A New Bayesian Nonparametric Mixture Model

Ruth Fuentes-Garcia, Ramses H Mena, Stephen G. Walker

A New Bayesian Nonparametric Mixture Model

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2008-0168.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	03-Sep-2009
Complete List of Authors:	Fuentes-Garcia, Ruth; UNAM Mena, Ramses; UNAM, IIMAS Walker, Stephen G.; IMSAS
Keywords:	Bayesian model, Mixture model, Geometric distribution, Gibbs sampler
Abstract:	We propose a new mixture model for Bayesian nonparametric inference. Rather than considering extensions from current approaches, such as the mixture of Dirichlet process model, we end up shrinking it, by making the weights less complex. We demonstrate the model and offer an explanation for the performance.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
newbnp_rev3.zip	

A New Bayesian Nonparametric Mixture Model

R. Fuentes-García*, R.H. Mena** and S.G. Walker***

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

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

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

Abstract

We propose a new mixture model for Bayesian nonparametric inference. Rather than considering extensions from current approaches, such as the mixture of Dirichlet process model, we end up shrinking it, by making the weights less complex. We demonstrate the model and discuss its performance.

Kewords: Bayesian model; Geometric weight; Gibbs sampler; Mixture model.

1. Introduction. There are two approaches to Bayesian nonparametric density estimation; the first is based on mixture models where the random density function $f(y)$ is constructed via

$$f(y) = \int K(y;\theta)\,\mathrm{d}P(\theta),$$

where $K(y;\theta)$ is a density function for each θ, P is a random distribution function, which is typically taken to be discrete, *e.g.* of the type

$$P(\mathrm{d}\theta) = \sum_{l=1}^{\infty} \tilde{w}_l\,\delta_{\theta_l}(\mathrm{d}\theta) \tag{1}$$

with a prior distribution assigned to $(\tilde{w}_l, \theta_l)_{l=1}^{\infty}$ and where δ_θ denotes the measure with mass 1 at the location θ. For example, the well known and widely used Dirichlet process results when $\{\theta_l\}$ are independent and identically distributed (iid) from some distribution G, the prior guess at the shape of P, $\tilde{w}_1 = v_1$, and for $l > 1$, $\tilde{w}_l = v_l \prod_{k<l}(1-v_k)$ with the $\{v_l\}$ being iid from Beta$(1, c)$, for some $c > 0$. See Sethuraman (1994) for the latter representation and Lo (1984) for definition and properties of mixture of Dirichlet process models. Some sampling techniques to infer from Bayesian nonparametric mixture models can be found in Escobar (1988, 1994), Escobar and West (1995), MacEachern (1994), MacEachern and Müller(1998), Neal (2000), Papaspiliopoulos and Roberts (2008), and Walker (2007).

[1]E-mail for correspondence: S.G.Walker@kent.ac.uk

The other approach to Bayesian density estimation is based on a finite mixture model with the number of mixtures being N and a prior distribution assigned to N. So

$$f(y|N) = \sum_{l=1}^{N} w_{l,N}\, K(y; \theta_{l,N}).$$

See, for example, Richardson and Green (1997). Algorithms for estimating this latter model are based on reversible jump Markov chain Monte Carlo (Green, 1995) or using birth and death Markov chain Monte Carlo; see Stephens (2000).

The algorithm of Walker (2007) for estimating the mixture of Dirichlet process (MDP) model, which is also applicable to more general stick–breaking processes, *e.g.* when the $v_l \sim \text{Beta}(a_l, b_l)$, starts by considering the density

$$f(y|u) = |A_u|^{-1} \sum_{l=1}^{\infty} \mathbf{1}(u < \tilde{w}_l)\, K(y; \theta_l)$$

where A_u denotes a random set defined as $A_u := \{j : \tilde{w}_j > u\}$ with u a uniform random variable. Given u, this random set is clearly finite and we denote its cardinality by $|A_u|$. It is worth emphasizing that A_u is a random finite subset of the set of positive integers.

We could consider a more general idea by constructing the following random density

$$f(y \mid A) = |A|^{-1} \sum_{l \in A} K(y; \theta_l) \tag{2}$$

where A denotes a different, perhaps more general, random set.

This is similar to the Richardson and Green (1997) model but with differences. First, notice that if we assume model (2) for each observation y_i, then there will be a random set A_i for each of them whereas in the Richardson and Green (1997) model the N suffices for all observations. This is one of the reasons for having complex weights and parameters specifications, $(w_{l,N}, \theta_{l,N})$'s, in their approach, *i.e.* to make a richer model. On the other hand for the MDP model it is sufficient to have uniform weights given A and a single sequence $\{\theta_l\}$. Second, the A used in the mixture of Dirichlet process model is not a consecutive sequence of integers from 1 to N, as it is with the Richardson and Green (1997) model. It seems clear to us that the use of consecutive sequences is likely to be more efficient when estimating the model using Markov chain Monte Carlo algorithms. In fact, there is no point in having the A to have gaps; the real question is why would one wish A to have gaps?

