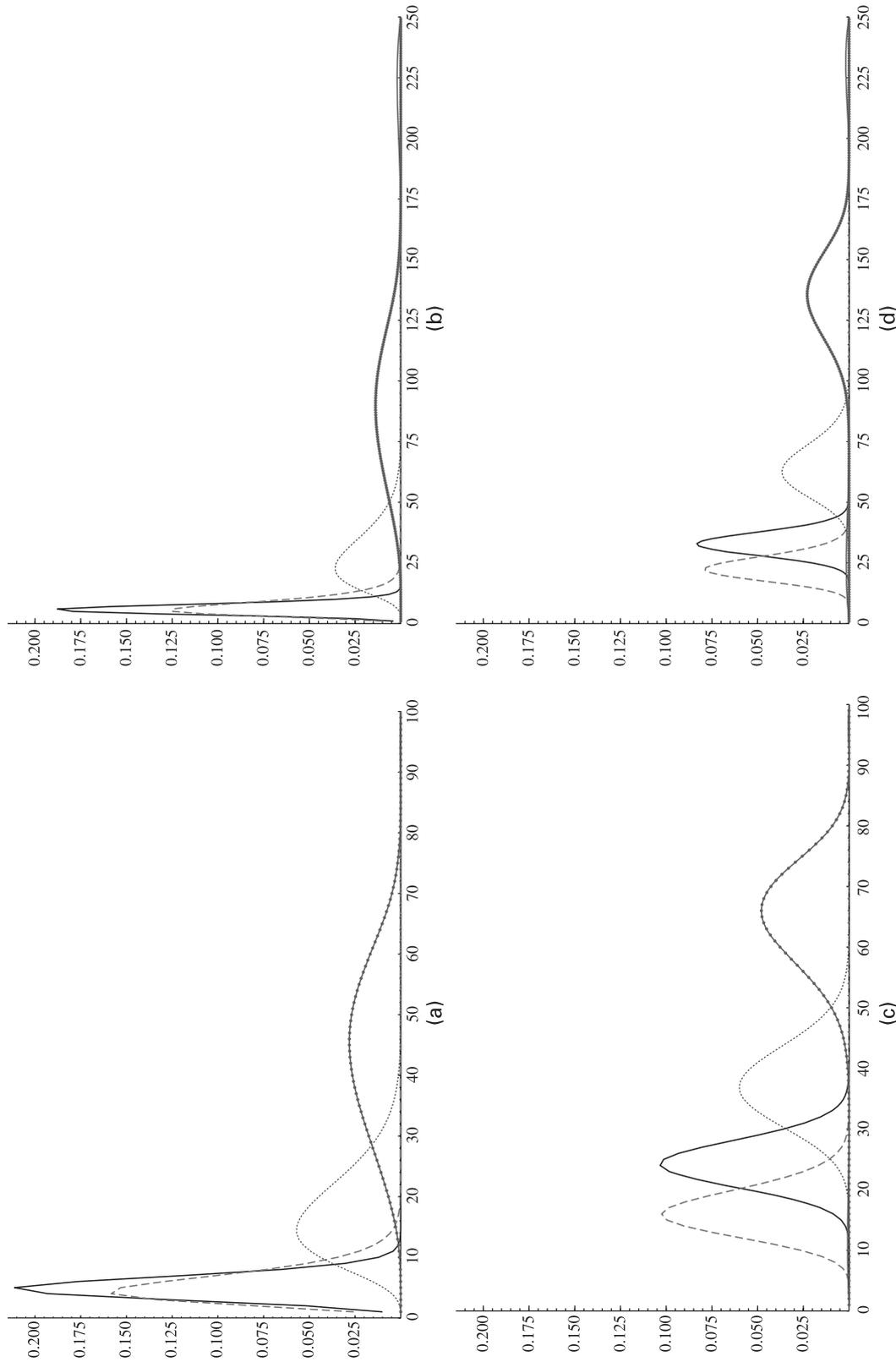
Fig. 1. Prior distribution of K_n for various values of β, σ and n: (a) $n = 100$ (———, Dirichlet ($a = 1$); · · · · · · , GG(1, 0.25); · · · · · · , GG(1, 0.5); ∘· · ·∘, GG(1, 0.75)); (b) $n = 250$ (———, Dirichlet ($a = 1$); · · · · · · , GG(1, 0.25); · · · · · · , GG(1, 0.5); ∘· · ·∘, GG(1, 0.75)); (c) $n = 100$ (———, Dirichlet ($a = 10$); · · · · · · , GG(10, 0.5); ∘· · ·∘, GG(10, 0.75)); (d) $n = 250$ (———, Dirichlet ($a = 10$); · · · · · · , GG(10, 0.5); ∘· · ·∘, GG(10, 0.75)); (e) $n = 100$ (———, GG(10, 0.25); · · · · · · , GG(1, 0.25); ———, GG(0.5, 0.25); · · · · · · , GG(10, 0.25); ∘· · ·∘, GG(50, 0.25); · · · · · · , GG(1, 0.25); (f) $n = 250$ (———, GG(10, 0.75); · · · · · · , GG(10, 0.25); ———, GG(0.5, 0.25); · · · · · · , GG(50, 0.25)); (g) $n = 100$ (———, GG(10, 0.75); · · · · · · , GG(10, 0.75); ∘· · ·∘, GG(50, 0.75)); (h) $n = 250$ (———, GG(1, 0.75); · · · · · · , GG(10, 0.75); ∘· · ·∘, GG(50, 0.75)))
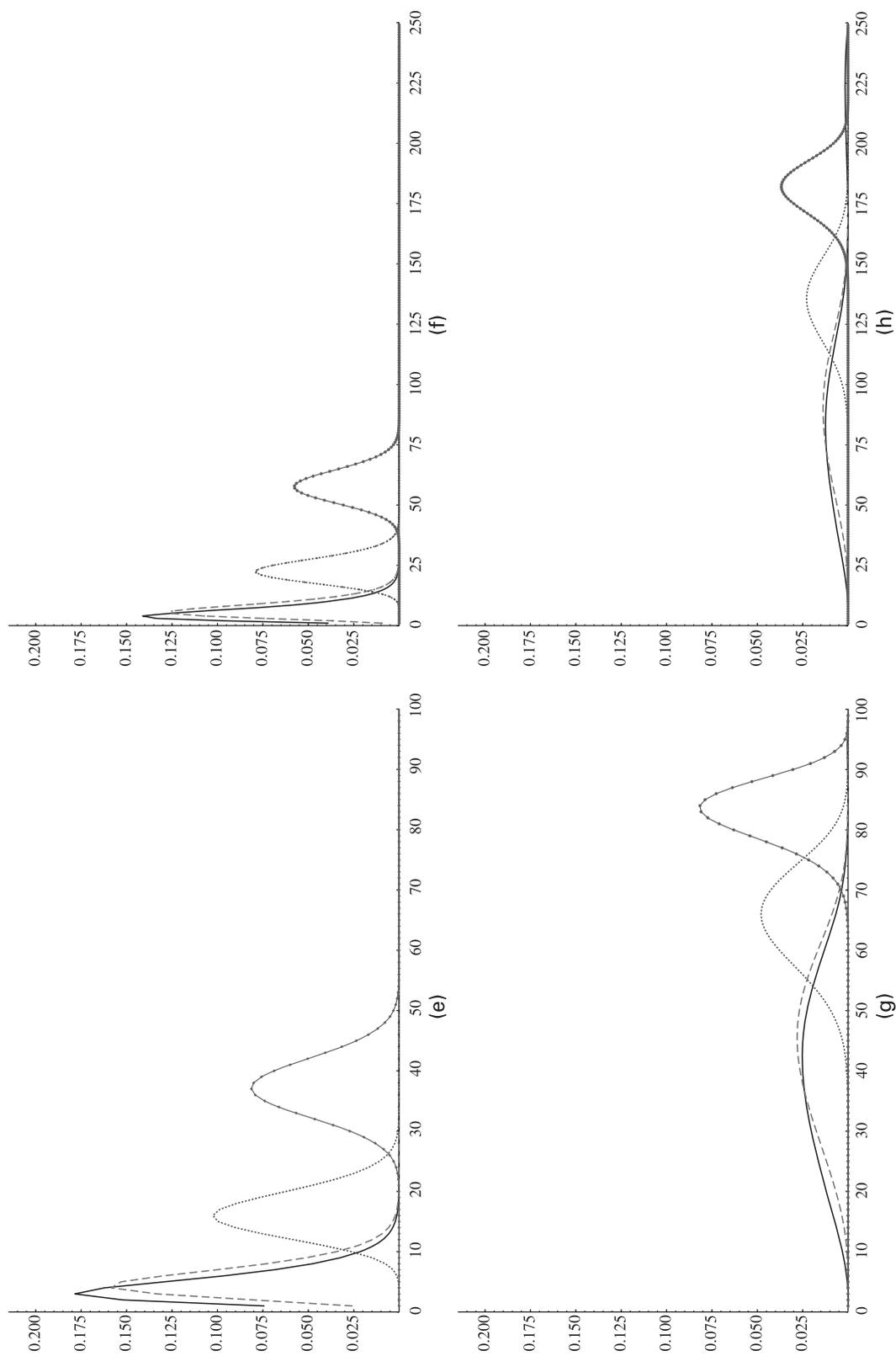
Fig. 1 (*continued*)

