A probability for classification based on the Dirichlet process mixture model

Ruth Fuentes–García*, Ramsés H. Mena** and Stephen G. Walker***

* Facultad de Ciencias, Universidad Nacional Autónoma de México. México, D.F. 04510, México.

** IIMAS, Universidad Nacional Autónoma de México. México, D.F. 04510, México.

*** [1] University of Kent, Canterbury, Kent, CT2 7NZ, UK.

Abstract

In this paper we provide an explicit probability distribution for classification purposes when observations are viewed on the real line and classifications are to be based on numerical orderings. The classification model is derived from a Bayesian nonparametric mixture of Dirichlet process model; with some modifications. The resulting approach then more closely resembles a classical hierarchical grouping rule in that it depends on sums of squares of neighboring values. The proposed probability model for classification relies on a numerical procedure based on a reversible Markov chain Monte Carlo (MCMC) algorithm for determining the probabilities. Some numerical illustrations comparing with alternative ideas for classification are provided.

Keywords: Classification; MCMC sampling; MDP model.

1. Introduction. Suppose we observe data $y := (y_1, \ldots, y_n)$ with each $y_i \in \mathbb{R}$. The aim is to classify these data into $k \leq n$ groups and to determine which ones are in the same group. This is a classic problem and current Bayesian approaches rely on mixture models, such as the one described in Richardson and Green (1997), or the mixture of Dirichlet process (MDP) model (see, for example, Lo (1984) and Escobar (1994)). In the Richardson and Green model the k is modeled explicitly via

$$p(y|k) = \sum_{j=1}^{k} w_{j,k}\, f(y; \theta_j),$$

[1] For correspondence: S.G.Walker@kent.ac.uk

where the $f(\,\cdot\,;\theta_j)$ is a density function, commonly chosen to be the Gaussian density, and $w_k := (w_{j,k})_{j=1}^k$ denotes a sequence of weights which sum to one and $\theta^{(k)} = (\theta_j)_{j=1}^k$ are parameters. Prior distributions are assigned to $(w_k, \theta^{(k)}|k)$ for each k and a prior is assigned to k, and inference is made possible, e.g., via reversible jump MCMC, (Green, 1995). The augmented likelihood function for n observations is given by

$$l(k, w, \theta; y, d) \propto \prod_{i=1}^{n} w_{d_i,k}\, f(y_i; \theta_{d_i}).$$

Here, the $d := (d_i)_{i=1}^n$ denotes a sequence of latent variables acting as classification labels; that is, d_i indicates the component, less than or equal to k, from which the ith observation is coming from.

On the other hand, an alternative approach is based on Bayesian non-parametric mixture models, specifically the widely used mixture of Dirichlet process (MDP) model which is set as

$$p(y) = \sum_{j=1}^{\infty} w_j\, f(y; \theta_j),$$

where the weights $(w_j)_{j=1}^{\infty}$ can be represented in a stick-breaking form as $w_j = v_j \prod_{l<j}(1 - v_l)$ with the v_l's being independent and identically distributed (i.i.d) as beta$(1, \alpha)$ random variables, and the θ_j's are also i.i.d. from π, the prior guess at the shape of the Dirichlet process (see Sethuraman, 1994). In this case the corresponding augmented likelihood function is given by

$$l(w, \theta; y, d) \propto \prod_{i=1}^{n} w_{d_i}\, f(y_i; \theta_{d_i}).$$

Hence, we emphasize that in the Richardson and Green model, the k is explicit, but in the MDP model it is implicit, and taken to be the number of distinct d_i's.

Both of these models classify roughly in terms of numerical orderings of the data. So two observations are in the same cluster if they are both close to one of the means of the mixing densities. Yet both models put positive probability on configurations which would not be chosen based on an ordering rule. Given only the data, we seek to classify the observations into groups which strictly adhere to an ordering constraint. We adjust the MDP model to do this; effectively putting zero probability on classifications which violate this rule.

We do acknowledge this is not the only way observations could be classified; there are others. Perhaps there are ways to adjust the MDP model to achieve other types of rules and this can be investigated in the future. But for this paper we concentrate solely on the ordering rule; because it is probably the most common and because it is implicit in the nature of the Bayesian models.