The idea of this paper is to suggest the model whereby $A_i := \{1, \ldots, N_i\}$, so

$$f(y|N) = N^{-1} \sum_{l=1}^{N} K(y; \theta_l)$$

where N is random, but with the same distribution for each observation; say $\mathrm{P}(N) = q_N$ and a prior distribution is assigned to $\{q_N\}$.

Our observation is on the choice of the $\{q_N\}$, which as we will see determines the structure of the weights $\{w_l\}$ corresponding to a random distribution characterized by (1). In principle, any distribution supported on the set of positive integers could be used. Particularly, we look for a choice of $\{q_N\}$ that lead us to a manageable structure for the weights and also to simple conditionals in the corresponding Gibbs sampler algorithm. A suitable choice for these purposes turns out to be a Neg-Bin$(2, \lambda)$ that, together with a beta prior distribution for λ, results in a well defined random distribution of the type (1). Our results based on this choice are remarkable considering the simplicity of the model and algorithm. In Section 2 we describe the model, resulting from our choice of random set A, in more detail and describe some properties. Section 3 details the Gibbs sampler for estimating the model and Section 4 is devoted to illustrations that aid to understand the potential and contribution of our approach. A discussion on our findings is presented in Section 5.

2. Properties of the model. If we write out the model by marginalizing over N then we have

$$f(y) = \sum_{N=1}^{\infty} \frac{1}{N} \sum_{l=1}^{N} K(y; \theta_l)\, q_N \tag{3}$$

which can be written as

$$f(y) = \sum_{l=1}^{\infty} w_l\, K(y; \theta_l) = \int K(y; \theta)\, \mathrm{d}P(\theta),$$

where

$$P(\mathrm{d}\theta) = \sum_{l=1}^{\infty} w_l\, \delta_{\theta_l}(\mathrm{d}\theta)$$

and the weights $\{w_l\}$ are given by

$$w_l = \sum_{N=l}^{\infty} q_N / N. \tag{4}$$

These weights clearly add up to one and, in contrast to the weights corresponding to the Dirichlet process, are always decreasing. Although, this model can be seen as a nonparametric mixture model with the above random distribution, we can write it in the hierarchical form (2) with the random set A chosen as $\{1, \ldots, N\}$ rather than being of the non–consecutive type as it is with the Dirichlet process model, as seen in Walker (2007).

Indeed, as we mentioned in the introduction, the $\{q_N\}$ can be given any arbitrary distribution supported on the set of positive integers; immediate choices could fall in the Poisson, negative

binomial or geometric families. However, note that an arbitrary choice for q_N does not necessarily leads to a simple analytic structure for the weights. For example, if

$$q_N = \binom{N+r-2}{r-1}\lambda^r(1-\lambda)^{N-1}, \quad N = 1, 2, \ldots$$

namely a negative binomial distribution (Neg-Bin(r, λ)) supported on the set of positive integers, then the corresponding weights take the form

$$w_l = \frac{1}{l}\binom{l+r-2}{r-1}\lambda^r(1-\lambda)^{l-1}\,{}_2F_1(1, l+r-1; l+1; \lambda), \tag{5}$$

where ${}_2F_1(a, b; c; \lambda)$ denotes the Gauss hypergeometric function.

In Section 3 we will detail the Markov chain Monte Carlo algorithm for estimating the model based on geometric weights. First, let us look at the conditional distribution for each N_i that motivates this choice for q_N. If we assume model (3) for a set of observations, $\{y_i\}_{i=1}^n$, and introduce a latent variable d_i that, given N_i, indicates from which component y_i comes from, then

$$\mathrm{P}(d_i = l|N_i) = N_i^{-1}\mathbf{1}\,(l \in \{1, \ldots, N_i\}).$$

Now, since we have $\mathrm{P}(N_i = l) = q_l$, then

$$\mathrm{P}(N_i = N|d_i) \propto \frac{q_N}{N}\ \mathbf{1}(N \geq d_i).$$

Hence, it is convenient that the sequence of probabilities $\{q_N/N\}$ take a form from which is relatively easy to sample truncated versions. As we will see, a special case of the negative binomial distribution, specifically a Neg-Bin$(2, \lambda)$, leads to a truncated geometric distribution for $P(N_i \mid d_i)$, which is clearly simple to simulate from. Furthermore, a beta prior distribution could be taken as a conjugate choice easing even more the implementation of the MCMC algorithm. It is worth mentioning that at the outset our plan was to start with this latter choice for q_N and then move to other, perhaps more general, choices but as we will see this appears unnecessary at least for mixture modelling aiming at density estimation.

3. Simulation algorithm. In order to see how to construct a Gibbs sampler for this model and our choice of random sets A, we write it in hierarchical form for a general choice of q_N

$$f(y_i|d_i, N_i) = K(y_i; \theta_{d_i})$$

$$\mathrm{P}(d_i = l|N_i) = N_i^{-1}\,\mathbf{1}(l \in \{1, \ldots, N_i\})$$

$$\mathrm{P}(N_i = N) = q_N,$$

where the $\{\theta_l\}$ are assumed to be i.i.d. from a distribution with density g and a prior is assigned to the parameters of q_n, namely $\pi(q)$.