Table 2. Ratio of the probabilities allocated to X_i^* observed n_i times and X_j^* observed only once for various choices of σ

Model	Results for the following values of n_i:			
	$n_i = 2$	$n_i = 10$	$n_i = 50$	$n_i = 100$
Dirichlet	2	10	50	100
$GG(\beta, \sigma = 0.25)$	2.33	13	66.33	133
$GG(\beta, \sigma = 0.50)$	3	19	99	199
$GG(\beta, \sigma = 0.75)$	5	37	197	397
$GG(\beta, \sigma \to 1)$	$\to \infty$	$\to \infty$	$\to \infty$	$\to \infty$

the sampling procedure tends to reinforce, among the observed clusters, those having higher frequencies. Table 2 provides an idea of the magnitude of the reinforcement.

Combining these remarks with the asymptotic results that have been achieved so far, we can expect a $GG(\beta, \sigma)$ process prior to behave as follows. For a large value of σ, a large number of clusters will be generated. Among the clusters most will have small size (a small value of n_i). Some of them will have a significant size. The reinforcement mechanism acts on the mass allocation among the clusters by penalizing the clusters of small size and favouring those exhibiting some empirical evidence. We provide a qualitative illustration of this intuition by resorting to a simple example.

3.2.1. Example 1

Consider a uniform mixture of three normal distributions with means -4, 0 and 8, and unit variance. We simulate 100 values from such a mixture and use the data to compare the performance of three different mixture models: the MDP model, the mixture of normalized inverse Gaussian process and the mixture of $GG(\beta, 0.75)$ process. In each of these models, the mixing kernel is Gaussian with fixed variance 1. In the three cases that are under consideration, the main parameters have been chosen in such a way that the expected number of clusters, among the 100 sampled values, is equal to 50. Thus the prior opinion is very far from the truth. This choice is motivated by the fact that we wish to highlight the reinforcement mechanism acting with a $GG(\beta, \sigma)$ process prior. The corresponding parameter values turn out to be $a = 39.13205$ for the Dirichlet, $\beta = 24$ for the normalized inverse Gaussian ($\sigma = 0.5$) and $\beta = 2.23$ for the generalized gamma model with $\sigma = 0.75$. For all three processes P_0 is set equal to $N(\cdot | \bar{Y}, t^2)$, where $\bar{Y}$ is the sample mean and t is the data range. Simulations were carried out by using the algorithm that is detailed in Section 4. The results are based on 20 000 iterations with 2000 burn-in sweeps. In Table 3 we have displayed the posterior probabilities on the number of components. Fig. 2 shows the corresponding Bayesian density estimates. In particular, the $GG(2.23, 0.75)$ process provides, in this case, a better fit.

In this example, the choice of a large expected value of the number of clusters has been done on purpose. It firstly allowed us to point out the benefits of the reinforcement mechanism. Secondly, it also allowed us to circumvent the problem of fixing a large value of σ and achieving a low value of $\mathbb{E}_{\beta,\sigma}[K_n]$. In other terms, there is a trade-off in the choice of σ. The best way to solve the issue seems to be the specification of a prior for σ. Hence, the data will select the appropriate reinforcement rate. This will be carried out in Section 4.

Table 3. Posterior distributions on the number of components arising from the three mixture models centred such that the prior expected value of the number of components is 50

| k	*Results for the following models:*		
	Dirichlet (a=39.13)	*GG(β=24, σ=0.5)*	*GG(β=2.23, σ=0.75)*
3	0.00205	0.06660	0.42490
4	0.01295	0.19095	0.36055
5	0.04000	0.25175	0.15555
6	0.08210	0.22095	0.04575
7	0.13690	0.14305	0.01090
8	0.16560	0.07395	0.00195
9	0.16450	0.03530	0.00035
10	0.14395	0.01100	0.00005
11	0.10725	0.00455	
$\geqslant 12$	0.1447	0.00190	
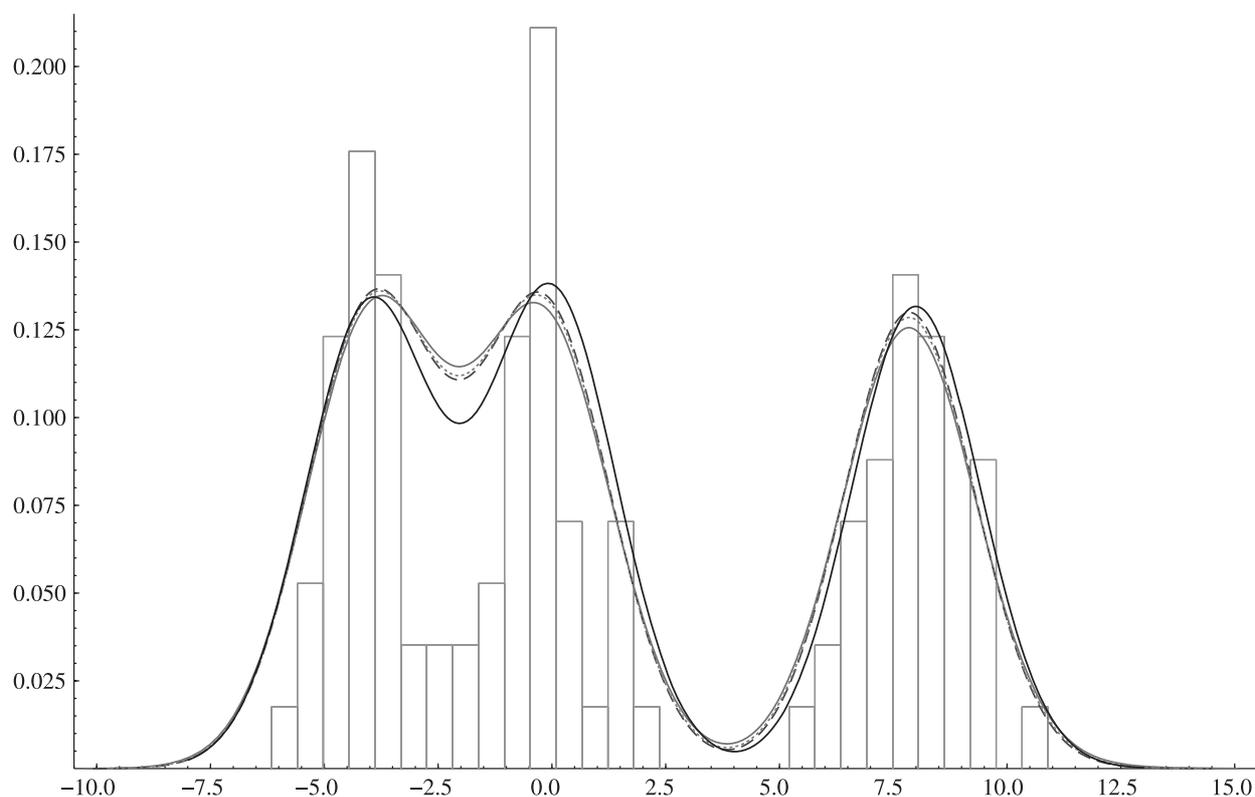
Fig. 2. Posterior density estimates arising from the MDP model, the mixture of normalized inverse Gaussian process and the mixture of generalized gamma process centred such that the expected value of the number of components is 50: ·······, Dirichlet ($a = 39.13205$); − − −, GG(2.23, 0.75); ··· ···, GG(24, 0.5); ———, true model