The MDP approach to the classification problem leads to posterior inference for the number of distinct groups k given by

$$p(k \mid y) \propto \sum_{\mathcal{P}^k_{[n]}} \left\{ \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^{k} (n_j - 1)! \right\} \int \prod_{j=1}^{k} \prod_{i \in \mathcal{J}_j} f(y_i \mid \theta)\, \pi(d\theta), \quad k = 1, \ldots, n, \tag{1}$$

where $n_j = \sum_{i=1}^{n} \mathbf{1}(d_i = j)$ denotes the number of observations in group j, $\mathcal{J}_j = \{i : d_i = j\}$ is the corresponding set of indices, and $\mathcal{P}^k_{[n]}$ is the set of all partitions of the set $[n] := \{1, \ldots, n\}$ into k groups, namely all possible partitions of the data y into k disjoint groups. We refer to Lijoi and Prünster (2009) for a recent account of Bayesian non-parametric methods. There are $S_{n,k}$ (a Stirling number of the second kind) ways to partition the data into k groups and to obtain the normalising constant a Bell number $B_n = \sum_k S_{n,k}$ of computations is required. This implies that exact computation of (1) is computationally unfeasible even for small values of n, e.g. $B_{10} = 115975$ and $B_{20} > 517 \times 10^{12}$. Therefore, the need to resort to suitable simulation schemes arise. Within the Bayesian framework, these are typically based on MCMC algorithms and specifically for the case of MDP models the preferred Markov chain to navigate through the space of set partitions is constructed via the Pólya urn scheme. See for instance Escobar and West (1995), Ishwaran and James (2001) and Lijoi *et al.* (2007). A problem of a similar dimension is faced when dealing with product partition models, see for instance Quintana and Iglesias (2003).

However, we are not convinced that either of these models are suitable for classification purposes when the aim is numerical ordered classification. Our point is better understood if we set f to be the Gaussian density $\mathrm{N}(\cdot; \mu, \lambda^{-1})$ and $\pi(\mu, \lambda) = \pi(\mu \mid \lambda)\pi(\lambda)$ a conjugate prior with

$$\pi(\mu|\lambda) = \mathrm{N}(\mu; 0, (c\lambda)^{-1}) \quad \text{and} \quad \pi(\lambda) = \mathrm{gamma}(\lambda; a, b).$$

Hereafter we work under these specifications.

The problem is that locations of the Gaussian distributions, the μ_j's, can be arbitrarily close to each other and therefore register as different clusters. They could be close to each other since the model is first and foremost modeling the data and this does not preclude two densities being close to each other if the data deem this appropriate. Locally, two normals model the true density at this place better than one, yet only one group is at this place. The density of this group may be skewed, for example. But two μ_j's close to each other still register as two clusters. Such a scenario may well happen, for example, as we have just mentioned, when clusters are not Gaussian based. So both methods would over–estimate the number of clusters when clusters are not Gaussian based. We will discuss this issue later in Section 4 when we do some numerical illustrations. Nevertheless, it is worth mentioning that the issue of overestimation of the number of clusters for the MDP model with a Gaussian kernel has already been known; see, for example, the results reported in McGrory and Titterington (2007).

Our approach is not model based yet the starting point is the MDP model. From the MDP model we compute $p(d|y)$, by integrating out the (w, μ, λ). But this $p(d|y)$ will include many classifications which are not suitable for the ordered case. For example, there is positive probability on $d_i = d_{i'}$ with y_i and $y_{i'}$ the largest and smallest observation, and observations in between these two extremes allocated to different groups. So, at this point we simply study $p(d|y)$ as a classification probability model and adjust it to eliminate groupings that do not preserve the ordering in the data.

Hence, we first order the y, so that y_1 is the smallest observation and y_n is the largest observation. We then constrain the d so that the d_i's are non–decreasing. This ensures that any group contains only consecutive y_i's. For example, group 1 would contain a number of the smallest observations; group 2 would contain a number of the next smallest observations; while group k would contain a number of the largest observations. Thus, for any trio of $(y_{i_1} < y_{i_2} < y_{i_3})$, if y_{i_1} and y_{i_3} are in the same group, then so is y_{i_2}. It follows then that our $p^*(d|y)$, with the ordered y's, is given by

$p^*(d|y) \propto p(d|y)\,\mathbf{1}(d_1 \leq \cdots \leq d_n)$. As we will see, this constrain reduces drastically the cardinality of the support when compared with the support of (1). Once this approach has been undertaken we then show how to sample from $p^*(d|y)$ in order to compute the classification probabilities, and obviously relevant features as the mode.

In Section 2 we derive and explain our probability model for classification. Section 3 then describes an MCMC algorithm for sampling from this probability model, since for large n the number of possible classes is prohibitively large to compute directly. Section 4 then presents numerical illustrations based on a toy example of 10 data points, whereby all probabilities can be computed. We also well evaluate the well known and widely studied galaxy data set. Finally, Section 5 provides a discussion.

