

Bayesian clustering in nonparametric hierarchical mixture models

Clustering bayesiano nei modelli mistura nonparametrici

Antonio Lijoi¹, Ramses H. Mena², Igor Prünster³

¹ Dipartimento di Economia Politica e Metodi Quantitativi, Università di Pavia
e-mail: lijoi@unipv.it

² Departamento de Probabilidad y Estadística IIMAS–UNAM, México

³ Dipartimento di Statistica e Matematica Applicata “D. de Castro” e ICER, Università di Torino

Riassunto: Il problema della stima del numero di componenti di una mistura che ha generato un insieme di dati può essere affrontato in ambito bayesiano mediante l'utilizzo di modelli mistura nonparametrici. Lo sviluppo di appropriati metodi di simulazione ha favorito la diffusione in ambito applicativo di modelli mistura gerarchici basati sul processo di Dirichlet. Vengono qui studiati modelli nonparametrici alternativi alla mistura del processo di Dirichlet e confrontate le diverse strutture di clustering dei dati. L'analisi è completata dallo studio di due insiemi di dati reali.

Keywords: Bayesian nonparametric inference, Clustering, Hierarchical mixture models.

1. Introduction

The issue of clustering has been the object of a thorough investigation within the area of Bayesian nonparametrics. The key model, the so-called mixture of the Dirichlet process (MDP), dates back to Lo (1984) who proposed a random mixture density of the type

$$\tilde{f}(\cdot) = \int_{\Theta} k(\cdot, \theta) \tilde{P}(d\theta) \quad (1)$$

where $\tilde{P}$ is a Dirichlet process and $k(\cdot, \cdot)$ is a suitable kernel. This model overcomes the drawback of the Dirichlet process prior which selects (almost surely) discrete distributions, hence, making it unappealing for modeling “continuous” data. Indeed, in (1) the discreteness of the Dirichlet process is fruitfully exploited for providing a flexible model for the clustering of items: the number of distinct values generated by $\tilde{P}$ identifies the number of different clusters into which the observations may be grouped. Lo (1984) provided also the Bayes estimator, under squared loss function, of $\tilde{f}$ in (1). However, such an expression, involving a sum over partitions, turns out to be inapplicable to concrete examples with real datasets. Consequently, the popularization of the model (1) is due to a simulation based approach, relying on the Markov Chain Monte Carlo (MCMC) techniques, set forth in Escobar and West (1995). Since then there has been a remarkable growth of the literature on the subject. See Müller and Quintana (2004) and references therein.

A natural extension of the model in (1) consists in replacing the Dirichlet process with a different random discrete probability measure. It is then natural to ask whether

such a change of the mixing measure might lead to a different clustering behavior with respect to the MDP. In the literature there have been recently proposed new discrete random probability measures which are natural candidates for replacing the Dirichlet process in (1): among them we mention species sampling models (Pitman, 1996), Poisson–Kingman models (Pitman, 2003), normalized random measures with independent increments (NRMI) (Regazzini, Lijoi and Prünster, 2003), spatial neutral to the right models (James, 2006a). Any specific and tractable prior contained in the classes listed above is a potential alternative to the Dirichlet process. For example, Ishwaran and James (2001, 2003) investigate formal properties of mixture models based on species sampling priors and concretely examine the mixture of the two parameter Poisson–Dirichlet process. James (2006b) deals with mixtures of spatial neutral to the right priors.

In the present paper we focus on mixture models based on specific cases of NRMI and provide a review of the results obtained in Lijoi, Mena and Prünster (2005a, 2005b, 2006). The reader is referred to these papers for a more exhaustive account. We will show that the answer to the question of whether or not the replacement of the Dirichlet process leads to a different clustering behavior is positive for the models we are going to consider. The outline of the paper is as follows. In Section 2 we present the mixture model in hierarchical form and give details about three different mixture models: the MDP, the mixture of the normalized inverse–Gaussian (N–IG) process and the mixture of a generalized Dirichlet process. Section 3 is devoted to the mixture of the generalized gamma process and an analysis of its clustering structure is carried out. Finally, in Section 5 we analyze two real datasets which are widely used in the context of mixtures for comparing different models.