3.3. The posterior distribution of the number of clusters

Since the main focus of the paper is the behaviour of the number of components K_n within mixture models defined as in expression (1), the interest naturally lies on the determination of the posterior distribution of K_n given the observations $Y_1, \ldots, Y_n$. Such a task is difficult, even though a formal analytic representation can be provided. Within the Dirichlet process set-up,

such an issue has been considered first in Liu (1996) where a derivation, which is related to the binomial kernel, is presented in his theorem 2. In the setting of contingency tables, with Dirichlet prior on cell probabilities, the problem was tackled in Quintana (1998). For species sampling mixture models, Ishwaran and James (2003) formally derived a representation for the posterior distribution of K_n and, then, focused attention on the two-parameter Poisson–Dirichlet process whose weights are obtained via a stick breaking procedure. An explicit form for normalized random measures with independent increments priors was obtained in James *et al.* (2005). To obtain an expression for the posterior distribution of K_n for a $GG(\beta, \sigma)$ mixture model, we first introduce the symbol $\mathscr{L}_k$ to denote the joint distribution of the observations $Y^{(n)} = (Y_1, \ldots, Y_n)$ and of the number of distinct values K_n among the n latent variables $X_1, \ldots, X_n$. In other words,

$$\mathscr{L}_k(y^{(n)}) \, \mathrm{d}y^{(n)} = \mathbb{P}(Y_1 \in \mathrm{d}y_1, \ldots, Y_n \in \mathrm{d}y_n, K_n = k).$$

By resorting to proposition 2, which determines $V_{n,k}$, we have

$$
\mathscr{L}_k(y^{(n)}) = \frac{\sigma^{k-1} \exp(\beta)}{\Gamma(n)} \left\{ \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right) \right\}
$$
$$
\times \sum_{(*)_k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \int_{\mathbb{R}} \prod_{i \in I_j} f(y_i|x) \, P_0(\mathrm{d}x) \tag{15}
$$

where the above sum is extended over all partitions of the integers $\{1, \ldots, n\}$ into k groups and I_j is the set of indices in the jth cluster of the partition. Having equation (15) at our disposal, we can provide an expression for the posterior distribution of K_n, given $Y^{(n)} = y^{(n)}$,

$$\mathbb{P}(K_n = k | Y^{(n)} = y^{(n)}) = \mathscr{L}_k(y^{(n)}) \bigg/ \sum_{i=1}^{n} \mathscr{L}_i(y^{(n)}) \qquad k = 1, \ldots, n. \tag{16}$$

At this point, any further computation requires the specification of the kernel $f(\cdot|x)$ and of the base-line measure P_0. We first consider a simple and widely used choice of these two quantities, namely

$$f(y|x) = \frac{1}{\sqrt{(2\pi)}} \exp\left\{ -\frac{1}{2} (y-x)^2 \right\},$$

$$P_0(\mathrm{d}x) = \frac{1}{v\sqrt{(2\pi)}} \exp\left\{ -\frac{1}{2v^2} (x-m)^2 \right\} \mathrm{d}x.$$

Such a choice connects with example 1 that we have considered before and highlights difficulties that are usually associated with the exact evaluation of equation (16). In this case, the integral appearing in equation (15) can be easily seen to coincide with

$$\int_{\mathbb{R}} \prod_{i \in I_j} f(y_i|x) \, P_0(\mathrm{d}x) = \frac{(n_j v^2 + 1)^{-1/2}}{(2\pi)^{(n_j+1)/2}} \exp\left[-\frac{1}{2} \left\{ \frac{m}{v^2} + \sum_{(j)} y_i^2 + \frac{v^2}{n_j v^2 + 1} \left(\frac{m}{v^2} + n_j \bar{y}_{(j)} \right)^2 \right\} \right],$$

where $\bar{y}_{(j)} = \Sigma_{(j)} y_i / n_j$ is the sample mean within the jth cluster of a partition of $\{1, \ldots, n\}$. From this, we note that the main difficulty arises when trying to compute the sum over all partitions of order k, i.e.

$$\sum_{(*)_k} \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \int_{\mathbb{R}} \prod_{i \in I_j} f(y_i|x) \, P_0(\mathrm{d}x).$$

This consideration suggests that in practice formula (16) cannot be used, apart from cases in which n is very small. Hence, we shall resort to suitable computational schemes to determine a numerical approximation of the posterior distribution of K_n. This will be an issue that is considered in the next section.

4.   The mixture model with a prior on σ

On the basis of the arguments of the previous sections, we introduce a mixture model with a further hierarchy on σ. Hence model (1) becomes

$$
\left.
\begin{aligned}
Y_i|X_i &\overset{\text{ind}}{\sim} f(\cdot|X_i), \\
X_i|\tilde{P} &\overset{\text{IID}}{\sim} \tilde{P}, \\
\tilde{P}|\sigma &\sim \mathscr{P}_{\beta,\sigma}, \\
\sigma &\sim q.
\end{aligned}
\right\}
\tag{17}
$$

Various sampling techniques are widely used for generating inferences from model (1), given the knowledge of the predictive distribution. See, for example, Escobar and West (1995), Liu (1996), MacEachern (1994), MacEachern and Müller (1998) and Ishwaran and James (2001). Model (17) requires a slight modification of any of these algorithms to take into account a step in which the value of σ is drawn. This is easily accomplished once we can evaluate the full conditional for σ. Indeed, it can be seen, on the basis of proposition 2, that such a full conditional is given by

$$
q(\sigma|Y^{(n)}, X^{(n)}) = q(\sigma|K_n = k, n_1, \ldots, n_k)
$$
$$
\propto q(\sigma)\sigma^{k-1} \left\{ \prod_{j=1}^{k} (1-\sigma)_{n_j-1} \right\} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right). \tag{18}
$$

It is clear from this expression that the full conditional for σ depends only on the clustering structure of the n latent variables X_i. None-the-less, the clustering is affected by the observed Y_is which, then, indirectly influence the sampled value of σ. Before focusing on the illustrative examples, let us briefly describe the sampling scheme that we resort to. First fix initial values $\sigma = \sigma_0$ and $X_i = X_{i,0}$, for $i = 1, \ldots, n$. These can be drawn from q and from P_0 respectively. At step $t \geqslant 1$ we can proceed in a similar fashion to that in Ishwaran and James (2001), i.e.

(a) draw σ_t from $q(\sigma|k_{t-1}, n_1, \ldots, n_{k_{t-1}})$, where k_{t-1} is the number of ties in $X_{t-1}^{(n)}$, and
(b) draw the latent variables $X_{1,t}, \ldots, X_{n,t}$ from the Pólya urn scheme as follows: for any i sample X_i from

$$
P(X_{i,t} \in \cdot | X_{-i,t}^{(n)}, Y^{(n)}, \sigma_t) = q_{i,0}^*(\sigma_t) P_0(\mathrm{d}X_{i,t}) f(Y_i|X_{i,t}) + \sum_{j=1}^{k_{i,t}} q_{i,j}^*(\sigma_t) \delta_{Z_j^*}(\cdot), \tag{19}
$$

where $X_{-i,t}^{(n)} = (X_{1,t}, \ldots, X_{i-1,t}, X_{i+1,t-1}, \ldots, X_{n,t-1})$ and Z_j^* are the $k_{i,t}$ distinct values in the vector $X_{-i,t}^{(n)}$. The mixing proportions are given by

$$
q_{i,0}^*(\sigma_t) \propto w_{i,0}^{(n)}(\sigma_t) \int_{\mathbb{X}} f(Y_i|x) P_0(\mathrm{d}x),
$$
$$
q_{i,j}^*(\sigma_t) \propto (n_j - \sigma_t) w_{i,1}^{(n)}(\sigma_t) f(Y_i|Z_j^*),
$$

subject to the constraint $\Sigma_{j=0}^{k_{i,t}} q_{i,j}^*(\sigma_t) = 1$.