2. The classification probability model. Given the outline in the Introduction, our first task is to compute $p(d|y)$ based on the MDP model. First let us notice that

$$p(d, y|\mu, \lambda, w) = \prod_{i=1}^{n} w_{d_i}\, \mathrm{N}(y_i; \mu_{d_i}, \lambda_{d_i}^{-1}),$$

where $\mu := (\mu_i)_{i=1}^{\infty}$ and $\lambda := (\lambda_i)_{i=1}^{\infty}$. Hence

$$\begin{aligned} p(d, y) \;&= \mathrm{E}\left\{\textstyle\prod_{i=1}^{n} v_{d_i} \prod_{l<d_i}(1-v_l)\right\} \;\prod_{j=1}^{\infty} \int \prod_{d_i=j} \mathrm{N}(y_i; \mu, \lambda^{-1})\, \pi(\mathrm{d}\mu, \mathrm{d}\lambda) \\ &= \textstyle\prod_{j=1}^{\infty} \left\{\alpha \int v^{n_j}(1-v)^{m_j+\alpha-1}\,\mathrm{d}v\right\} \left\{\int \prod_{d_i=j} \mathrm{N}(y_i; \mu, \lambda^{-1})\, \pi(\mathrm{d}\mu, \mathrm{d}\lambda)\right\}. \end{aligned}$$

Here, n_j is defined as before and $m_j = \sum_{i=1}^{n} \mathbf{1}(d_i > j)$. The first term in the product is given by

$$\frac{\alpha\,\Gamma(1+n_j)\Gamma(\alpha+m_j)}{\Gamma(1+\alpha+n_j+m_j)}$$

and the second term is easily found to be given by

$$\frac{\Gamma(a+n_j/2)\, b^a\, \sqrt{c}}{\left\{b + S_j^2/2\right\}^{a+n_j/2}\, \sqrt{c+n_j}\, \Gamma(a)},$$

where

$$S_j^2 = \sum_{d_i=j} y_i^2 - \frac{n_j \bar{y}_j^2}{1 + c/n_j}$$

and

$$\bar{y}_j = n_j^{-1} \sum_{d_i=j} y_i.$$

Hence,

$$p(d|y) \propto \prod_{j=1}^{\infty} \frac{\alpha\,\Gamma(1+n_j)\Gamma(\alpha+m_j)}{\Gamma(1+\alpha+n_j+m_j)} \frac{\Gamma(a+n_j/2)\,b^a\,\sqrt{c}}{\left\{b+S_j^2/2\right\}^{a+n_j/2}\,\sqrt{c+n_j}\,\Gamma(a)}.$$

This then is the classification probability based on the MDP model.

At this point, we simply focus on $p(d|y)$ and assess it as a probability model for classification. So, without loss of generality, we take the y's to be ordered, with y_1 being the smallest observation and y_n being the largest. Then, for reasons given in the Introduction, we would now for ordered, or unsupervised, classification purposes, only wish to consider the (d_i) to be non–decreasing. Hence, we obtain

$$p^*(d|y) \propto \mathbf{1}(d_1 \leq \cdots \leq d_n)\, p(d|y).$$

We also impose the constraint that if there are k distinct d_i's then $d_n = k$, namely there are no gaps.

Our observation now is that given the current parameter and data values d is completely determined by $(k, n_1, \ldots, n_k)$ whereby k is the number of distinct d_i's and n_j is the number of the d_i's equal to j. Hence,

$$p(k, n_1, \ldots, n_k) = \mathsf{K} \prod_{j=1}^{k} \frac{\alpha\,\Gamma(1+n_j)\Gamma(\alpha+m_j)}{\Gamma(1+\alpha+n_j+m_j)} \frac{\Gamma(a+n_j/2)\,b^a\,\sqrt{c}}{\left\{b+S_j^2/2\right\}^{a+n_j/2}\,\sqrt{c+n_j}\Gamma(a)} \tag{2}$$

where K denotes the corresponding normalizing constant, $m_j = n - n_1 - \cdots - n_j$ and

$$S_j^2 = \sum_{i=n_{j-1}^*+1}^{n_j^*} y_i^2 - \frac{n_j\,\bar{y}_j^2}{1+c/n_j}$$

and

$$\bar{y}_j = n_j^{-1} \sum_{i=n_{j-1}^*+1}^{n_j^*} y_i,$$

with $n_j^* = n_1 + \cdots + n_j$ and $n_0^* = 0$. Note that, for a given sample size n, the support of this probability is substantially reduced and is in bijection with the set of compositions $\mathcal{C}_n$, rather than