2. Hierarchical mixture models

We first present the mixture model in hierarchical form and introduce some notation. Let $(X_i)_{i \geq 1}$ be a sequence of exchangeable observable random variables with values in $\mathbb{X}$, whereas $(\theta_i)_{i \geq 1}$ is a sequence of latent random variables with values in Θ . Model (1) can then be formulated in hierarchical form as

$$\begin{aligned} X_i | \theta_i &\stackrel{\text{ind}}{\sim} k(\cdot, \theta_i) \\ \theta_i | \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P} \\ \tilde{P} &\sim \mathcal{P} \end{aligned} \tag{2}$$

where $\mathcal{P}$ stands for the distribution of the discrete random probability measure $\tilde{P}$. The discreteness of $\tilde{P}$ clearly leads to coincidences among the θ_i 's. Accordingly, for any $n \geq 1$ there are k distinct values $\theta_1^*, \dots, \theta_k^*$ within the set $\{\theta_1, \dots, \theta_n\}$ and these identify k clusters $C_1, \dots, C_k$ in which the observed data are grouped. The grouping is such that any two observations X_i and X_j belong to the same cluster if and only if the corresponding θ 's coincide, i.e. $\theta_i = \theta_j$. Our interest then focuses on the properties of the random variable K_n denoting the number of groups among n observations. In particular, one may be interested in: (i) determining the prior distribution of the number of clusters K_n induced by $\mathcal{P}$; (ii) studying the asymptotic behavior of K_n as the sample size n increases; (iii) evaluating the conditional distribution of K_n given the observables $X_1, \dots, X_n$, which represents the posterior distribution of the number of components in the mixture. Let

us now consider three different mixture models and highlight the corresponding prior distributions of K_n .

2.1. Mixtures of the Dirichlet process

As mentioned in the Introduction, the most popular nonparametric prior in the context of mixture modelling is the MDP, which is based on a Dirichlet process with parameter measure α such that $\alpha(\Theta) = a \in (0, \infty)$ and $\alpha(\{\theta\}) = 0$ for any θ in Θ . The clustering structure induced by a Dirichlet process has been already investigated by Antoniak (1974). The resulting prior distribution for K_n turns out to be

$$P[K_n = k] = \frac{a^k}{(a)_n} |s(n, k)| \quad (3)$$

where $(a)_n = a(a+1) \cdots (a+n-1)$ and $|s(n, k)|$ is the signless Stirling number of the first kind. This represents a variation of the celebrated Ewens sampling formula. It is worth noting that (3) is a highly peaked distribution, for any fixed value of $a > 0$. This is commonly seen as a drawback which can be avoided by randomizing a as done, e.g., in Escobar and West (1995).

2.2. Mixtures of the normalized inverse-Gaussian process

The normalized inverse-Gaussian process (N-IG) is a nonparametric prior that has been studied in Lijoi, Mena and Prünster (2005a). It is constructed via a family of consistent finite-dimensional distributions whose density functions on the simplex $\Delta_{m-1} = \{(w_1, \dots, w_{m-1}) : w_i \geq 0, \sum_{i=1}^{m-1} w_i \leq 1\}$ are given by

$$f(w_1, \dots, w_{m-1}) = \frac{e^{\sum_{i=1}^m \alpha_i} \prod_{i=1}^m \alpha_i}{2^{\frac{m}{2}-1} \pi^{\frac{m}{2}}} \times \frac{K_{-\frac{m}{2}}(\sqrt{\mathcal{A}_m(w_1, \dots, w_{m-1})})}{w_1^{\frac{3}{2}} \dots w_{m-1}^{\frac{3}{2}} (1 - \sum_{j=1}^{m-1} w_j)^{\frac{3}{2}} [\mathcal{A}_m(w_1, \dots, w_{m-1})]^{\frac{m}{4}}} \quad (4)$$

where $\mathcal{A}_m(w_1, \dots, w_{m-1}) = \sum_{i=1}^{m-1} \frac{\alpha_i^2}{w_i} + \frac{\alpha_m^2}{1 - \sum_{j=1}^{m-1} w_j}$, K denotes the modified Bessel function of the third type and $\alpha_i = \alpha(A_i) > 0$, the sets $\{A_i : i = 1, \dots, m\}$ being a measurable partition of Θ . Thus (4) is seen as the density function of the vector $(\tilde{P}(A_1), \dots, \tilde{P}(A_{m-1}))$ and the random probability measure $\tilde{P}$ is termed *normalized inverse-Gaussian process* with parameter measure α on Θ . The a priori clustering structure arising from such a process is described via the probability distribution of K_n which coincides with

$$P[K_n = k] = \binom{2n-k-1}{n-1} \frac{e^a (-a^2)^{n-1}}{2^{2n-k-1} \Gamma(k)} \sum_{r=0}^{n-1} \binom{n-1}{r} (-a^2)^{-r} \Gamma(k+2+2r-2n; a)$$

where $\Gamma(\nu; a) = \int_a^\infty x^{\nu-1} e^{-x} dx$ is the incomplete gamma function.