The full conditional for θ_j is then given by

$$f(\theta_j|\cdots) \propto g(\theta_j) \prod_{d_i=j} K(y_i;\theta_j)$$

which in particular is easy to sample when $K(y;\theta)$ and $g(\theta)$ form a conjugate pair. The full conditional for d_i is given by

$$\mathrm{P}(d_i = l|\cdots) \propto K(y_i;\theta_l)\, \mathbf{1}(l \in \{1,\ldots,N_i\})$$

which being a discrete distribution with finite support it is easy to sample. The full conditional for N_i has already been considered in Section 2, repeated here, as

$$\mathrm{P}(N_i = N|\cdots) \propto \frac{q_N}{N}\, \mathbf{1}(N \geq d_i).$$

Finally, the full conditional for q is given by

$$\pi(q|\cdots) \propto \left\{\prod_{i=1}^{n} q_{N_i}\right\} \pi(q).$$

As we mentioned in Section 2, in order to simplify the sampling of the full conditional for N_i an easy form for q_N/N is required. In particular, this is attained by setting q_N to take Neg-Bin$(2,\lambda)$ distribution, *i.e.* with density

$$q_N = N\,\lambda^2(1-\lambda)^{N-1}$$

which, following (4), results in weights given as

$$w_l = \lambda(1-\lambda)^{l-1}.$$

That is, the decreasing weights take a geometric distribution. We then assign a beta hyper–prior distribution for λ.

As we will corroborate in the following section this model, although at first sight simplistic and not flexible enough, will prove to perform well for density estimation purposes. Before going into the numerical illustrations let us adjust and complete the details of the Gibbs sampler for sampling the full conditional distributions of N_i and λ. Notice that assuming a Neg-Bin$(2,\lambda)$ for q_N results in

$$\mathrm{P}(N_i = N|\cdots) \propto (1-\lambda)^{N-1}\, \mathbf{1}(N \geq d_i)$$

which is a truncated geometric distribution and is easy to sample. The full conditional for λ, assuming a Beta(a, b) prior, is given by

$$\pi(\lambda|\cdots) \propto \left\{\prod_{i=1}^{n} \lambda^2(1-\lambda)^{N_i-1}\right\} \lambda^{a-1}(1-\lambda)^{b-1}$$

which is also a beta distribution with parameters $a+2n$ and $b-n+\sum_{i=1}^{n} N_i$.

Note that with this choice of q_N we could see our approach as a Bayesian nonparametric mixture model with a mixing discrete random distribution, of the type (1), with geometric weights.

In fact, weights defined as above can be thought of as the expected value of the weights corresponding to Dirichlet process. That is, if we use the stick breaking representation of the Dirichlet process, *i.e.* with weights $\tilde{w}_l = v_l \prod_{k<l}(1-v_k)$ with the $\{v_l\}$ being iid from Beta$(1, c)$ and $c > 0$, then

$$\mathrm{E}[\tilde{w}_l] = \frac{1}{c+1}\left(\frac{c}{c+1}\right)^{l-1},$$

which is a simple re-parametrization of $\lambda\,(1-\lambda)^{l-1}$ when $\lambda = (c+1)^{-1}$. When implementing models based on the Dirichlet process, typically the assignation of a prior distribution on the total mass parameter, c, is needed to achieve good results. Therefore, our approach could be seen as the removal of a hierarchical level from the Dirichlet process model by replacing the random $\{v_l\}$ with their expected values.

4. Numerical Illustrations. In this section we consider 3 examples; 2 simulated data sets and one real data set; the well known galaxy data set.

4.1 *Location modeling for simulated data.* Here we generate 200 iid data points, $\{y_i\}_{i=1}^{200}$, coming from a minitial values for the $\{d_i\}_{i=1}^{200}$, chosen uniformly from the set $\{1, \ldots, 10\}$, the full conditional distributions are given by ixture of two normal distributions; N$(0, 1)$ and N$(6, 1)$, with corresponding weights 0.3 and 0.7. For the implementation of the method described in Section 3, let us assume that $K(y;\theta) = \mathrm{N}(y;\theta,1)$, $g(\theta) = \mathrm{N}(\theta;\mathsf{m},1/\mathsf{v})$, $q_N(N;\lambda) = N(1-\lambda)^{N-1}\lambda^2$ and $\lambda \sim \mathrm{Be}(\mathsf{a},\mathsf{b})$. Hence, for a given set of hyper–parameters $(\mathsf{m},\mathsf{v},\mathsf{a},\mathsf{b})$, starting configurations for the $\{N_i\}_{i=1}^{200}$ and the

$$\pi(\theta_j \mid \cdots) = \mathrm{N}\left(\theta_j \mid \frac{\mathsf{mv}+s_j}{\mathsf{v}+n_j}, \frac{1}{\mathsf{v}+n_j}\right),$$

where $n_j := \sum_{d_i=j} 1$ and $s_j := \sum_{d_i=j} y_i$,

$$\mathrm{P}(d_i = l \mid \cdots) = \frac{\mathrm{N}(y_i \mid \theta_l, 1)}{\sum_{k=1}^{N_i} \mathrm{N}(y_i \mid \theta_k, 1)}\,\mathbf{1}(l \in \{1, \ldots, N_i\}),$$

$$\mathrm{P}(N_i = j \mid \cdots) = \lambda\,(1-\lambda)^{j-1}\,\mathbf{1}(j \geq d_i)$$

and

$$\pi(\lambda \mid \cdots) = \mathrm{Be}\left(\lambda; \mathsf{a} + 2n, \mathsf{b} + \sum_{i=1}^{n} N_i - n\right).$$