on the number of partitions of a set with n elements as in the MDP or other exchangeable partition probability settings encountered in the Bayesian nonparametric literature. This, for instance, would exclude group configurations of the sort $\mathcal{G}^0 := \{(y_1, y_2, y_8, y_4), (y_3, y_5, y_6, y_7, y_9, y_{10})\}$. Although much smaller than the set of all partitions of $[n]$, i.e. $\mathcal{P}_{[n]} = \bigcup_{k=1}^{n} \mathcal{P}_{[n]}^{k}$ (with $B_n = \#\mathcal{P}_{[n]}$), the cardinality of the set of compositions is 2^{n-1} and therefore the need of MCMC methods to approximate (2) is still evident. However, the fact that the support of such probabilities is restricted to the ordering of the data makes it more acceptable to employ split and merge samplers. In the next section we present an algorithm to perform such a task.

From an unsupervised perspective, this probability model for classification is a suitable and, we believe, useful adaptation of the probability model for classification based on the MDP model. It can be seen to depend fundamentally on the sample variances of the observations in the same group, so the lower the sample variances, the higher the probability assigned to the corresponding partition. The rule of having $k = n$ groups is countered by the probability being a product of k terms. For our particular choice of f and π our classification probability depends on the parameters (α, a, b, c), which have a similar interpretation as if we were using a MDP model for the data. So, for example, if α is big, which implies a large number of groups in the MDP model, its role can be seen explicitly in $p(k, n_1, \ldots, n_k)$, since we would have the term α^k and so encourages large k. See the discussion in Section 4.

We also note that attempts have been made to emphasize configurations, i.e. $(k, n_1, \ldots, n_k)$, that are suitable for classification purposes by using alternative nonparametric mixing prior distributions to the Dirichlet process and which have extra parameters to modulate the weight across the different data partitioning (Lijoi *et al.*, 2007). However, even in those more general Bayesian nonparametric settings positive mass is still being put to overlapping configurations. Our approach, in light of this, is remarkably obvious in that we put zero weight on all but those configurations preserving the order of the data.

Here we also mention the problem of what happens if a new piece of data arrives. Our approach is not to assume a classification for the existing data has been set and to decide into which group, possibly a new one, the extra piece of data should be put; but rather we merely recompute $p(k, n_1, \ldots, n_k)$ with all the data, including the additional piece. We do not see any other approach as being relevant here.

We will compare our approach with a routine in the package R, a hierarchical clustering routine based on local sums of squares that in principle is not unlike the idea of working with sample variances. The routine is labeled `hclust` in R and is based on an original algorithm appearing in Ward (1963). We use this method for comparison primarily as it is in R.

3. Sampling the model. The basic idea for sampling from $p(k, n_1, \ldots, n_k)$ will be a split–merge MCMC algorithm. So at each iteration one of two types of move will be proposed: a split, whereby a group of size bigger than one is divided into two groups so k is increased by one; and a merge, whereby two groups are combined into one group so k is decreased by one. The idea for sampling from $p(k, n_1, \ldots, n_k)$ can be seen as a reversible jump MCMC algorithm, and for ease of exposition we will describe the algorithm using latent variables and the specification of a joint density for a configuration conditional on a k: so let $n^{(j)}$ for $j = 1, \ldots, n$ be a clustering for j groups, and consider

$$p(k, n^{(1)}, \ldots, n^{(n)}) = p(k, n_1, \ldots, n_k) \prod_{j=k+1}^{n} p(n^{(j)}|n^{(j-1)}) \prod_{j=1}^{k-1} p(n^{(j)}|n^{(j+1)}),$$

which is based on ideas provided, for example, in Godsill (2001).

The key here is that the marginal density for $(k, n_1, \ldots, n_k)$ is unchanged. Now given a k and $n^{(k)}$ we propose a move to $k+1$ with probability $1/2$ and to $k-1$ with probability $1/2$ (with obvious modifications if $k = 1$ or $k = n$). We need to therefore sample $n^{(k+1)}$ from $p(n^{(k+1)}|k, n^{(k)})$ and $n^{(k-1)}$ from $p(n^{(k-1)}|k, n^{(k)})$. The former is achieved by finding an existing group with size bigger than 1, and then we split this group into 2. If uniform distributions are used for both operations