2.3. Mixtures of a generalization of the Dirichlet process

An interesting generalization of the Dirichlet process has been proposed in Regazzini, Lijoi and Prünster (2003), where it is defined as a normalized random measure with

independent increments. A description in terms of a consistent family of finite-dimensional distributions is given in Lijoi, Mena and Prünster (2005b). It is shown that such distributions are expressible in terms of the confluent form of the fourth Lauricella function. See Exton (1976) for details on Lauricella functions. In order to deduce the clustering mechanism, let us specify some useful notation. For any $\gamma > 1$ let $\mathbf{J}_{\gamma-1} = (1, \dots, \gamma-1)'$ denote $(\gamma-1)$ -dimensional vector and, for each $q = 1, \dots, \gamma$, introduce the measures

$$\alpha_q^*(\cdot; \mathbf{n}, \mathbf{r}^k) := \alpha(\cdot) + \sum_{i=1}^k n_i \delta_{x_i^*}(\cdot) \delta_q(\{r_i\}) \quad (5)$$

where $\mathbf{r}^k = (r_1, \dots, r_k)' \in \{1, \dots, \gamma\}^k$, $\mathbf{n} = (n_1, \dots, n_k)$ and δ_c is, for any c , the point mass at c . Finally, $a_q^*(\mathbf{n}, \mathbf{r}^k) = a + \sum_{i=1}^k n_i \delta_q(\{r_i\})$ denotes the total mass of the measure $\alpha_q^*(\cdot; \mathbf{n}, \mathbf{r}^k)$ and $\mathbf{a}^*(\mathbf{n}, \mathbf{r}^k) := (a_1^*(\mathbf{n}, \mathbf{r}^k), \dots, a_{\gamma-1}^*(\mathbf{n}, \mathbf{r}^k))$. If $A_{k,n} = \{(n_1, \dots, n_k) : n_i \geq 1, \sum_{i=1}^k n_i = n\}$, then the prior distribution of the number of clusters is given by

$$P[K_n = k] = \frac{1}{k!} \frac{(\gamma!)^a \Gamma(n)}{\gamma^{\gamma a} (\gamma a)_n} \sum_{(n_1, \dots, n_k) \in A_{k,n}} \frac{n!}{n_1 \cdots n_k} \sum_{\mathbf{r}^k} F_D^{(\gamma-1)} \left(\gamma a, \mathbf{a}^*(\mathbf{n}, \mathbf{r}^k); \gamma a + n; \frac{1}{\gamma} \mathbf{J}_{\gamma-1} \right).$$

Here $F_D^{(\gamma-1)}$ is the fourth Lauricella hypergeometric function. Unfortunately we are not aware of a closed form expression for the above summation and the distribution of K_n does not simplify as in the previous cases.

3. Clustering behavior with the generalized gamma process

In a recent paper, Lijoi, Mena and Prünster (2006) examine mixture models based on a discrete random probability derived from the generalized gamma process studied in Brix (1999). Such a random probability measure can be constructed by suitably normalizing the generalized gamma process according to the technique introduced in Regazzini, Lijoi and Prünster (2003) and represents a particularly tractable case of NRMI. The resulting generalized gamma prior is henceforth denoted by $\text{GG}(\sigma, \beta)$ where $\sigma \in (0, 1)$ and $\beta \in (0, \infty)$ are two parameters indexing the prior. It is worth noting that both the Dirichlet and the N-IG processes are special cases of the $\text{GG}(\sigma, \beta)$, corresponding to a suitable limit as $\sigma \rightarrow 0$ and to the choice $\sigma = 1/2$, respectively. The main advantage of embedding these two priors within this more general model relies on the availability of the parameter σ , which will be shown to control the clustering behaviour of the whole mixture model.

Let us first consider the prior distribution for the number of components K_n , which turns out to be of the form

$$P[K_n = k] = \frac{e^\beta \mathcal{C}(n, k, \sigma)}{\sigma \Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \beta^{i/\sigma} \Gamma\left(k - \frac{i}{\sigma}; \beta\right), \quad (6)$$

for any $k \in \{1, \dots, n\}$, where

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

is known as *generalized factorial coefficient*. We also agree that $\mathcal{C}(0, 0, \sigma) = 1$ and $\mathcal{C}(n, 0, \sigma) = 0$ for all $n \geq 1$. See Charalambides (2005) for a review and further details on generalized factorial coefficients. A mere graphical analysis of the distribution of K_n suggests that β plays the role of a location parameter: increasing β shifts the distribution of K_n to the right for any given σ . On the other hand, σ controls the flatness of the distribution of K_n : as σ increases the distribution of K_n tends to become flatter. Hence, loosely speaking, one can say that a large value of σ yields a non-informative prior for K_n .

As far as the asymptotic behavior of K_n is concerned, the following result provides the rate of growth of K_n as n increases. Its proof can be found in Lijoi, Mena and Prünster (2006).