Figure 1 shows the Monte Carlo density estimate,

$$\hat{f}(y) = \frac{1}{\mathrm{M}} \sum_{k=1}^{\mathrm{M}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{N_i^k} \sum_{l=1}^{N_i^k} \mathrm{N}(y \mid \theta_l^k, 1),$$

resulting from $M = 10,000$ iterations after a $2,000$ burn–in period. The choice of hyper–parameters is given by $(\mathsf{m}, \mathsf{v}, \mathsf{a}, \mathsf{b}) = (3, 1, 1, 1)$ and was obtained by inspecting the data, *i.e.* by preserving the original mean and variance of the model that generated the data. As in any other nonparametric mixture model, changing this values radically would lead to different estimations. However, this could be extended to a more complex model, possibly with further hierarchies as done in the following subsection. In this subsection we keep it simple in order to better understand the role of having an unlimited number of θ's.

Figure 2 shows an estimator of the number of θ's below and above a cutoff point, set at $y = 3$ where visually we could locate a separation between the 2 cluster locations. The plot could be seen as representing the number of θ's used to capture each mode in the mixture. Hence, the following interpretation follows: for the mode with higher probability (located around $y = 6$) we observe a tendency to represent it with a small number of θ's, namely the bar plot shows a mode in two θ's. Whereas for the mode with the smaller probability (located around $y = 0$) the tendency points towards a higher number, the bar plot shows a mode in four θ's. This effect can be explained by the fact that the weights, w_l's, are decreasing. That is, if a small weight needs to be increased, this is achieved by increasing the number of θ's used to represent the corresponding mode.

We believe that this assignation of θ's to represent a particular cluster location is reinforced in a more ordered fashion than the assignation founded when using non-consecutive sets, as in the Dirichlet process case. This argument is better illustrated in the following subsection.

4.2 *Location and scale modeling for simulated data.* Here we consider a more complicated data set that allows us to highlight the flexibility of our approach. We generate 240 data points coming from a mean-variance mixture of six normal distributions with weights $(0.17, 0.08, 0.125, 0.2, 0.125, 0.21)$ and mean-variance parameters given by $(-18, 2), (-5, 1), (0, 1), (6, 1), (14, 1)$ and $(23, 125)$. Similar to the previous section, for our modeling approach, we assume $K(y; \theta) = \mathrm{N}(m, 1/v)$, so $\theta := (m, v)$, and a conjugate prior distribution given by

$$g(\theta) = \mathrm{N}(m; \mu, \tau v^{-1}) \, \mathrm{Ga}(v; \alpha, \beta).$$

The only substantial difference for this example is that the posterior density is given by

$$\begin{aligned}\pi(m_j, v_j \mid \ldots) &= \mathrm{N}\left(m_j \mid \frac{\tau\, n_j\, \bar{y}_j + \mu}{\tau\, n_j + 1}; \frac{\tau}{v\,(\tau\, n_j + 1)}\right) \\ &\times \mathrm{Ga}\left(v_j \mid \frac{n_j}{2} + \alpha; \frac{n_j(\bar{y}_j - \mu)^2}{2(\tau\, n_j + 1)} + \frac{D_j}{2} + \beta\right)\end{aligned}$$

where $\bar{y}_j = s_j/n_j$, s_j and n_j are as before and $D_j = \sum_{d_i=j}(y_i - \bar{y}_j)^2$.

Figures 3 and 4 show the dynamics of the density estimator for the first 100 iterations based on our approach, here termed geometric, and on the MDP model respectively. From Figure 3 we note that the availability of an unlimited number of θ_j's to represent a particular cluster location always results in a improvement in subsequent iterations. Whereas in the MDP case, Figure 4, the algorithm might require several iterations to obtain a good candidate for the θ_j representing a particular location. This feature is better appreciated in the mode welling around -18, which can be thought as being far from the overall mean of the data. This can be also observed at the tails of the density estimators in Figure 3, where for the initial iterations a bigger mass, than that shown for the MDP, is allocated.

In fact this drawback of the MDP and other mixtures based on more general random distributions, has received considerable attention in the Bayesian nonparametric literature resulting in algorithms that aim to accelerate the identification of good candidates for the θ_j's identifying particular cluster locations. See for instance MacEachern (1998). It is worth emphasizing that, despite these efforts, this issue is not fully resolved.

Figure 5 shows the estimates for both, the Dirichlet process and our geometric approach, at a convergent stage. This figure also compares the true model that generated the observations, as we can see both approaches can be thought as being relatively satisfactory, however our approach appears to be closer to the true model.

It is then clear that the decreasing order of the weights results in a more ordered and faster convergence of the density estimation.