The weights $w_{i,r}^{(n)}$, with $r \in \{0, 1\}$, are the same as those given in equations (10) and (11). Moreover, it is useful to implement an acceleration step which is aimed at a faster mixing. Such a variation of the MCMC algorithm for MDP models has been proposed by MacEachern (1994) and Bush and MacEachern (1996). See also MacEachern (1998) and Ishwaran and James (2001). The idea is to add a further step after step (b) in the previous description of the algorithm. Indeed, step (b) is used to fix the number of clusters and the cluster memberships. To generate the representative of each cluster, i.e. the unique distinct values Z_j^*, we proceed as follows. Suppose that from step (b) we have k_t clusters with memberships identified by the sets of indices $I_{1,t}, \ldots, I_{k_t,t}$. Then

   (c) draw the unique values $Z_{1,t}^*, Z_{2,t}^*, \ldots, Z_{k_t,t}^*$ from the full conditional

$$P(Z_{j,t}^* \in \mathrm{d}x \mid Y^{(n)}, X^{(n)}) \propto \prod_{i \in I_{j,t}} f(y_i|x)\, P_0(\mathrm{d}x).$$

We can see that an important point of the algorithm is the evaluation of the weights $q_{i,0}^*$. To obtain an explicit form for them, we can choose a conjugate pair $\{f(\cdot|\cdot), P_0\}$. This computational scheme has been used in all the examples in the paper. In particular, in example 1 step (a) is skipped since a prior for σ has not been specified. In all the other examples the following specifications are made:

   (a) $f(\cdot|\cdot)$ coincides with a normal density and P_0 is conjugate and
   (b) as a prior for σ we choose a discrete distribution q, which is useful for reducing the computational burden of the algorithm.

The final subsection provides some insight into the numerical issues that are raised by the algorithm and how they have been tackled.

4.1. Simulated data for a simple mixture

We first consider the simple mixture that we have already dealt with in example 1 and we, now, specify a prior for σ. This mixture, even though simple, turns out to be quite useful for identifying some specific features of putting a hierarchy on σ. We set q as the discrete uniform distribution over the points $j/100$ for $j = 1, \ldots, 99$. To draw a comparison with the previously examined cases (fixed σ), we select β such that

$$\mathbb{E}_\beta[K_{100}] = \frac{1}{99} \sum_{j=1}^{99} \mathbb{E}_{\beta,\,j/100}[K_{100}] = 50$$

which corresponds to $\beta = 22.8$. It is worth noting that the prior that is induced on the number of components K_n is essentially uniform with prior probabilities on the ks of about 0.01. To be more precise, the distribution is bimodal with modes in $k = 19$ and $k = 100$ with probabilities of 0.01076 and 0.01163 respectively. The MCMC sampling scheme that we have adopted provides an output that can be used to provide posterior summaries for K_n and for the density of the data and we run it for 20000 iterations with 2000 burn-in sweeps. The posterior probabilities for K_n are displayed in Table 4.

Table 4 suggests that the performance, in terms of the ability to detect the correct number of clusters in the data, of a generalized gamma mixture model with random σ is clearly better than the model that is achieved with a fixed σ. For instance, note that the best model with fixed σ that we have considered, i.e. $\sigma = 0.75$, yields a posterior probability of 0.4249 on the correct number $k = 3$ of components. This is remarkably lower than the value of 0.8811 that we obtain by putting a uniform prior on σ. The density estimate does provide a fit that is indistinguishable from the fit that is featured by the $\mathrm{GG}(\beta, \sigma)$ model with $\sigma = 0.75$.

Table 4. Prior and posterior probabilities for the number of components corresponding to the $GG(\beta = 22.8, \sigma)$ process with σ uniformly distributed over $j/100$ for $j = 1, \ldots, 99$, $n = 100$

		Results for the following values of k:					
		$k \leqslant 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k \geqslant 7$
$GG(\beta = 22.8, \sigma)$	Prior	0.0121	0.0092	0.0096	0.0098	0.01	0.9495
$\Pr(\sigma = j/100) = 1/99$	Posterior		0.8811	0.1105	0.0078	0.00055	0.00005

These preliminary findings allow us to draw some comments about the role of σ in this more general mixture model. First, note that, when the configuration of the latent variables consists of a relatively large number of groups, then it is more likely to sample a relatively large value of σ. This can be seen from Fig. 3 where we can find graphs of the full conditional of σ that is given in expression (18) for different configurations which highlight the influence of both the number of clusters K_n and the balancedness or unbalancedness of the corresponding composition. In particular

(a) Fig. 3(a) displays the full conditionals for $K_{100} = 3$ with compositions $(n_1, n_2, n_3) = (33, 34, 33)$ and $(n_1, n_2, n_3) = (98, 1, 1)$,
(b) Fig. 3(b) displays pictures of the full conditional for $K_{100} = 20$ with compositions $(n_1, \ldots, n_{20}) = (5, \ldots, 5)$ and $(n_1, \ldots, n_{20}) = (81, 1, \ldots, 1)$ and
(c) Fig. 3(c) concerns the case of $K_{100} = 50$ with compositions $(n_1, \ldots, n_{50}) = (2, \ldots, 2)$ and $(n_1, \ldots, n_{50}) = (51, 1, \ldots, 1)$.

Having sampled a large value of σ, this generates a large number K_n of clusters in the Pólya urn scheme and concentrates the observations in those few groups having larger sizes. Hence, there is a remarkable fraction of groups having small sizes. The first behaviour, namely having a large K_n, can be motivated in the light of proposition 3: indeed large values of σ lead to a large value of K_n. The second effect, i.e. the concentration structure of the observations, can be explained by corollary 1. Indeed, according to such a result, the larger σ the higher is the proportion of classes having small sizes. However, we also need to consider the action of the kernel f, which in this case is the normal distribution. Indeed, the effect of a large σ generating a large K_n is compensated by such a kernel. From a qualitative point of view, we can say that the smoothing effect of the normal kernel tends to 'merge' the classes with small sizes to classes having large sizes. If as a final outcome we have a reduction in K_n, then the next sampled value of σ is likely to be smaller.