Proposition 1. *If K_n is the number of clusters induced by a generalized gamma process prior with parameters (σ, β) , then*

$$\frac{K_n}{n^\sigma} \rightarrow S_\sigma \quad (7)$$

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

$$g_{\beta, \sigma}(s) = e^{\beta - (\beta/s)^{1/\sigma}} \frac{f_\sigma(s^{-1/\sigma})}{\sigma s^{1+1/\sigma}} \quad (8)$$

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

The previous result clearly suggests that σ is responsible for the number of distinct latent variables generated by the generalized gamma prior: the bigger σ , the larger number of clusters are going to be generated among the latent θ_i 's. It is interesting to compare this result with an analogous concerning the Dirichlet process proved by Korwar and Hollander (1973): if K_n is the number of clusters induced by a Dirichlet process, then almost surely

$$\frac{K_n}{\log(n)} \rightarrow a$$

where, as previously, $a = \alpha(\Theta)$. See also Pitman (2005) for similar results related to the two parameter Poisson–Dirichlet process. It is also possible to describe the asymptotic behaviour of the number of clusters of fixed size. Indeed, one easily deduces from Proposition 1 the following

Corollary 1. *If K_n and $K_{n,j}$ are the number of clusters and the number of clusters of size j , generated by a generalized gamma process prior with parameters (σ, β) , then*

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

almost surely, for any $j = 1, 2, \dots$, as $n \rightarrow +\infty$.

Hence, the asymptotic proportion of clusters of size j equals $p_{\sigma,j}$ and, in particular, the asymptotic proportion of clusters of size 1 coincides with σ . As for the tail behaviour of $p_{\sigma,j}$, one has a power law decay of index $(\sigma + 1)$ for the proportion of clusters of size j , as j increases.

Let us now consider the posterior distribution of the number of components in the mixture, which represents the main object of study. Indeed, it turns out that it is possible to determine an expression for it, but since it depends on a sum over partitions it is hardly applicable. Consequently, one has to resort to a simulation based approach. For instance, the generalized Pólya urn sampler of Ishwaran and James (2001) is easily adapted to the present framework since it just requires the knowledge of the predictive distributions. It is precisely by looking closely at the predictive distributions that we are able to give an heuristic explanation of the posterior behavior of a $\text{GG}(\sigma, \beta)$ process and to highlight a reinforcement mechanism driven by σ . They are given by

$$P[\theta_{n+1} \in d\theta \mid \theta_1, \dots, \theta_n] = w_0^{(n,k)} P_0(d\theta) + w_1^{(n,k)} \sum_{j=1}^k (n_j - \sigma) \delta_{\theta_j^*}(d\theta) \quad (9)$$

where $P_0(d\theta) = E[\tilde{P}(d\theta)]$ and $(\theta_1^*, \dots, \theta_k^*)$ represent the k distinct values in $(\theta_1, \dots, \theta_n)$. Denote by n_i the frequency of θ_i , for $i = 1, \dots, k$. For an exact expression of the weights $w_0^{(n,k)}$ and $w_1^{(n,k)}$ see Lijoi, Mena and Prünster (2006). The predictive distribution in (9) can be seen as the result of a two stage procedure: the latent variable θ_{n+1} is either “new” (i.e. not coinciding with any of the previously observed θ_i^* ’s) or “old” with probability depending on n and k but not on the frequencies n_i ’s. Given θ_{n+1} is “new”, it is sampled from P_0 . Given θ_{n+1} is “old” (i.e. θ_{n+1} is equal to one of the already sampled θ_i^* ’s), it will coincide with a particular θ_j^* with probability $(n_j - \sigma)/(n - k\sigma)$. These probabilities are determined by the size n_j of each cluster and by σ . At this stage, a reinforcement mechanism driven by σ takes place. Indeed, one can see that the ratio of the probabilities assigned to any pair of (θ_i^*, θ_j^*) is given by $(n_i - \sigma)/(n_j - \sigma)$. As $\sigma \rightarrow 0$, 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 shifted from θ_j^* to θ_i^* . This means that the sampling procedure tends to reinforce, among the observed clusters, those having higher frequencies and correspondingly penalize clusters of small size.

Combining these remarks with the asymptotic results, one can expect a $\text{GG}(\sigma, \beta)$ mixture to behave as follows. For a large value of σ , a large number of clusters is going to be generated. Among these clusters, most will have small size. The reinforcement mechanism acts on the mass allocation among the clusters by penalizing clusters of small size and favoring those few groups containing a large number of elements. Such a behavior is of great interest for inferential purposes and a large value of σ seems a desirable feature for a model. However, there are some constraints on the choice of σ : since, as noted previously, σ controls also the flatness of the prior, a large σ is often incompatible with the natural strategy for prior specification which consists in tuning the parameters on the expected number of components. This will become clear in the next section. The solutions to this problem are represented by the choice of the largest possible σ compatible with the desired prior specification or by the randomization of σ , in which the data select the appropriate reinforcement rate. The latter is not pursued here and the reader is deferred to Lijoi, Mena and Prünster (2006).