4.3 *Galaxy data set.* In this section we consider the galaxy data set; see *e.g.* Roeder (1990). This data set has been widely used as an example in mixture modeling and in particular as a benchmark when proposing or comparing new Bayesian nonparametric mixture models. See for instance Escobar and West (1995) and Lijoi *et. al.* (2005). Some discussion on the advantages and drawbacks of using mixtures of Dirichlet processes for density estimation can be found in Green and Richardson (2001).

The galaxy data are typically captured by relatively complex Bayesian nonparametric mixture

models, *e.g.* with Gaussian kernels and where the mixing parameters are both, mean and variance. These parameters are then modelled non-parametrically through random distributions where the corresponding mean measure needs further hierarchies in order to improve the density estimation. For simple choices of random distributions, as in the Dirichlet process case, the total mass parameter also requires a further randomization through the assignation of a prior distribution, as done in Escobar and West (1995). Further discussions on the galaxy data and its modelling through Bayesian nonparametric mixtures can be found in Lijoi *et. al.* (2005,2006).

Our intention here is not to claim a superiority of our approach neither to give an exhaustive comparison of existing models, since this would basically consist of restating arguments already well established in the literature. We rather aim at illustrating the simplicity of our approach and how this performs in the particular case of this widely discussed data set. Having said that, we do compare it with the Dirichlet process.

Hence, with this purpose in mind, we have used the same framework, *i.e* same model with same parameter and hyper-parameter specifications, used in Section 4.2. We have based the estimations using our approach on 1,000 iterations, and we have used 10,000 iterations plus a 2000 for burning in the sample in the case of the Dirichlet process. For this latter, we have additionally randomized the total mass parameter with a gamma distribution as done in Escobar and West (1995) and make use of the acceleration step suggested by MacEachern (1998). The additional sampling used in this latter approach is needed since otherwise compromised the convergence and therefore the quality of the corresponding density estimation. As we can see in Figure 6, both fits are comparable and satisfactory. Though we emphasize the estimate we have used is obtained with less parameters, less iterations, in less than a tenth of the computational time and with a much simpler implementation.

5. Discussion. The current trend in Bayesian nonparametric mixture modelling focuses on generalizations of the Dirichlet process model. However, in most cases these generalizations result in complex models hard to implement and to apply in real situations, see Lijoi *et. al* (2005b) for an example of how complex these generalizations can be. In this context, our results are appealing, in that we have proposed a simple approach and yet competent for Bayesian nonparametric density estimation. To some extent we could say that we have gone in the opposite direction of such a trend by proposing a simpler approach when using a random distribution with geometric weights. As stated in Section 3, these weights can be seen as the expected values of those corresponding to the Dirichlet process, hence removing a hierarchy from this latter one, which is apparently unnecessary, at least for density estimation purposes. As a byproduct of this simplification, a relatively easy

Gibbs sampler algorithm is available, which results in a simpler alternative to those typically used for Bayesian nonparametric mixtures. It is worth noting that most of these MCMC algorithms are based on the Pólya-urn construction of the Dirichlet process, or its extensions when dealing with other random distributions. In contrast to our algorithm, these algorithms are based on almost sure approximations to the random distribution through exchangeable sequences. See Blackwell and MacQueen (1973) for the Pólya urn scheme.

One might wonder why an approach based on such a simple construction works, at least as well, as other methods based on more complex models, as those based on Dirichlet or Poisson-Dirichlet processes models where an infinite number of beta variables is needed for the corresponding stick-breaking construction. Our explanation is quite straightforward. While the weights would be practically useless if we could only have one of the θ_j's identifying a particular cluster location, the fact that there are an infinite number of possible θ_j's implies that we can have an unlimited number of them supporting this location. Hence the weights for a particular cluster location are obtained via a combination of the geometric weights and the number of θ_j's supporting that cluster location.

For a more mathematical explanation consider the following: Let

$$P = \sum_{j=1}^{\infty} \rho_j \delta_{\phi_j}$$

be a Dirichlet process where the weights have been ordered to be decreasing, so $\rho_1 > \rho_2 > \cdots$. The ϕ_j are iid from some density defined on the real line. Now consider our random distribution function

$$P_G = \sum_{j=1}^{\infty} w_l(\lambda) \delta_{\theta_j}$$

where the θ_j are also iid from the same density as the $\{\phi_j\}$ and the $\{w_j(\lambda)\}$ are the geometric weights, with λ being assigned a distribution.

We can use our model to arbitrarily approximate the Dirichlet model when for some sequence $n_1, n_2, \ldots$ we have

$$|w_1(\lambda) + \cdots + w_{n_1}(\lambda) - \rho_1|,\ |w_{n_1+1}(\lambda) + \cdots + w_{n_2}(\lambda) - \rho_2|, \ldots$$

are all suitably small and correspondingly

$$\max\{|\theta_1 - \phi_1|, \ldots, |\theta_{n_1} - \phi_1|\},\ \max\{|\theta_{n_1+1} - \phi_2|, \ldots, |\theta_{n_2} - \phi_2|\}, \ldots$$

are all suitably small. Since there is positive probability on these events we can see that we do not need the weights to be that exotic; such as those obtained via stick–breaking construction.

Hence, our conclusion is that the

$$P_G = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} \delta_{\theta_j}$$

is sufficient for mixture modeling and that the current trend of further elaborating on Dirichlet process mixture models must be compromised with an application that potentially would require more complicated weights specifications.