Let us now focus on the specific example of this subsection. From the MCMC output we can extract information about σ. Indeed the average σ, which we interpret as a posterior estimate of σ, turns out to be 0.0267. A comparison with the results that were obtained in example 1 can be of interest. In that example, it turned out that the value $\sigma = 0.75$ was much better than the value corresponding to the Dirichlet process, i.e. $\sigma \to 0$. This might seem to be in contrast with what we have obtained with random σ. Indeed, it is not. This is because in this example the algorithm detects very quickly the correct number of components and these are well separated. Hence, the effect of the normal kernel is dominating and there is no need for reinforcement once the correct number of components has been identified. Such a small value of K_n leads to sampling small values of σ, thus explaining the outcome that we obtained.
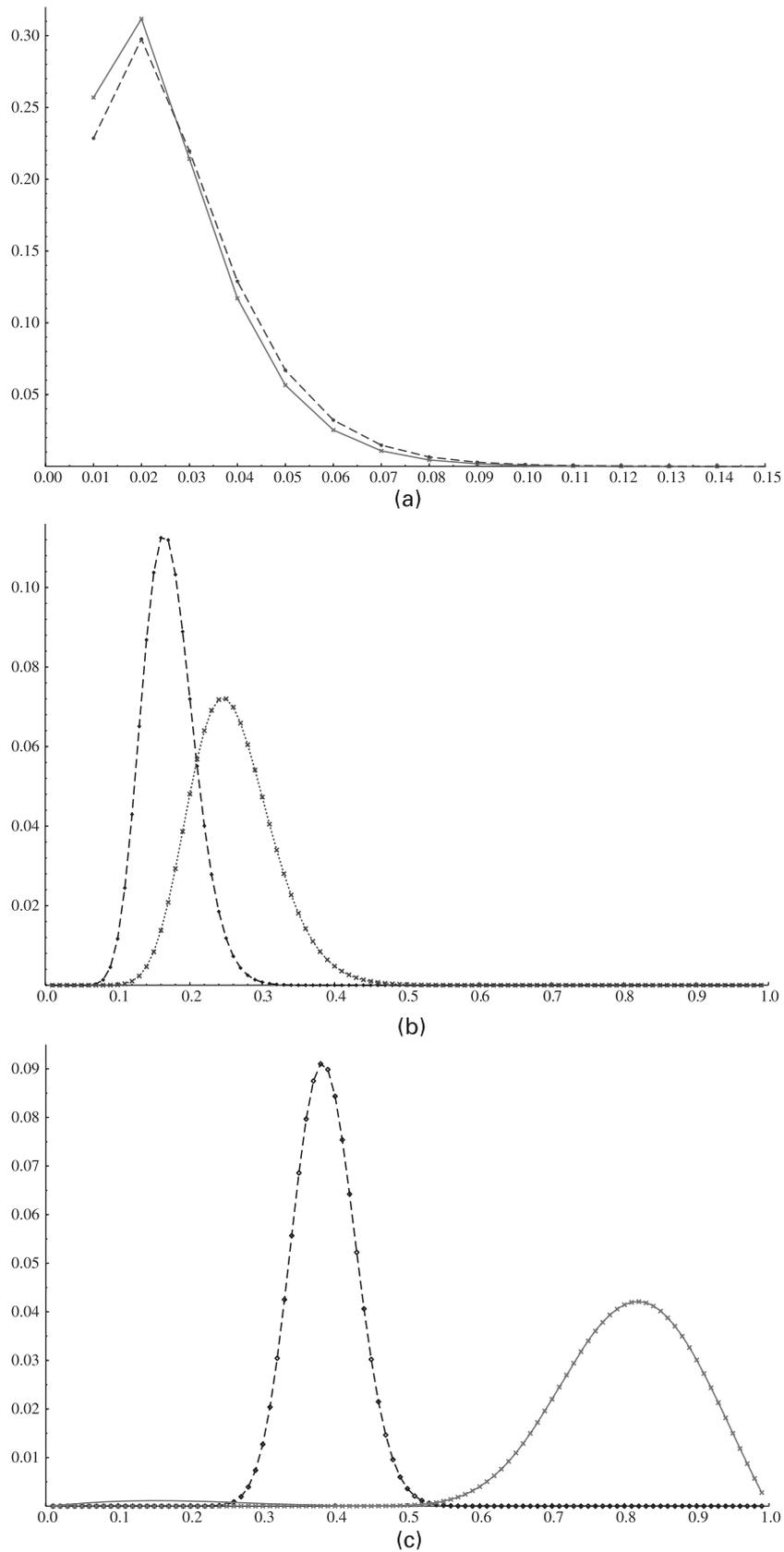
Fig. 3. Conditional distributions of σ for various configurations of K_n and $n_1, \ldots, n_{K_n}$: (a) $k = 2$ (×, $(n_1, n_2, n_3) = (33, 34, 33)$; •, $(n_1, n_2, n_3) = (98, 1, 1)$); (b) $k = 20$ (•, $(n_1, \ldots, n_{20}) = (5, \ldots, 5)$; ×, $(n_1, \ldots, n_{20}) = (81, 1, 1)$); (c) $k = 50$ (•, $(n_1, \ldots, n_{50}) = (2, \ldots, 2)$; ×, $(n_1, \ldots, n_{50}) = (51, 1, \ldots, 1)$)

4.2. A mixture with non-uniform weights and unequal variances

We now apply our mixture model to a more complicated setting to test its performance further. We shall consider a mixture of normal distributions with unequal variances and mixing weights of the kind

$$\sum_{i=1}^{7} \omega_i \, N(\cdot | \theta_i, \lambda_i^2)$$

where $\omega_1 = \omega_2 = 0.1$, $\omega_3 = 0.15$, $\omega_4 = \omega_5 = 0.125$, $\omega_6 = \omega_7 = 0.2$, $(\theta_1, \ldots, \theta_7) = (-8, -5, 0, 5, 7, 11, 15)$ and $(\lambda_1^2, \ldots, \lambda_7^2) = (0.8, 0.5, 2, 0.25, 0.05, 1, 1.25)$. Our hierarchical mixture model (17) can be translated in a semiparametric form as

$$(Y_i \, | \, m_i, V_i) \overset{\text{ind}}{\sim} N(Y_i \, | \, m_i, V_i^{-1}),$$

$$(m_i, V_i \, | \, \tilde{P}) \overset{\text{IID}}{\sim} \tilde{P},$$

$$\tilde{P} | \sigma \sim \mathscr{P}_{(\beta, \sigma)},$$

$$\sigma \sim q$$

for $i = 1, \ldots, n$. The $GG(\beta, \sigma)$ process is centred at P_0 of the form

$$P_0(dm, dv) = N(m | \bar{Y}, t^2/v) \, Ga(v | 1, \tfrac{1}{2}) \, dm \, dv, \tag{20}$$

where $Ga(\cdot | c, d)$ is the density corresponding to a gamma distribution with mean c/d and, as before, t stands for the range of the data. The last specification that we need to make concerns σ and β. Unlike example 1, we now wish to consider the case in which the prior expected number of components is lower than the correct number. Hence, we aim at fixing the expected value for K_{200} equal to 3. As has already been observed in Table 1, having some mass on values of σ that are sufficiently close to 1 prevents us from obtaining a small expected value for K_{200}. For this reason, the prior q for σ has been chosen as a discretized beta distribution with parameters (1,9). In particular, we consider discretizations yielding supports for q either with nine points corresponding to $\{0.1, 0.2, \ldots, 0.9\}$ or 99 points corresponding to $\{0.01, 0.02, \ldots, 0.99\}$. These two choices are made to check the sensitivity of the posterior inferences on the degree of coarseness of the support of the prior. At this point, the targeted $\mathbb{E}_\beta[K_{200}] = 3$ is attained for $\beta = 0.485$ (for the nine-points support of q) and $\beta = 0.841$ (for the 99-points support of q).

To evaluate the performance of the $GG(\beta, \sigma)$ mixture, we generate a sample $Y_1, \ldots, Y_{200}$ from the seven normal distributions that were described before with proportions coinciding with the correct weights ω_i, for $i = 1, \ldots, 7$. In Table 5 we summarize the results about posterior probabilities on the number of components. Fig. 4 depicts the plot of the posterior density estimates.

As for the behaviour of σ, we have a posterior estimate equal to 0.16871 and 0.14440 in the nine-points support and 99-points support case respectively.

The results show a good performance of the mixture model that is driven by a $GG(\beta, \sigma)$ process with random σ. The model can detect the correct number of components for a data set that was generated by a complicated mixture. Also the posterior density estimates feature a satisfactory fit to the density that has generated the data. This argument is further strengthened by the fact that we are using a sample of relatively small size if compared with the complex structure of the true mixture.

As for the sensitivity of posterior inferences to the degree of coarseness of the support of the prior q for σ, we can note that there is no substantial difference between the two cases. This is true when considering the posterior distribution of K_{200}, the density estimate of the mixture and the Bayes estimate of σ. Hence, we find that the model has a satisfactory degree of robustness with respect to the prior for σ.