4. Some real data examples

We now analyze two real data sets and highlight some of the features of mixtures based on the generalized gamma process. The model we adopt is a normal mixture where both means and variances are random and chosen according to a $\text{GG}(\sigma, \beta)$ prior, i.e.,

$$\begin{aligned} (Y_i | m_i, V_i) &\stackrel{\text{ind}}{\sim} N(Y_i | m_i, V_i), \quad i = 1, \dots, n \\ (m_i, V_i | \tilde{P}) &\stackrel{\text{iid}}{\sim} \tilde{P} \\ \tilde{P} &\sim \text{GG}(\sigma, \beta), \end{aligned}$$

where N is a normal kernel. We aim at comparing the different behaviors as σ varies. Hence, given σ , the parameter β , which controls the location, is fixed basing on the prior guess on the number of components present in the mixture: in particular, if one expects K_n to be k , then β is chosen in such a way that the mode of the prior distribution of K_n is in k . Indeed, there is an interval (β', β'') satisfying this requirement and $\beta^* \in (\beta', \beta'')$ is fixed such that $P_{(\beta^*, \sigma)}(K_n = k) = \max_{\beta \in (\beta', \beta'')} P_{(\beta, \sigma)}(K_n = k)$. For the remaining part concerning P_0 , we employ the nowadays quite standard semiparametric prior specification of Escobar and West (1995), namely $P_0(dx dv) = N(x | \mu, \tau v^{-1}) \text{Ga}(v | 2, 1) dx dv$, where $\text{Ga}(\cdot | c, d)$ is the density corresponding to a gamma distribution with mean c/d . A further hierarchy is assumed for μ and τ , i.e. $\mu \sim N(\cdot | 0, .001)$ and $\tau^{-1} \sim \text{Ga}(\cdot | 1, 100)$.

Simulations were carried out adopting a generalized Pólya urn sampling scheme similar to the one set forth in Ishwaran and James (2001). All inferences are based on 10000 iterations after a burn-in period of 2000 sweeps.

4.1. Galaxy data

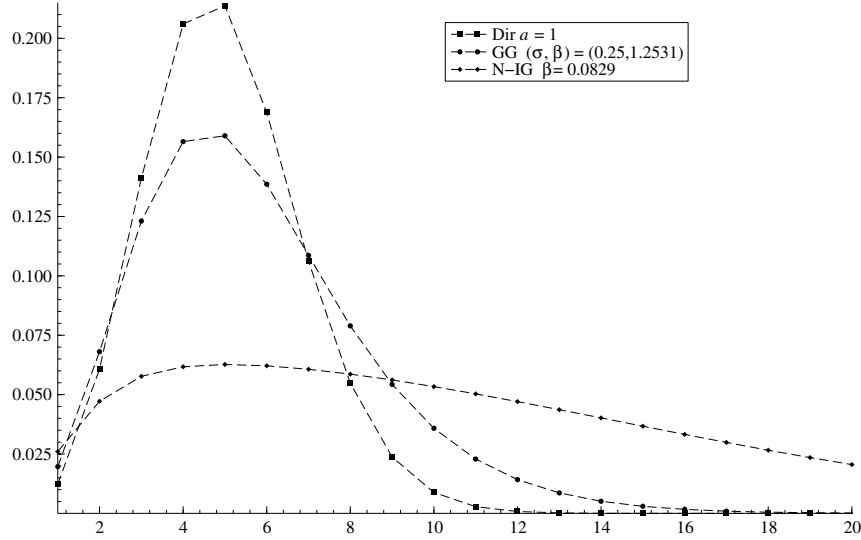
In this section we reanalyze an astronomy data set consisting of the velocities (in 10^3 km/s) of 82 distant galaxies diverging from our own galaxy.

Table 1: Prior and posterior probabilities for the number of components K_{82} corresponding to the Dirichlet, the GG and the N-IG mixtures for the 82 galaxy data. Their parameters are fixed such that the mode of the prior distribution of K_{82} is in 5.

$n = 82$	Dirichlet		GG		N-IG	
	$a = 1$		$(\sigma, \beta) = (.25, 1.2531)$		$\beta = .0829$	
	Prior	Posterior	Prior	Posterior	Prior	Posterior
$k \leq 3$	.2140		.2108	.0585	.1309	.003
$k = 4$	.2060	.0465	.1565	.1189	.0617	.0754
$k = 5$	.214	.125	.159	.2089	.0627	.168
$k = 6$	.169	.2524	.1386	.2435	.0621	.2301
$k = 7$	.106	.2484	.1086	.1775	.0607	.2338
$k = 8$	.0548	.2011	.0789	.1058	.0586	.1225
$k = 9$	.0237	.08	.0543	.0522	.0562	.0941
$k = 10$	.0088	.019	.0358	.0225	.0534	.0352
$k \geq 11$	.0038	.0276	.0574	.0122	.4537	.0379

The physical problem can be sketched as follows: after the Big Bang, it is believed that the matter expanded a tremendous rate and galaxies were formed due to the local

Figure 1: Prior distributions for the number of components K_{82} corresponding to the Dirichlet, the GG and the N-IG mixtures for the 82 galaxy data. Their parameters are specified such that the mode of the prior distribution of K_{82} is in 5. The probabilities are connected by lines only for visual simplification.