then

$$p(n^{(k+1)}|k, n^{(k)}) = \frac{1}{n_g^{(k)}(n_s^{(k)} - 1)},$$

where $n_g^{(k)}$ is the number of groups of size > 1, from $(n_1, \ldots, n_k)$, and $n_s^{(k)}$ is the size of this chosen group. The latter is obtained by merging two neighboring groups and so with a uniform distribution we have

$$p(n^{(k-1)}|k, n^{(k)}) = 1/(k-1).$$

Therefore, the sampler carries out each step through a Metropolis-Hastings scheme. When at state $x^{(k)} = (k, n_1, \ldots, n_k)$ the acceptance probability to move to state $x^{(k+1)} = (k+1, n_1, \ldots, n_{k+1})$ is

$$\alpha(x^{(k)}, x^{(k+1)}) = \min\left\{1, \frac{1-q}{q}\frac{p(x^{(k+1)})}{p(x^{(k)})}\frac{n_g^{(k)}(n_s^{(k)} - 1)}{k}\right\}, \tag{3}$$

where, e.g., $q = 1/2$. On the other hand, if instead two groups, $(n_{s1}^{(k)}, n_{s2}^{(k)})$, in $x^{(k)}$ are selected and we attempt to merge them, then the acceptance probability for this move is

$$\alpha(x^{(k)}, x^{(k-1)}) = \min\left\{1, \frac{q}{1-q}\frac{p(x^{(k-1)})}{p(x^{(k)})}\frac{k-1}{(n_{s1}^{(k)} + n_{s2}^{(k)} - 1)n_g^{(k-1)}}\right\}, \tag{4}$$

where $n_g^{(k-1)}$ denotes the cardinality of the set containing all groups with more than one observation in $x^{(k-1)}$.

To improve the algorithm we then can shuffle the $n^{(k)}$ by selecting adjacent groups, (n_{s1}, n_{s2}) and attempting to change them into (n_{s1}^*, n_{s2}^*) in such a way that both n_{s1}^* and $n_{s2}^* \geq 1$. The shuffle is based on the idea of putting the two groups together and then uniformly splitting the combined group into two. The acceptance probability is then given by

$$\alpha(x, x^*) = \min\left\{1, \frac{p(x^*)}{p(x)}\frac{(n_{s1}^* + n_{s2}^* - 1)}{(n_{s1} + n_{s2} - 1)}\right\}. \tag{5}$$

These acceptance probabilities all follow from the expression $p(k, n^{(1)}, \ldots, n^{(n)})$ and the cancelations which occur when evaluating the ratios of neighboring k.

The sampling algorithm can then be summarized with the following three steps:

1. For a set of fixed parameters (α, a, b, c), initialize the chain with a configuration $(k, n_1, \ldots, n_k)$.

2. Select a type of move, either to split or merge and accept it with its corresponding acceptance probability.

3. Perform a shuffle and accept it with probability given by (5).

The algorithm is effectively a joint Metropolis–Hastings and Gibbs algorithm, rather than a reversible jump MCMC algorithm, and for connections between the two see Godsill (2001). However, notice that under this perspective the dimension of the state space is fixed and so no special considerations arise on this issue. The only necessary consideration is that if $p(n^{(k)}|n^{(k-1)}) > 0$ then $p(n^{(k-1)}|n^{(k)}) > 0$, and vice versa. Neither, if one is even needed, have we had to worry about a Jacobian, since we are not basing the moves on transformations of variables between different dimensions. We believe it is more explicit to understand transdimensional samplers from this perspective.

3.1. An alternative sampler. Here we consider a more general idea for sampling, based on the notion of a joint density

$$p(k, n^{(1)}, \ldots, n^{(n)}) = p(k, n^{(k)})\, p(n^{(2)}, \ldots, n^{(k-1)}, n^{(k+1)}, \ldots, n^{(n-1)} |\, k, n^{(k)}). \tag{6}$$

It is easy to see how the reversible jump MCMC arises from this model, however we can seek alternative and more general strategies, and one such is based on the idea of

$$p(n^{(2)}, \ldots, n^{(k-1)}, n^{(k+1)}, \ldots, n^{(n-1)}|n^{(k)}) = p(n^{(2)}) \prod_{j=3}^{k-1} p(n^{(j)}|n^{(j-1)}) \prod_{j=k+1}^{n-1} p(n^{(j)}|n^{(j-1)}),$$