Table 5. Posterior distributions on the number of components arising from the GG(β, σ) process mixtures with nine and 99 points in the support of q

k	*Results for the following models:*	
	GG ($\beta=0.485, \sigma$), $\sigma \sim q$, supp(q) = {0.1, 0.2, ..., 0.9}	*GG ($\beta=0.841, \sigma$), $\sigma \sim q$, supp(q) = {0.01, 0.02, ..., 0.99}*
6	0.25705	0.30230
7	0.33760	0.32785
8	0.19820	0.19695
9	0.09835	0.09660
10	0.05590	0.04235
11	0.02725	0.01820
12	0.01435	0.00900
$\geqslant$13	0.01130	0.00675
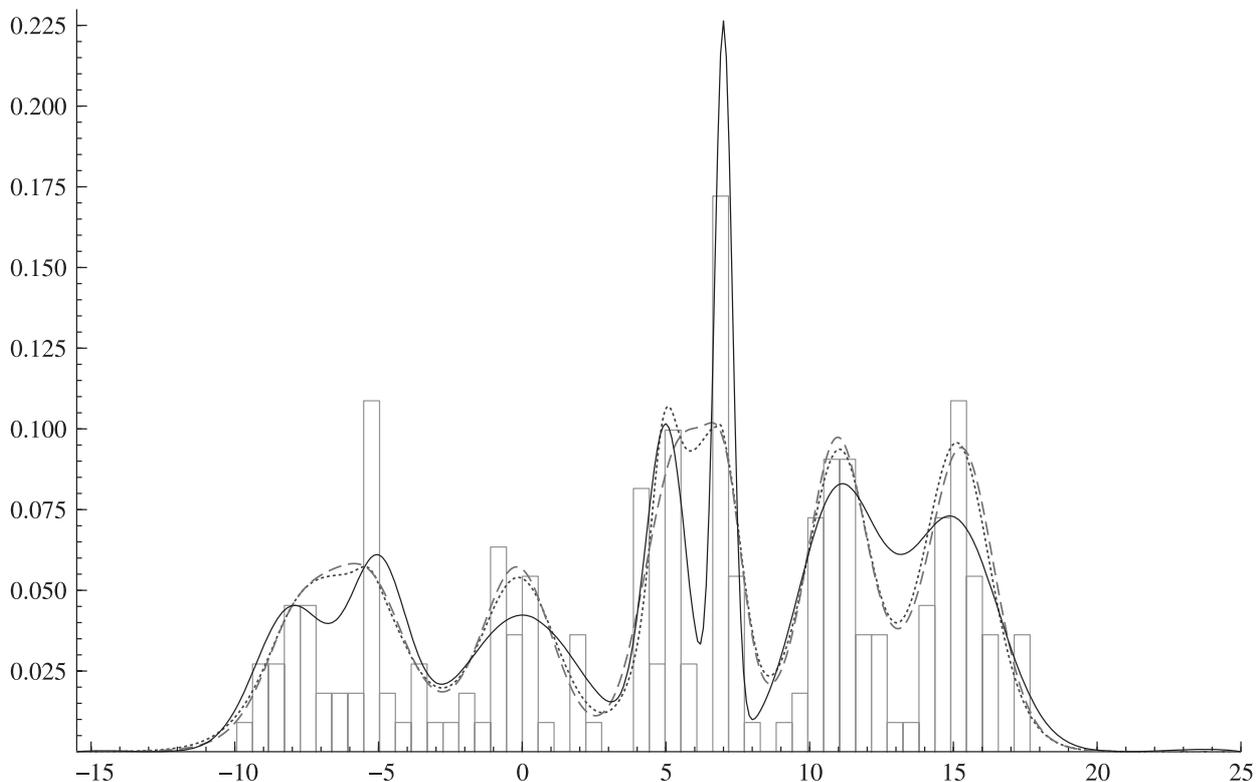
Fig. 4. Posterior density estimates arising from GG(β, σ) process mixtures with $\mathbb{E}[K_{200}] = 3$ and $\sigma \sim q$, where q is a discretized beta distribution with support {0.1, 0.2, ..., 0.9} and {0.01, 0.02, ..., 0.99}; ······, GG($\beta = 0.485$); − − −, GG($\beta = 0.841$); ———, true model

A concluding comment concerns the possibility of adding further hierarchies to the mixture model that we have used. For example, we could have introduced a further hierarchy for the parameters in P_0 in equation (20). This strategy has been pursued in other references concerning the MDP model such as Escobar and West (1995). From a computational point of view, the addition of this step is straightforward and does not complicate the algorithm, and also for GG(β, σ) process mixtures. We have, indeed, experimented with models with a larger number of hierarchies: however, we have found a strong sensitivity of inferential results to the

specification of the hyperparameters. In particular, for some choices of these parameters we can achieve the same performance, for the detection of the correct number of components, and a slight improvement for the density fit. In contrast, other choices of the hyperparameters yield a drastic worsening in the posterior inference on the number of components. These findings suggest that the addition of a further hierarchy is reasonable when the prior specification for the hyperparameters is advised by particular features of the real problem being examined. See, for example, Escobar and West (1995) and the motivation that they provided for the choice of the hyperparameters, for the MDP model, in the example with the galaxy data. In other situations, a good strategy would be a default specification as the one that we have adopted in this example which leads to robust inferences about K_n.

4.3. Numerical issues

Just as in the case of MDP models, inferences that are based on mixtures of $GG(\beta, \sigma)$ models strongly rely on the ability to draw samples from such a random measure. Although many of the methods that are available for MDP models are also potentially feasible for mixtures of $GG(\beta, \sigma)$ models, some computational issues need to be discussed. In the examples that are contained in this work, we have used the generalized Pólya urn Gibbs sampler which requires the computation of the weights $w_0^{(n)}$ and $w_1^{(n)}(n_j - \sigma)$, $j = 1, \ldots, k$, which are given in equations (10) and (11). These quantities clearly take a more complex form than those corresponding to the Dirichlet process, since they require the computation of sums of incomplete gamma functions. However, it is not only the computation of these functions that makes such quantities difficult to handle in standard packages but also the magnitude of the resulting summands.

To have a better picture of this issue, let us consider the following set of parameters: $n = 200$, $K_{200} = 10$, $\beta = 20.5$ and $\sigma = 4/3$. For these values, the summands that are contained in $w_0^{(n)}$ eventually need the evaluation of $\Gamma(-256.667, 20.5) = 9.336349705 \times 10^{-349}$. With the same parameters, but with $K_{200} = 150$, the generalized Pólya urn Gibbs sampler eventually falls into the evaluation of $\Gamma(151, 20.5) = 5.713383956 \times 10^{262}$. Large values for the number of ties K_n are typically observed for the first iterations when the algorithm has not been able to disentangle the right number of components and still preserves many of the samples from P_0 or the data. Even though, for most purposes, such quantities could be considered as 0 and ∞ respectively, doing this in the computation of the weights would lead to a significative bias, since they must add up to 1.

One might attempt to rescale the big summands that are required to compute $w_0^{(n)}$ and $w_1^{(n)}$. However, such an approach would depend on a given set of parameters (β, σ and n) and would therefore be difficult to implement in general. Truncating small and big quantities might lead to bad approximations. However, the availability of arbitrary precision packages such as Mathematica, Mupad, Maple and PARI among others aids in the precise computation of small and big numbers. Given the nature of the algorithm, which is required to implement the generalized Pólya urn Gibbs sampler, we decided to use the PARI C library, which is freely available at `http://pari.math.u-bordeaux.fr/`, to compute, not only the weights, but also the generalized Stirling numbers that are required in expressions (5) and (8). To facilitate the implementation of the Gibbs sampler we programmed in OX (Doornik, 2002) and perform a call to the PARI C library when necessary. Other packages such as R could also be used as a front end environment. Alternatively one could develop the whole procedure in the C language. The programs that were used for this work are available from the authors.

5. Concluding remarks

When dealing with inferential problems of clustering or of density estimation in a Bayesian non-parametric framework two alternative approaches can be followed to achieve the necessary flexibility for fitting the data:

(a) employ the MDP model with a suitable number of hierarchies on the parameters or
(b) look for extensions of the MDP model by replacing the Dirichlet process with a more general prior.

The latter aims at a more parsimonious specification with a reduced number of hierarchies. Within this second approach, in this paper we have considered an important extension of the MDP model. The main advantage is the availability of an additional parameter with a precise meaning in the context of mixture modelling. Indeed, σ greatly affects the clustering behaviour of the latent variables.

Finally, it is worth remarking that, apart from the MCMC sampling scheme which we have adopted, one can resort to alternative algorithms such as sequential importance sampling as set out in Liu (1996). Future work will focus on the implementation of the sequential importance sampling algorithm for $GG(\beta, \sigma)$ mixture models and on the comparison, in terms of efficiency, with MCMC sampling.