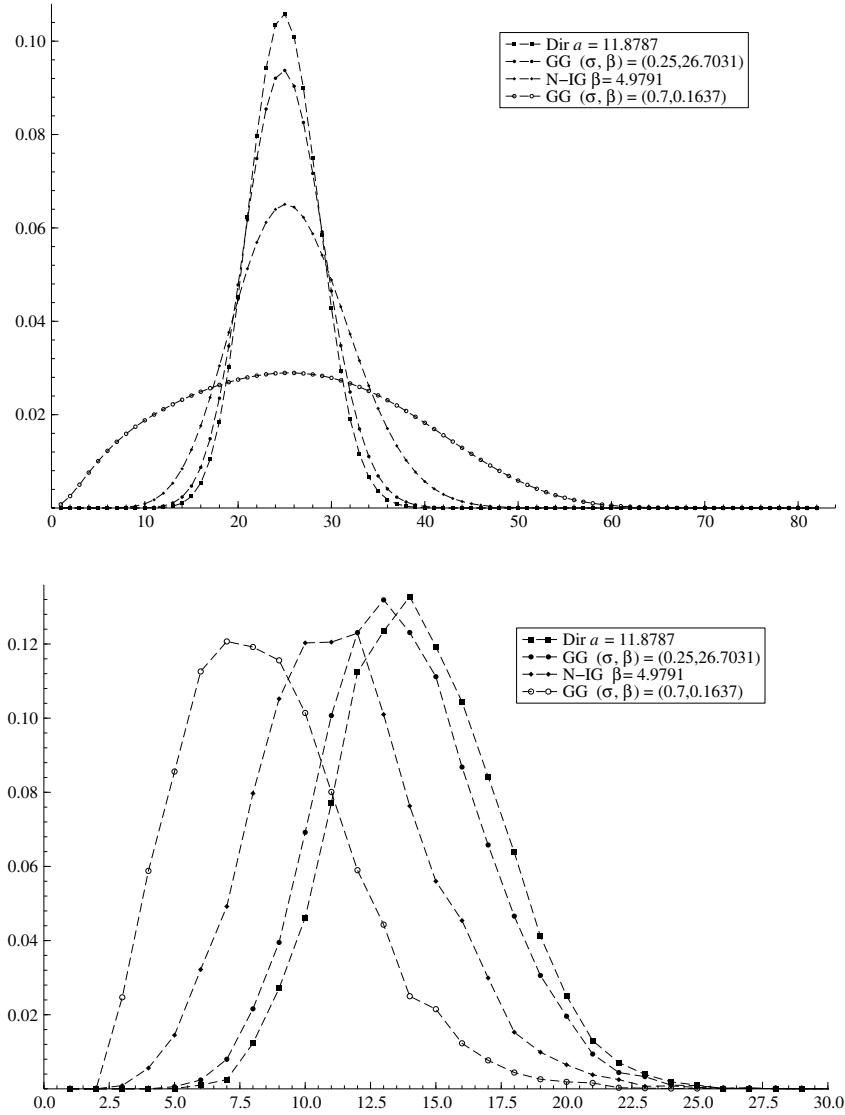
attraction of matter. Astronomers predicted that the gravitational pull would lead to some clustering of galaxies, but there are data which suggest the presence of superclusters surrounded by large voids, a phenomenon which is yet not understood. Hence, it is of interest to investigate the distribution of the velocities since multimodality might suggest the presence of superclusters. More details can be found in Roeder (1990). These data have been widely studied using different mixture models. See, e.g., Escobar and West (1995), Richardson and Green (1997) and McGrory and Titterton (2006). The results in the literature range from 3 up to 8 clusters.

This dataset was analyzed also in Lijoi, Mena and Prünster (2005a) comparing a Dirichlet with a N-IG mixture. The parameters were tuned such that the prior distributions for K_{82} have mode in 5. Here we compare these results with a generalized gamma process with $\sigma = .25$ which leads to select $\beta = 1.2531$. Note that no value of $\sigma > .52$ is compatible with the constraint on the mode. Figure 1 displays the prior distributions for K_{82} induced by the three processes, whereas Table 1 reports prior and posterior probabilities for K_{82} .