Acknowledgements. The authors are grateful to the Associate Editor and two referees for their valuable comments and suggestions. The second author thanks CONACYT for providing the Grant No. J50160-F. The work was completed during a visit by the first two authors to the University of Kent.

References

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn scheme. *Annals of Statistics*. **1**, 353–355.

Escobar, M.D. 1988. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

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.

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

Green, P.J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavean Journal of Statistics* **28**, 355–375.

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

Lijoi, A., Mena, R.H. and Prünster (2005b). Bayesian nonparametric analysis for a generalized Dirichlet process prior *Statistical Inference for Stochastic Processes* **8**, 283–309.

Lijoi, A., Mena, R.H. and Prünster (2006). Bayesian clustering in nonparametric hierarchical mixture models *Proceedings of XLIII Meeting of the Italian Statistical Society* **Vol I**, 449–460.

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

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

MacEachern, S.N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical non-parametric and semiparametric Bayesian statistics* (eds D. Dey, P. Müller and D. Sinha), 23–43. New York: Springer.

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

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

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

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Stephens, M. (2000). Bayesian Analysis of mixture Models with an unknown Number of Components- an alternative to Reversible Jump Methods. *Annals of Statistics.* **28**, 40–74.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation* **36**, 45–54.

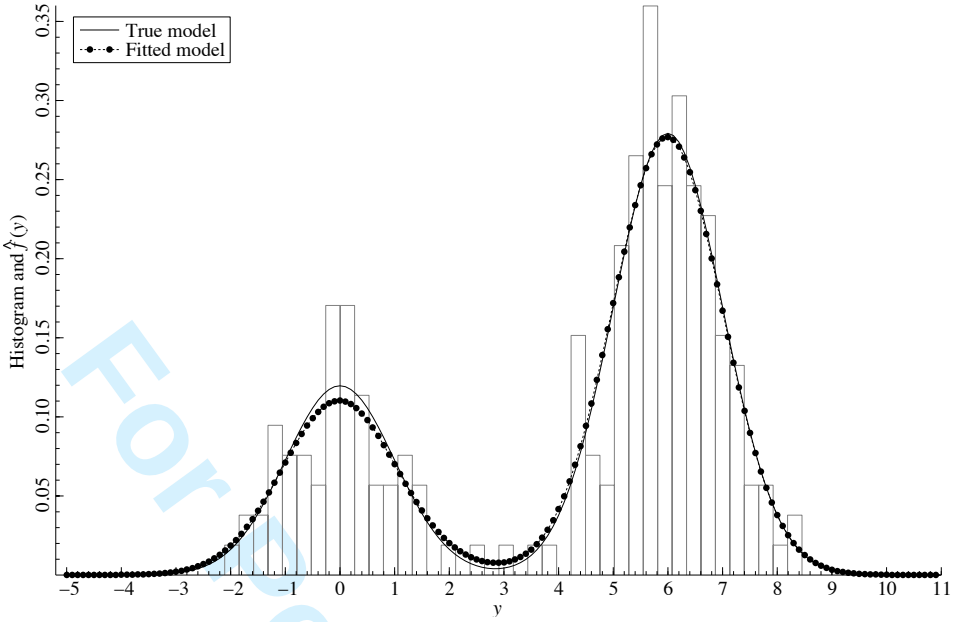

Figure 1: Fitted density (solid line) to simulated data set (histogram) with $n = 200$. The estimate is based on $10,000$ iterations after a burn in of $2,000$ iterations and hyperparameters $(\mathsf{m}, \mathsf{v}, \mathsf{a}, \mathsf{b}) = (3, 1, 1, 1)$, $N_i = 10$ and $d_i \in 1, \ldots, 10$ for all $i = 1, \ldots, 200$.