Acknowledgements

Appendix A

A.1. Proof of proposition 2
Note that the species sampling model (7) can also be represented as

$$\tilde{P}(A) = \frac{\displaystyle\int_{\mathbb{R}^+ \times A} s\, N\,(\mathrm{d}s, \mathrm{d}x)}{\displaystyle\int_{\mathbb{R}^+ \times \mathbb{X}} s\, N\,(\mathrm{d}s, \mathrm{d}x)} = \frac{\xi(A)}{\xi(\mathbb{X})},$$

where ξ is a random measure with independent increments such that

$$\mathbb{E}[\exp\{-\lambda\, \xi(A)\}] = \exp\left[-a\, P_0(A) \int_{\mathbb{R}^+} \{1 - \exp(-\lambda s)\}\, \nu(\mathrm{d}s)\right], \tag{21}$$

where ν is given in equation (6). The prior $\tilde{P}$ is, thus, a normalized random measure with independent increments, a general class of priors that was introduced in Regazzini *et al.* (2003) and extended to more general spaces $\mathbb{X}$ in James (2002). It also belongs to the family of Poisson–Kingman models that was introduced by Pitman (2003). The Lévy–Khintchine representation in equation (21) plays an important role in our proof. Indeed, we note that the joint distribution of the number of components K_n and of the absolute frequencies that we wish to determine coincides with the so-called *exchangeable partition probability function*; see Pitman (1995). If we let $\mu_n(u) := \int_0^\infty s^n \exp(-us)\, \nu(\mathrm{d}s)$, $n = 1, 2, \ldots$, for any positive u, and $\Pi_k^{(n)}(n_1, \ldots, n_k)$ denote the exchangeable partition probability function that is associated with a normalized random measure with independent increments having intensity $a\, P_0(\cdot)\, \nu(\cdot)$, then

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{a^k}{\Gamma(n)} \int_0^\infty u^{n-1} \mathbb{E}[\exp\{-u\,\xi(\mathbb{X})\}]\,\mu_{n_1}(u)\ldots\mu_{n_k}(u)\,\mathrm{d}u.$$

See corollary 6 in Pitman (2003) and, also, the proof of proposition 3.6 in Prünster (2002). In the case of a GG(β,σ) model we shall show that $\Pi_k^{(n)}$ can be decomposed in a product form, thus yielding a Gibbs-type exchangeable random partition for which $V_{n,k}$ can be exactly determined. If ν is as in equation (6), then

$$\mu_{n_j}(u) = \frac{1}{\Gamma(1-\sigma)} \int_0^\infty s^{n_j-\sigma-1} \exp\{-(\tau+u)s\}\,\mathrm{d}s = \frac{(1-\sigma)_{n_j-1}}{(\tau+u)^{n_j-\sigma}}$$

and, from equation (21),

$$\mathbb{E}[\exp\{-u\,\xi(\mathbb{X})\}] = \exp\left\{-a\frac{(\tau+u)^\sigma - \tau^\sigma}{\sigma}\right\}.$$

Hence, denoting the exchangeable partition probability function of a GG(β,σ) prior by $\Pi_{\beta,\sigma}$ to emphasize its dependence on the parameters (β,σ) yields

$$\Pi_{\beta,\sigma} = \frac{a^k \prod_{j=1}^k (1-\sigma)_{n_j-1}}{\Gamma(n)} \int_0^\infty u^{n-1} \exp\left[-\frac{a}{\sigma}\{(\tau+u)^\sigma - \tau^\sigma\}\right](\tau+u)^{-n+k\sigma}\,\mathrm{d}u.$$

If we set $\beta = a\tau^\sigma/\sigma$ and use an appropriate change of variable, we obtain

$$\Pi_{\beta,\sigma} = \frac{a^k \exp(\beta) \prod_{j=1}^k (1-\sigma)_{n_j-1}}{\sigma\,\Gamma(n)} \int_{\tau^\sigma}^\infty (y^{1/\sigma}-\tau)^{n-1} y^{k-n/\sigma+1/\sigma-1} \exp\left(-\frac{ay}{\sigma}\right)\mathrm{d}y$$

$$= \frac{a^k \exp(\beta) \prod_{j=1}^k (1-\sigma)_{n_j-1}}{\sigma\,\Gamma(n)} \sum_{i=0}^{n-1}\binom{n-1}{i}(-\tau)^i \int_{\tau^\sigma}^\infty y^{k-1-i/\sigma} \exp\left(-\frac{ay}{\sigma}\right)\mathrm{d}y$$

$$= \frac{\sigma^{k-1} \exp(\beta) \prod_{j=1}^k (1-\sigma)_{n_j-1}}{\Gamma(n)} \sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i \beta^{i/\sigma}\Gamma\left(k-\frac{i}{\sigma};\beta\right)$$

and the result desired easily follows.

A.2. Proof of proposition 3

The first part of the proof of proposition 3 exploits a martingale convergence theorem along the same lines as in theorem 8 of Pitman (2006). Let, first, Π_∞ stand for the infinite exchangeable random partition that is induced by a GG(β,σ) process and let $\mathbb{P}_{\beta,\sigma}$ denote its distribution. The random partition Π_n of the set of integers $\{1,\ldots,n\}$ is the corresponding restriction of Π_∞ and $\mathscr{F}_n$ is the σ-algebra that is generated by Π_n. Consider the Radon–Nikodym derivative of $\mathbb{P}_{\beta,\sigma}$ with respect to $\mathbb{P}_{0,\sigma}$ restricted to $\mathscr{F}_n$, i.e.

$$M_{\beta,\sigma,n} = \left.\frac{\mathrm{d}\mathbb{P}_{\beta,\sigma}}{\mathrm{d}\mathbb{P}_{0,\sigma}}\right|_{\mathscr{F}_n} = \frac{\exp(\beta)}{\Gamma(K_n)}\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\beta^{i/\sigma}\Gamma\left(K_n-\frac{i}{\sigma};\beta\right).$$

Now, $(M_{\beta,\sigma,n},\mathscr{F}_n)_{n\geqslant 1}$ is a positive martingale with respect to $\mathbb{P}_{0,\sigma}$. Note that $\mathbb{P}_{\beta,\sigma}$ and $\mathbb{P}_{0,\sigma}$ are mutually absolutely continuous on $\mathscr{F}_\infty$. Hence, by theorem 35.7 in Billingsley (1995) we have that $M_{\beta,\sigma,n} \to M_{\beta,\sigma}$ almost surely with respect to $\mathbb{P}_{0,\sigma}$ and

$$M_{\beta,\sigma} = \frac{\mathrm{d}\mathbb{P}_{\beta,\sigma}}{\mathrm{d}\mathbb{P}_{0,\sigma}}$$

is the Radon–Nikodym derivative of $\mathbb{P}_{\beta,\sigma}$ with respect to $\mathbb{P}_{0,\sigma}$ on $\mathscr{F}_\infty$. Moreover, we obviously have

$$\mathbb{E}_{0,\sigma}[M_{\beta,\sigma}] = 1. \tag{22}$$

This convergence result is now exploited to prove the asymptotic behaviour in expression (12). For this, let $(E_n)_{n \geqslant 1}$ be a sequence of IID random variables having a negative exponential distribution with parameter 1. Moreover, suppose that the E_ns are independent of K_n. Set $\mathscr{E}_n := \Sigma_{j=1}^{K_n} E_j$ and note that, conditionally on K_n, $\mathscr{E}_n$ has a gamma distribution with expected value K_n. We can then rewrite $M_{\beta,\sigma,n}$ as

$$M_{\beta,\sigma,n} = \frac{\exp(\beta)}{\Gamma(K_n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \int_\beta^\infty y^{K_n - i/\sigma - 1} \exp(-y) \, \mathrm{d}y$$

$$= \frac{\exp(\beta)}{\Gamma(K_n)} \int_\beta^\infty y^{K_n - 1} \exp(-y) \left(1 - \frac{\beta^{1/\sigma}}{y^{1/\sigma}}\right)^{n-1} \mathrm{d}y$$

$$= \exp(\beta) \mathbb{E}\left[\mathbb{I}_{(\beta,+\infty)}(\mathscr{E}_n) \left(1 - \frac{\beta^{1/\sigma}}{\mathscr{E}_n^{1/\sigma}}\right)^{n-1} \middle| K_n\right],$$