where $p(n^{(k)}|n^{(k-1)})$ is the probability density for a split move described earlier, and $p(n^{(2)})$ is the correct density for $n^{(2)}$ given $k = 2$, which is easy to sample since $n^{(2)}$ can be represented by a single number between 1 and $n - 1$. Then it is easy to see that a move from $x^{(k)}$ to $x^{(k')}$, with $k' \in \{k - 1, k + 1\}$ can be achieved by first sampling $x^{(k+1)}$ from $p(n^{(k+1)}|n^{(k)})$ and $x^{(k-1)}$ from

the density $p(n^{(2)}) \prod_{j=3}^{k-1} p(n^{(j)}|n^{(j-1)})$, which is done by sampling $n^{(2)}$, then $n^{(3)}$, and so on, up to $n^{(k-1)}$. If $p(n^{(k)}|n^{(k-1)}) = 0$ then the proposed move is rejected and $(k, n^{(k)})$ is kept. On the other hand, if $p(n^{(k)}|n^{(k-1)}) > 0$ then a move to $k+1$, proposed with probability $1/2$ is accepted with probability

$$\min\left\{1, \frac{p(k+1, n^{(k+1)})\, p(n^{(k)}|n^{(k-1)})}{p(k, n^{(k)})\, p(n^{(k+1)}|n^{(k)})}\right\},$$

or else a move to $k-1$ is proposed and is accepted with probability

$$\min\left\{1, \frac{p(k-1, n^{(k-1)})\, p(n^{(k)}|n^{(k-1)})}{p(k, n^{(k)})\, p(n^{(k-1)}|n^{(k-2)})}\right\}.$$

While in this particular case we do not claim an improvement using this algorithm, our point is that there are alternatives to be considered not necessarily falling in the reversible jump MCMC methodology, and in particular formulated by the notion of a joint density of the kind of (6).

4. Numerical illustrations. In order to underline the kind of results that can be obtained by our approach, we first consider a small data set; small enough ($n = 10$) so that we can provide exact computations of probabilities for all $(k, n_1, \ldots, n_k)$. We then illustrate our approach with a real data set; the galaxy data set.

4.1. Small data set. Suppose the set of ordered observations is y = (-1.522, -1.292, -0.856, -0.104, 2.388, 3.080, 3.313, 3.415, 3.922, 4.194), a histogram of the data is shown in Figure 1, from this it is evident that 2 groups are the most likely option. Table 1 gives the probabilities of having k groups using the MDP approach, posterior probability (1), and our proposed approach, probability (2). In both cases we use the prior specification of parameters as $\alpha = a = b = 1$ and $c = 0.1$. For the MDP model we computed the exact probabilities for each of the $B_{10} = 115975$ possible partitions of y, and then used them to obtain the exact posterior probabilities for each $k \in \{1, \ldots, 10\}$. The highest probability in this case is allocated to $k = 3$. Further inspection among the partitions indicates that the highest posterior probability, of 0.332, corresponds to the classification involving two groups; $\{[y_1, y_2, y_3, y_4], [y_5, \ldots, y_{10}]\}$, which corresponds to an integer partition $(n_1, n_2) = (4, 6)$. However, it

is worth emphasising that group configurations such as the $\mathcal{G}^0$ mentioned in Section 2 still receive positive posterior probability under this approach.

On the other hand, the exact probabilities (2), $p^*(k)$, computed over all the 512 possible configurations, assign the highest probability to $k = 2$. As with the MDP case, the classification with the highest probability corresponds to $\{[y_1, y_2, y_3, y_4], [y_5, \ldots, y_{10}]\}$; but in this case with the considerably higher probability of 0.833. Clearly, considering the order of the y limits the support considerably, from set partitions to integer compositions, withdrawing all inadequate partitions for unsupervised classification purposes and hence leading to an improved estimator for the number of groups.

The last four columns in Table 1 show the estimates of $p^*(k)$ based on the MCMC schemes described in Section 3, with $10,000$ and $100,000$ iterations following a burn in period of $1,000$ and $10,000$ iterations respectively. The results match the exact probabilities and from these it is evident that both algorithmic schemes are valid, although the first converges faster than the second.

The prior distribution on the different number of groups, in a sample of size n, corresponding to the Dirichlet process is given by

$$P[K_n = k] = \frac{\alpha^k}{(\alpha)_n} |s_{n,k}|, \quad k = 1, \ldots, n,$$

where $|s_{n,k}|$ stands for the sign-less Stirling number of the first kind. This distribution is highly peaked, (Lijoi *et al.*, 2007), with the parameter α modulating the location of the mode. Hence, when using a MDP model a careful choice of the parameter α must be made. In some situations, it is even convenient to assign a hyper–prior distribution to such a parameter, (see for instance Escobar and West, 1995).