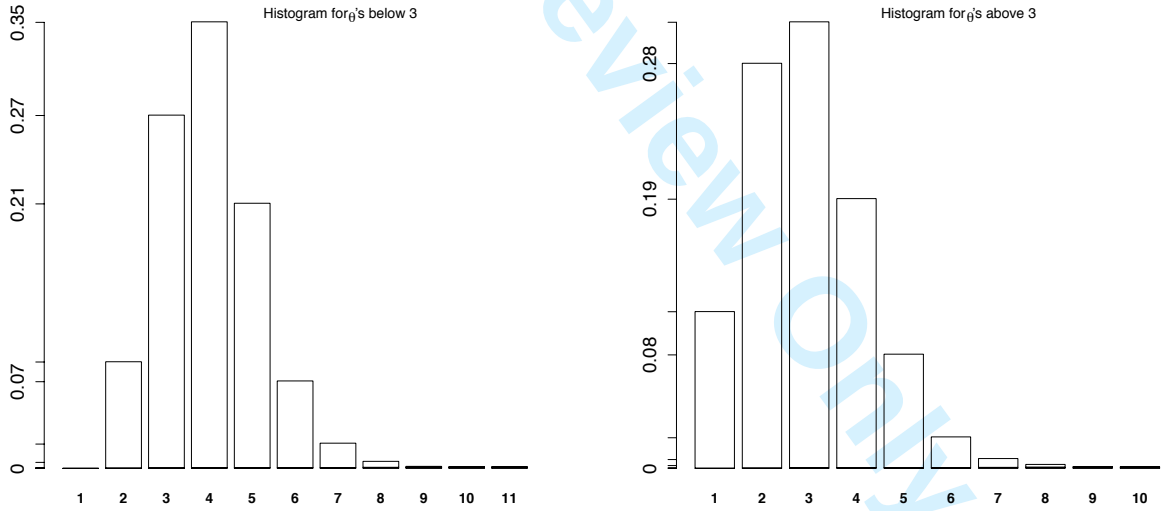

Figure 2: Proportion of the number of θ's needed to capture a particular cluster location, corresponding to the two modal simulated dataset. The histogram on the left hand side corresponds to the θ's needed to capture the cluster located around $y = 0$ and the histogram on the right hand side those needed to capture the cluster located around $y = 6$.

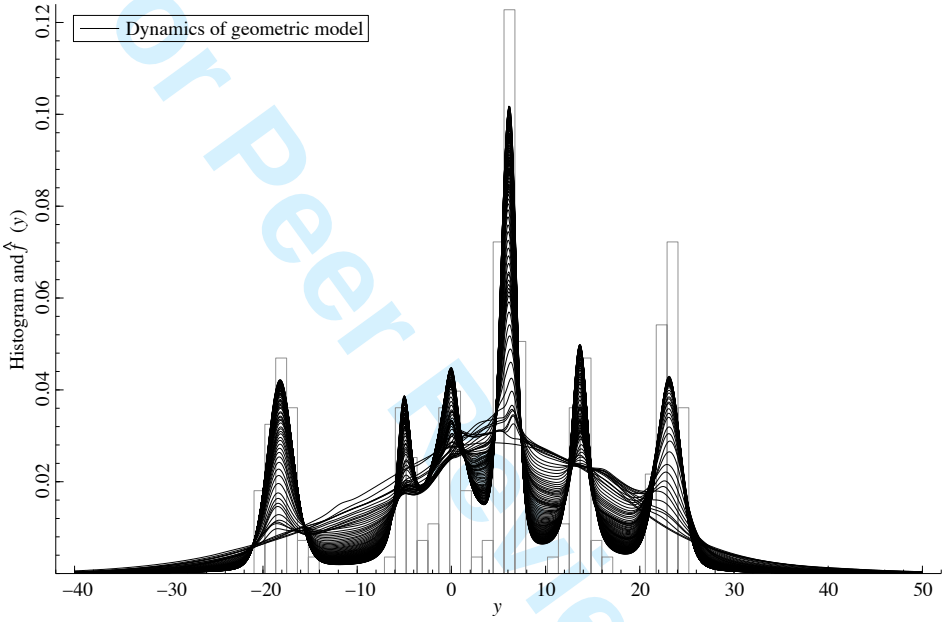

Figure 3: Dynamics of the density estimator, based on the geometric model, through the first 100 iterations of the Gibbs sampler algorithm for the mean-scale mixtures data set. The hyper–parameters are given by $(\mu, \tau, \alpha, \beta, \mathsf{a}, \mathsf{b}) = (0, 100, 0.5, 0.5, 0.5, 0.5)$ and $N_i = 10$ and $d_i \in 1, \ldots, 10$ for all $i = 1, \ldots, 240$.

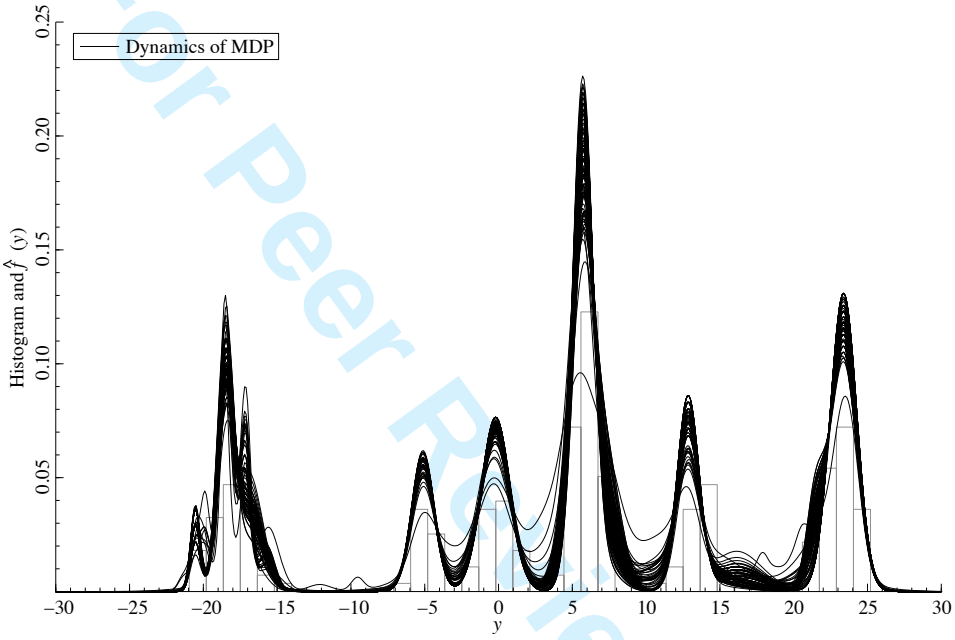

Figure 4: Dynamics of the density estimator, based on the MDP model, through the first 100 iterations of the Gibbs sampler algorithm for the mean-scale mixtures data set. The hyper–parameters are given by $(\mu, \tau, \alpha, \beta, \mathsf{a}, \mathsf{b}) = (0, 100, 0.5, 0.5, 0.5, 0.5)$ and $N_i = 10$ and $d_i \in 1, \ldots, 10$ for all $i = 1, \ldots, 240$.