where $\mathbb{I}_A$ denotes, as usual, the indicator function of a set A. From the law of large numbers, $\mathscr{E}_n/K_n \to 1$, almost surely and conditionally on K_n. This entails that $\mathscr{E}_n$ is eventually larger than β, whatever the choice of β. Hence, the indicator function in the conditional expectation is eventually 1. Using the dominated convergence theorem, we have

$$M_{\beta,\sigma,n} \sim \exp(\beta) \left\{1 - \frac{\beta^{1/\sigma}}{K_n (\mathscr{E}_n/K_n)^{1/\sigma}}\right\}^n \sim \exp(\beta) \left(1 - \frac{\beta^{1/\sigma}}{K_n^{1/\sigma}}\right)^n \sim \exp\left(\beta - \beta^{1/\sigma} \frac{n}{K_n^{1/\sigma}}\right)$$

as $n \to \infty$. Since $M_{\beta,\sigma,n} \to M_{\beta,\sigma}$ almost surely with respect to $\mathbb{P}_{0,\sigma}$, then there is some random variable, say Z_σ, such that

$$n/K_n^{1/\sigma} \to Z_\sigma$$

almost surely, with respect to $\mathbb{P}_{0,\sigma}$, as $n \to \infty$. From equation (22), it follows that

$$\mathbb{E}_{0\sigma}[\exp(-\beta^{1/\sigma} Z_\sigma)] = \exp(-\beta)$$

and, then, Z_σ must coincide with a σ-stable random variable, with respect to the distribution $\mathbb{P}_{0,\sigma}$. This implies that $K_n/n^\sigma \to Z_\sigma^{-\sigma} \equiv S_\sigma$, almost surely with respect to $\mathbb{P}_{0,\sigma}$. Since the $\mathbb{P}_{0,\sigma}$-law of Z_σ is σ stable, the $\mathbb{P}_{0,\sigma}$-law of S_σ is the σ-Mittag–Leffler distribution whose density is given by

$$g_\sigma(s) = f_\sigma(s^{-1/\sigma})/\sigma s^{1+1/\sigma}$$

and f_σ is the density function of Z_σ. Moreover, since the probability measures $\mathbb{P}_{\beta,\sigma}$ and $\mathbb{P}_{0,\sigma}$ are mutually absolutely continuous, almost sure convergence holds true with respect to $\mathbb{P}_{\beta,\sigma}$, as well. To deduce the $\mathbb{P}_{\beta,\sigma}$-law of S_σ, it is sufficient to exploit a change of measure that is suggested by

$$\mathbb{P}_{\beta,\sigma}(A) = \int_A \frac{\mathrm{d}\mathbb{P}_{\beta,\sigma}}{\mathrm{d}\mathbb{P}_{0,\sigma}} \, \mathrm{d}\mathbb{P}_{0,\sigma}$$

and by the fact that

$$\frac{\mathrm{d}\mathbb{P}_{\beta,\sigma}}{\mathrm{d}\mathbb{P}_{0,\sigma}} = M_{\beta,\sigma} = \exp(\beta - \beta^{1/\sigma} Z_\sigma).$$

The change of variable $Z_\sigma = S_\sigma^{-1/\sigma}$ leads to the desired conclusion in equation (13). Finally, for the case $\sigma = \frac{1}{2}$, some algebra leads to the simple density that is given in equation (14).

References

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

Arratia, R., Barbour, A. D. and Tavaré, S. (2003) *Logarithmic Combinatorial Structures: a Probabilistic Approach*. Zürich: European Mathematical Society.

Barry, D. and Hartigan, J. A. (1993) A Bayesian analysis for change point problems. *J. Am. Statist. Ass.*, **88**, 309–319.

Berry, D. A. and Christensen, R. (1979) Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.*, **7**, 558–568.

Billingsley, P. (1995) *Probability and Measure*, 3rd edn. New York: Wiley.

Brix, A. (1999) Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.*, **31**, 929–953.

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

Charalambides, C. (2005) *Combinatorial Methods in Discrete Distributions*. New York: Wiley.

Charalambides, C. A. and Singh, J. (1988) A review of the Stirling numbers, their generalizations and statistical applications. *Communs Statist. Theory Meth.*, **17**, 2533–2595.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An ANOVA model for dependent random measures. *J. Am. Statist. Ass.*, **99**, 205–215.

Doornik, J. A. (2002) *Object-oriented Matrix Programming Using Ox*, 3rd edn. Oxford: Timberlake.

Epifani, I., Lijoi, A. and Prünster, I. (2003) Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, **90**, 791–808.

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

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theor. Popln Biol.*, **3**, 87–112.

Gnedin, A. and Pitman, J. (2005) Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. St Peterburg. Otdel. Mat. Inst. Steklov.*, **325**, 83–102.

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.*, **28**, 355–375.

Hartigan, J. A. (1990) Partition models. *Communs Statist. Theory Meth.*, **19**, 2745–2756.

Hjort, N. L. (2000) Bayesian analysis for a generalized Dirichlet process prior. *Statistics Research Report 7*. University of Oslo, Oslo.

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Ass.*, **96**, 161–173.

Ishwaran, H. and James, L. F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sin.*, **13**, 1211–1235.

James, L. F. (2002) Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Technical Report*. Hong Kong University of Science and Technology, Hong Kong. (Available from http://arxiv.org/pdf/math/0205093.)

James, L. F. (2006) Poisson calculus for spatial neutral to the right processes. *Ann. Statist.*, **34**, 416–440.

James, L. F. (2007) Spatial neutral to the right species sampling mixture models. In *Essays in Honor of Kjell A. Doksum* (ed. V. Nair), pp. 425–440. Singapore: World Scientific Press.

James, L. F., Lijoi, A. and Prünster, I. (2005) Bayesian inference for classes of normalized random measures. *Working Paper 5/05*. International Centre for Economic Research, Turin.

Kingman, J. F. C. (1975) Random discrete distributions (with discussion). *J. R. Statist. Soc.* B, **37**, 1–22.

Korwar, R. M. and Hollander, M. (1973) Contributions to the theory of Dirichlet processes. *Ann. Probab.*, **1**, 705–711.

Lijoi, A., Mena, R. H. and Prünster, I. (2005) Hierarchical mixture modeling with normalized inverse Gaussian priors. *J. Am. Statist. Ass.*, **100**, 1278–1291.

Liu, J. S. (1996) Nonparametric hierarchical Bayes via sequential imputations. *Ann. Statist.*, **24**, 911–930.

Lo, A. Y. (1984) On a class of Bayesian nonparametric estimates: I, density estimates. *Ann. Statist.*, **12**, 351–357.

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

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

MacEachern, S. N. (1999) Dependent nonparametric processes. *Proc. Bayes. Statist. Sci. Sect. Am. Statist. Ass.*, 50–55.

MacEachern, S. N. and Müller, P. (1998) Estimating mixture of Dirichlet process models. *J. Computnl Graph. Statist.*, **7**, 223–239.

Marin, J. M., Mengersen, K. and Robert, C. P. (2005) Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics*, vol. 25 (eds D. Dey and C. R. Rao), pp. 459–507. Amsterdam: North-Holland.

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

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Reltd Flds*, **102**, 145–158.

Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell* (eds T. S. Ferguson, L. S. Shapley and J. B. MacQueen), pp. 245–267. Hayward: Institute of Mathematical Statistics.

Pitman, J. (2003) Poisson–Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed* (ed. D. R. Goldstein), pp. 1–35. Hayward: Institute of Mathematical Statistics.

Pitman, J. (2006) Combinatorial stochastic processes. *Lect. Notes Math.*, **1875**.

Prünster, I. (2002) Random probability measures derived from increasing additive processes and their application to Bayesian statistics. *PhD Thesis*. University of Pavia, Pavia.

Quintana, F. A. (1998) Nonparametric Bayesian analysis for assessing homogeneity in $k \times l$ contingency tables with fixed right margin totals. *J. Am. Statist. Ass.*, **93**, 1140–1149.

Quintana, F. A. and Iglesias, P. L. (2003) Bayesian clustering and product partition models. *J. R. Statist. Soc.* B, **65**, 557–574.

Regazzini, E., Lijoi, A. and Prünster, I. (2003) Distributional results for means of random measures with independent increments. *Ann. Statist.*, **31**, 560–585.