Table 2 reports the exact probabilities for both the MDP model and our approach, with the parameter α selected in such a way that the $E[K_{10}] = 2.1$ ($\alpha = 0.5$) and $E[K_{10}] = 5.8$ ($\alpha = 5$). The first being the correct number and other being far from it. As expected, for the MDP model with $\alpha = 0.5$ the mode of the posterior distribution is located at $k = 2$, improving the result encountered

in Table 1. In contrast, the mode for the case when $\alpha = 5$ is far from $k = 2$. For our approach we obtain a mode at $k = 2$ in both cases.

Our results are in agreement with the hierarchical agglomerative clustering, using Ward's (1963) approach, which reduces the number of groups from k to $k-1$ by minimizing the local sum of squares; see Figure 2. It is obvious from this that two groups are by far the preferred choice.

4.2. Galaxy data set. Here we consider the galaxy data set, first studied in Roeder (1990). It is widely used in the literature to illustrate methodology for mixture modeling and cluster analysis. In this case the sample size is $n = 82$ and so we would need to compute 2^{81} probabilities to obtain all the possible configurations.

Therefore, we will use the first MCMC algorithm proposed in Section 3 to obtain the probabilities. We undertake this approach, with the same parameter specifications as in the small data set example and perform 10000 iterations after a burn–in period of 1000 iterations. The MCMC estimates result in $p^*(k = 3) = 0.997$ and $p^*(k = 4) = 0.003$, with the highest probability of 0.677 on the configuration $(n_1, n_2, n_3) = (7, 72, 3)$. The same results are obtained with the second scheme of Section 3, but with a higher number of simulations required. It is worth remembering that most Bayesian nonparametric approaches, such as the MDP or those based on more general random probability measures, typically favour between 5 and 6 components. See, for example, Escobar and West (1995) and Lijoi *et al.* (2005). Similar results are achieved in the finite mixture setting, as found in Richardson and Green (1997) where a mode is put in $k = 6$ with posterior probability 0.199 followed by $k = 5$ with 0.182, when using an uniform prior for k. All of these approaches seem to be overestimating the number of groups, as noted from results reported in McGrory and Titterington (2007).

5. Discussion. We have proposed a probability for classification based on the mixture of Dirichlet process model. The main idea is to restrict the support of all possible data grouping to those

preserving the order, which for unsupervised classification purposes appears to be more acceptable. This led us to consider probabilities with a much smaller support, namely those in bijection with the set of compositions of the integer n instead of the typical set of partitions of a set with n elements. An MCMC algorithm, that overcomes some of the complications typically found with the reversible jump approach, is employed to sample the relevant classification probabilities.

Here we have focused on MDP based on Gaussian distributions. Similar results might be obtained for other more general non-parametric priors, e.g. two parameter Poisson-Dirichlet process, and other symmetric location-scale invariant kernels. We will devote future research to investigate this.

Acknowledgements. The first author gratefully acknowledges the Mexican Mathematical Society and the Sofia Kovalevskaia Fund, and the second author gratefully acknowledges CONACYT for Grant No. J50160-F, for allowing them to travel to the UK, where the work was completed during a visit to the University of Kent. The authors gratefully acknowledge the comments of 3 referees which have improved the paper.

References.

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions.* Hoboken, NJ: Wiley.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, **89**, 268–277.

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

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, **10**, :230–248.

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

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Lijoi, A. and Mena, R. H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse Gaussian priors. *Journal of the American Statistical Association*, **100**, 1278–1291.

Lijoi, A. and Mena, R. H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B*, **69**, 715–740.

Lijoi, A. and Prünster, I. (2009). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, (Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G. Eds.), Cambridge University Press, in press.

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

McGrory, C.A. and Titterington, D.M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis* **51**, 5352–5367.

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

Quintana, F.A. and Iglesias, P.L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society, Series B*, **65**, 557–574.

Roeder, K. (1990). Density estimation with confidence sets exemplified by super–clusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617–624.