In order to appreciate the reinforcement mechanism inherent to a generalized gamma process and at the same time investigate the sensitivity we consider also a different setup. To this end we specify the parameters in such a way that the mode of the prior distributions for K_{82} is in 25, hence far away from the estimated number of components in previous studies. This leads to consider a Dirichlet process with $a = 11.8787$, a GG process with $(\sigma, \beta) = (.25, 26.7031)$, a N-IG process with $\beta = 4.9791$ and a GG process with $(\sigma, \beta) = (.7, .1637)$. Table 2 and Figure 2 display the prior and posterior probabilities on K_{82} ; the posterior probabilities greater than .1 are highlighted in bold in the table. It is worth remarking that the higher σ the more the posterior probabilities move away from the “wrong” prior specification. This is due to a combination of the flatness of the

prior on K_{82} and the reinforcement mechanism. As noted in the previous section a large value of σ is desirable though not always selectable due to the constraints it imposes on the distribution of K_n : hence, it seems a good advice to select the largest possible σ compatible with the prior information on the number of clusters.

Figure 2: Prior and posterior distributions for the number of components K_{82} corresponding to the Dirichlet, the N-IG and the two GG mixtures for the 82 galaxy data. Their parameters are specified such that the mode of the prior distribution of K_{82} is in 25. The probabilities are connected by lines only for visual simplification.



4.2. Acidity data

The second data set we analyze concerns the environmental problem of acidification. The data consist of measurements of an acid neutralizing capacity (ANC) index in a

Table 2: Prior and posterior probabilities for the number of components K_{82} corresponding to the Dirichlet, the N-IG and the two GG mixtures for the 82 galaxy data. Their parameters are set such that the mode of the prior distribution of K_{82} is in 25.

$n = 82$	Dirichlet $a = 11.8787$		GG $(\sigma, \beta) = (.25, 26.7031)$		N-IG $\beta = 4.9791$		GG $(\sigma, \beta) = (.7, .1637)$	
	Prior	Posterior	Prior	Posterior	Prior	Posterior	Prior	Posterior
$k \leq 3$						.0009	.0083	.0247
$k = 4$						.0056	.0076	.0588
$k = 5$		.0001		.0006		.0145	.0100	.0856
$k = 6$		.0010		.0024		.0322	.0122	.1126
$k = 7$		.0025		.0080	.0001	.0492	.0142	.1207
$k = 8$		.0123		.0216	.0002	.0797	.0159	.1192
$k = 9$		.0271		.0395	.0004	.1052	.0174	.1156
$k = 10$		.0462	.0001	.0692	.0009	.1203	.0188	.1014
$k = 11$		.0770	.0002	.1007	.0018	.1205	.0200	.0801
$k = 12$	.0001	.1124	.0004	.1231	.0032	.1229	.0212	.0590
$k = 13$	.0004	.1234	.0011	.1319	.0053	.101	.0222	.0443
$k = 14$	.0010	.1328	.0023	.1231	.0084	.0763	.0232	.0250
$k = 15$	.0025	.1193	.0047	.1112	.0125	.056	.0241	.0215
$k = 16$	.0053	.1044	.0087	.0868	.0176	.0454	.0249	.0123
$k = 17$	.0104	.0842	.0148	.0658	.0237	.0299	.0257	.0077
$k = 18$	.0185	.0638	.0235	.0466	.0305	.0153	.0264	.0044
$k = 19$	.0302	.0412	.0347	.0306	.0376	.0099	.0270	.0026
$k = 20$	.0452	.0251	.0478	.0196	.0447	.0065	.0275	.0019
$k = 21$	.0624	.0130	.0617	.0094	.0512	.0038	.0280	.0016
$k = 22$	.0796	.0071	.0749	.0044	.0569	.0025	.0283	.0003
$k = 23$	.0942	.0039	.0855	.0033	.0612	.0007	.0286	.0002
$k = 24$	.1035	.0019	.0921	.0010	.0640	.0008	.0288	.0001
$k = 25$	.1059	.0010	.0937	.0009	.0651	.0004	.0289	.0001
$k = 26$	.1010	.0001	.0904	.0001	.0644	.0002	.0289	.0001
$k = 27$	.0899	.0001	.0826	.0002	.0623	.0003	.0288	.0001
$k = 28$	.0750	.0001	.0717		.0587		.0286	.0001
$k = 29$	.0585		.0592		.0542		.0283	
$k = 30$	.0428		.0465		.0488		.0279	
$k \geq 31$	.0736		.1035		.2264		.3681	

sample of 155 lakes in North–Central Wisconsin, USA. A low value of ANC can lead to a loss of biological resources. The identification of clusters of lakes is important for the determination of lake characteristics which can be used to predict higher acidification. Also these data were studied by several authors and were considered on a log–scale as we do. Richardson and Green (1997) obtain a fairly equal support for 3–5 components. In the discussion of the above paper McLachlan and Peel (1997) and Schlattmann and Böhning (1997) have shown how different frequentist approaches tend to favor 2–3 components in accordance with the Bayesian study of Crawford (1994). Recently, exploiting a variational approximation in finite Bayesian mixtures, McGrory and Titterton (2006) support the existence of only two clusters.