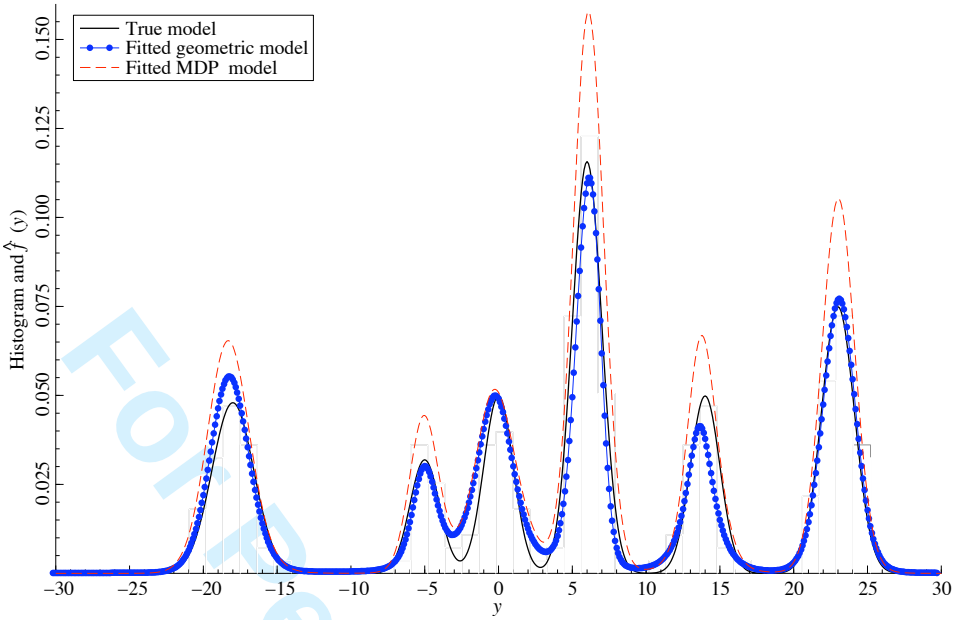

Figure 5: Density estimates for the 6 modes simulated data set based on both the geometric model and the MDP model. The estimates are based on 10,000 after a burn in period of 2000 iterations. The hyper–parameters are given by $(\mu, \tau, \alpha, \beta, \mathsf{a}, \mathsf{b}) = (0, 100, 0.5, 0.5, 0.5, 0.5)$ and $N_i = 10$ and $d_i \in 1, \ldots, 10$ for all $i = 1, \ldots, 240$.

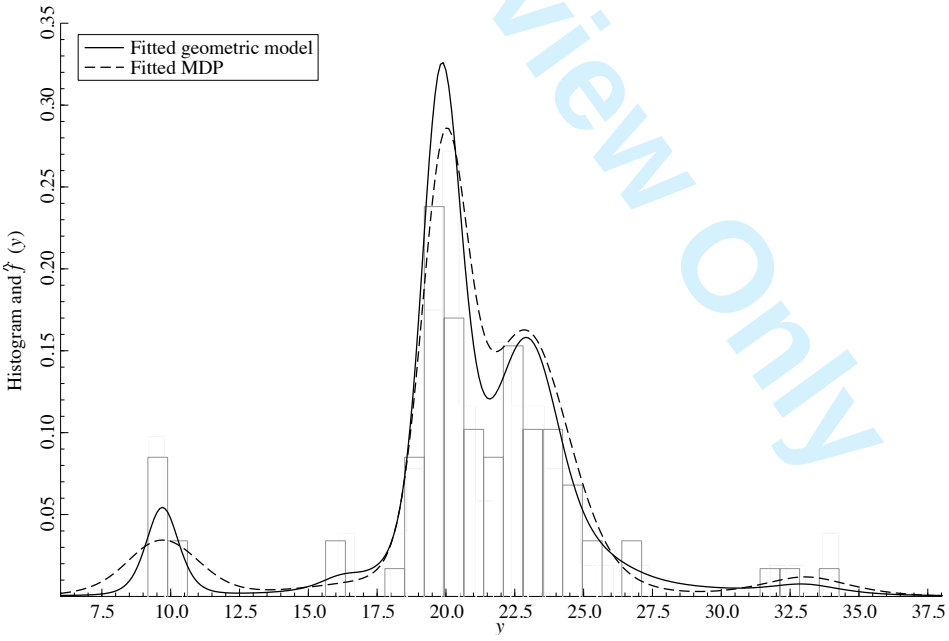

Figure 6: Density estimators for the galaxy data.