Sethuraman, J. (1990). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Ward, J.H. (1990). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

k	MDP	Exact $p^*(k)$	M1 (A) $\widehat{p}^*(k)$	M1 (B) $\widehat{p}^*(k)$	M2 (A) $\widehat{p}^*(k)$	M2 (B) $\widehat{p}^*(k)$
1	0.00619	0.04535	0.04760	0.04709	0.04450	0.04794
2	0.37634	0.88622	0.88480	0.88375	0.84770	0.88376
3	0.39729	0.06597	0.06710	0.06652	0.10130	0.06482
4	0.17298	0.00240	0.00050	0.00250	0.00650	0.00348
5	0.04088	0.00006	–	0.00011	–	–
6	0.00578	1.00 E−6	–	0.00003	–	–
7	0.00051	1.31 E−8	–	–	–	–
8	0.00003	1.22 E−10	–	–	–	–
9	8.38 E−7	7.44 E−13	–	–	–	–
10	1.12 E−8	2.26 E−14	–	–	–	–

Table 1: Probabilities on the different number of groups for the small data set example. The MDP results correspond to exact posterior probabilities. The probabilities $p^*(k)$ and $\widehat{p}^*(k)$ for the classification model correspond to the exact and MCMC estimates, respectively. The columns labeled M1 and M2 refer to the two sampling schemes described in Section 3 with (A) 10000 iterations after a 1000 burn in period and (B) 100000 iterations after a 10000 burn in period.

	$\alpha = 0.5$		$\alpha = 5$	
k	MDP	$p^*(k)$	MDP	$p^*(k)$
1	0.019469	0.08342	0.000071	0.01292
2	0.591630	0.87837	0.021504	0.80256
3	0.312288	0.03742	0.113509	0.16689
4	0.067986	0.00078	0.247113	0.01652
5	0.008033	0.00001	0.291972	0.00105
6	0.000568	1.06 E−7	0.206592	0.00005
7	0.000025	7.84 E−10	0.090763	1.64 E−6
8	6.74 E−7	4.10 E−12	0.024486	4.10 E−8
9	1.03 E−8	1.38 E−14	0.003740	7.10 E−10
10+	6.85 E−11	2.29 E−17	0.000249	6.26 E−12

Table 2: Probabilities on the different number of groups for the small data set example. All the results, for the posterior probabilities corresponding to the MDP model and for the classification model correspond to the exact probabilities.

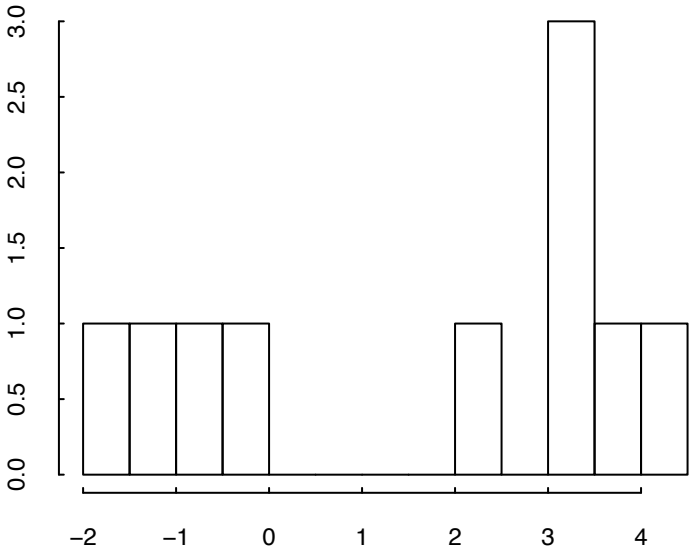

Figure 1: Histogram for the small data set.

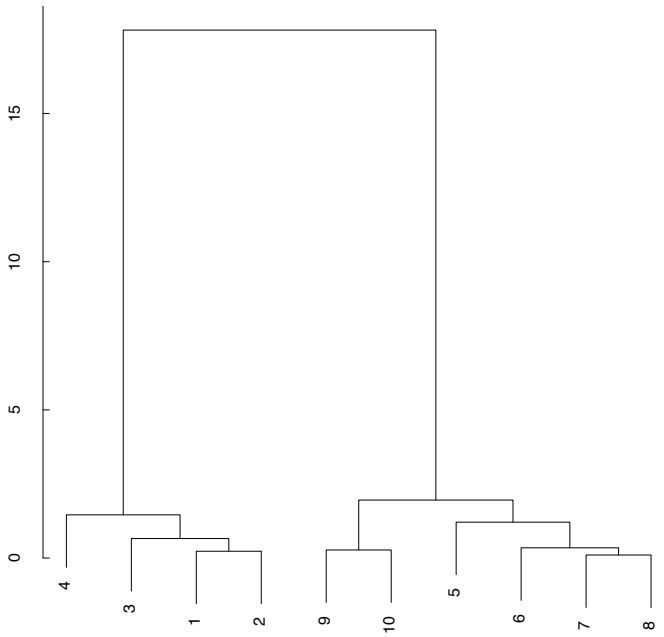

Figure 2: Dendogram for the small data set example..