In order to compare our model with previous contributions we set our prior guess on K_{155} relatively high: we tune the parameters in order to achieve the mode of the distribution of K_{155} in 8. This results in $a = 1.634$ for the Dirichlet mixture, $\beta = 2.1734$ for the GG mixture with $\sigma = .25$ and $\beta = .1464$ for the N-IG mixture. Note that no value of $\sigma > .56$ is compatible with this choice. Table 3 reports the prior and posterior probabilities on the number of clusters. Although we tuned the priors having mode in 8, it is apparent that all three mixtures favor neatly 2 components and allocate probability greater than .9 to 2 or 3 components. Thus, our results are in complete accordance with McGrory and Titterton (2006).

Table 3: Prior and posterior probabilities for the number of components K_{155} corresponding to the Dirichlet, the GG and the N-IG mixtures for the 155 acidity data. Their parameters are specified such that the mode of the prior distribution of K_{155} is in 8.

$n = 155$	Dirichlet $a = 1.634$		GG $(\sigma, \beta) = (.25, 2.1734)$		N-IG $\beta = .1464$	
	Prior	Posterior	Prior	Posterior	Prior	Posterior
$k = 1$	.0004		.0014		.0071	
$k = 2$	.0035	.7197	.0083	.6673	.0185	.6711
$k = 3$	.0154	.2285	.0250	.2580	.0291	.2382
$k = 4$	.0426	.0444	.0516	.0585	.0367	.0650
$k = 5$	.0849	.0067	.0820	.0140	.0413	.0198
$k = 6$	.1301	.0006	.1074	.0021	.0439	.0037
$k = 7$	.1605	.0001	.1214	.0001	.0451	.0015
$k = 8$	.1642		.1222		.0455	.0005
$k = 9$	.1426		.1121		.0452	.0001
$k \geq 10$	.2558		.3686		.6875	.0001

References

- Antoniak C.E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics*, 2, 1152–1174.
- Brix A. (1999) Generalized gamma measures and shot-noise Cox processes, *Advances in Applied Probability*, 31, 929–953.

- Charalambides C. (2005) *Combinatorial methods in discrete distributions*, Wiley, New York.
- Crawford, S.L. (1994) An application of the Laplace method to finite mixture distributions, *Journal of the American Statistical Association*, 89, 259–267.
- Escobar M.D., West, M. (1995) Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association*, 90, 577–588.
- Exton H. (1976). *Multiple hypergeometric functions and applications*, Ellis Horwood, Chichester.
- Ishwaran H., James L. F. (2001) Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, 96, 161–173.
- Ishwaran H., James L. F. (2003) Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica*, 13, 1211–1235.
- James L.F. (2006a) Poisson calculus for spatial neutral to the right processes, *The Annals of Statistics*, to appear.
- James L.F. (2006b) Spatial neutral to the right mixture models, in: *A Festschrift in honor of Kjell Doksum*, IMS Lecture Notes Monograph Series, to appear.
- Korwar R.M., Hollander M. (1973) Contributions to the theory of Dirichlet processes, *The Annals of Probability*, 1, 705–711.
- Lijoi A., Mena R.H., Prünster, I. (2005a) Hierarchical mixture modeling with normalized inverse Gaussian priors, *Journal of the American Statistical Association*, 100, 1278–1291.
- Lijoi A., Mena R.H., Prünster I. (2005b) Bayesian nonparametric analysis for a generalized Dirichlet process prior, *Statistical Inference for Stochastic Processes*, 8, 283–309.
- Lijoi A., Mena R.H., Prünster, I. (2006) Controlling the reinforcement in Bayesian nonparametric mixture models, *Tech. Report*.
- Lo A.Y. (1984) On a class of Bayesian nonparametric estimates: I Density estimates *The Annals of Statistics*, 12, 351–357.
- McGrory C.A., Titterton D.M. (2006) Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions, *Tech. Report* 06–01, University of Glasgow.
- Müller P., Quintana, F.A. (2004) Nonparametric Bayesian data analysis, *Statistical Science*, 19, 95–110.
- Pitman J. (1996) Some developments of the Blackwell-MacQueen urn scheme, in: *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, Ferguson T.S., et al. (Eds.), IMS Lecture Notes Monograph Series, 30, 245–267.
- Pitman J. (2005) *Combinatorial stochastic processes*, Lecture notes for St. Flour Summer School.
- Pitman J. (2003) Poisson–Kingman partitions, in: *Science and Statistics: A Festschrift for Terry Speed*, Goldstein D.R. (Ed.), IMS Lecture Notes Monograph Series, 40, 1–35.
- Regazzini E., Lijoi A., Prünster I. (2003) Distributional results for means of random measures with independent increments, *The Annals of Statistics*, 31, 560–585.
- Richardson, S., Green, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society Series B*, 59, 731–792.