On the Bayesian analysis of species sampling mixture models for density estimation

J.E. Griffin[*]

Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.

Abstract

The mixture of normals model has been extensively applied to density estimation problems. This paper proposes an alternative parameterisation that naturally leads to new forms of prior distribution. The parameters can be interpreted as the location, scale and smoothness of the density. Priors on these parameters are often easier to specify. Alternatively, improper and default choices lead to *automatic Bayesian density estimation*. The ideas are extended to multivariate density estimation.

Keywords: Density Estimation, Species sampling models, Dirichlet process mixture models, Mixtures of normals.

1 Introduction

The problem of density estimation has a long history in the statistical literature. We assume that $y_1, \dots, y_n$ are i.i.d. draws from a distribution F, with density f, that must be estimated. In some recent work the focus has shifted from the distribution of observables to the distribution of unobserved random quantities. For example, Bush and MacEachern (1996) consider an unknown distribution of the block effect in a two-way analysis of variance and Müller and Rosner (1997) estimate the distribution of a random effect nonparametrically. In both cases we would be interested in replacing the standard parametric assumption of a normal distribution by a more flexible nonparametric choice which is *centred* over the standard parametric

[*] Corresponding author: Jim Griffin, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. Tel.: +44-2476-574808; Fax: +44-2476-524532; Email: J.E.Griffin@warwick.ac.uk.

form. However, in neither paper is the nonparametric model (using mixtures of normals) centred over the standard model since the hyperparameter of the distribution have different prior distributions under the two models. This paper attempts to address this issue by proposing a structure for the nonparametric model which allows the model to be centred.

A number of approaches and priors have been proposed in the Bayesian literature, which are reviewed in Walker *et al* (1999) and Müller and Quintana (2004) and include: mixture distributions, Dirichlet process priors, Polya trees, and random histograms. In this paper I will concentrate on modelling the unknown distribution by a species sampling model mixture of normals.

$$f(y) = \int \mathrm{N}(y|\mu, \sigma^2)\, dG(\mu, \sigma^2) \tag{1}$$

where

$$G = \sum_{i=1}^{q} p_i \delta_{\mu_i, \sigma_i^2}. \tag{2}$$

The number of components q is an integer or infinity, $\sum_{i=1}^{q} p_i = 1$, and $\mathrm{N}(y|\mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2, which is often called a component of the mixture. The concept of a species sampling model was introduced by Pitman (1996) and makes the assumption that p is *a priori* independent of μ and σ^2, which are i.i.d. from some distribution H. The class includes: finite mixture models (Richardson and Green 1997), Dirichlet process mixtures (Ferguson 1983, Lo 1984), normalized random measures (Nieto-Barajas *et al* 2004) and many more. In this paper, it will be assumed that q is infinite. Recent work on infinite-dimensional mixture models has concentrated on specifying alternative species sampling models to the Dirichlet process, see *e.g.* normalized inverse gaussian processes (Lijoi *et al* 2005) and Poisson-Dirichlet processes (Ishwaran and James 2002). In fact the only non-species sampling model prior developed is the spatial neutral to the right model (James 2006). The mixture of normals is a standard choice and I will assume it throughout the paper (although the ideas are readily extended to other continuous component distributions). The Bayesian analysis of mixture models is reviewed in Marin *et al* (2006) who describe in detail the possible computational approaches to inference and the potential pitfalls in their use.

It is useful to draw a distinction between density estimation and clustering. In the latter our interest often focus on the number of clusters and the allocation of observations to each cluster. It is natural to include prior information about the size, orientation and location of the clusters. However, in density estimation it is not clear that the number of clusters and allocation of observations to each cluster are relevant quantities of interest and will often only enter our thinking in terms of the statistical properties of the estimation procedure. In a subjective framework, the natural quantities on which to place prior information are the unknown density f and perhaps the smoothness of the density or the number of modes.

This paper follows Robert and Titterington (1998) by using uninformative prior for location and scale whilst placing prior information (with possible "benchmark" values) on other aspects. In combination these benchmark values define automatic or semi-automatic Bayesian density estimation procedures. By providing prior information about the unknown density directly, we hope to sensibly provide a compromise between prior and data. The framework also allows us to think sensibly about shrinkage effects, which are inherent in any Bayesian procedure. The approach allows us to replace a parametric distribution in the model by a nonparametric distribution whilst retaining the prior structure on hyperparameters.

This paper will concentrate on specification of H, and the prior distribution of its parameters, rather than the more commonly studied specification of the prior for p. There have been several choices previously discussed. The orignal work of Ferguson (1983) and Lo (1984) assumes that the component variances σ_i^2 in equation (2) share a common value σ_k^2 and to unify notation I will write their prior as $H(\mu, \sigma^2) = \mathrm{N}(\mu|\mu_0, \frac{\sigma_k^2}{n_0})\delta_{\sigma_i^2=\sigma_k^2}$. This prior has recently been studied by Ishwaran and James (2002). Typically a hyperprior would be assumed for σ_k^2 which can be made vague. This prior distribution will act as a starting point for the suggestions in this paper. A drawback with this model is the single variance hyperparameter σ_k^2 which may be an overly restrictive assumption. If parts of the density can be well-represented by a normal distribution with different variances then imposing this constraint will lead to the introduction of many extra normal distributions to model areas with larger spreads. Therefore, it is useful to also consider models where the variance is allowed to vary over the components. A popular choice is a conjugate model for each component, discussed by Escobar and West (1995) where $H(\mu, \sigma^2) = \mathrm{N}(\mu|\mu_0, \frac{\sigma^2}{n_0})\mathrm{IG}(\sigma^2|\alpha, \beta)$ where IG is an inverted Gamma distribution with mean $\frac{\beta}{\alpha-1}$ and variance $\frac{\beta^2}{(\alpha-1)(\alpha-2)}$ if they exist. Its attraction stems from the analytic form of the predictive density of an observation y_{pred} which is equal to $\int \mathrm{N}(y_{pred}|\mu, \sigma^2)h(\mu, \sigma^2)d\mu d\sigma^2$, which plays a key role in standard computational methods. A drawback in the mixture context is the role of n_0. It is not clear why a component with a larger variance should be associated with more uncertain means and unlike the usual normal model we cannot set n_0 to be "small" leading to a "default" analysis since the choice has serious implications for inference about the unknown distribution. Escobar and West (1995) suggest interpreting n_0 as a smoothness parameter and the idea will be developed in this paper. It is also often difficult to choose α and β. A further alternative, discussed by MacEachern and Müller (1998) removes the conjugacy $H(\mu, \sigma^2) = \mathrm{N}(\mu|\mu_0, \sigma_\mu^2)\mathrm{IG}(\sigma^2|\alpha, \beta)$. An important problem is the choice of the hyperparameter and the effect on the posterior distribution. If we consider how these priors enter the model it becomes clear that although density estimation problems are commonly approached using this model, the parameterisation and structure of $H(\mu, \sigma^2)$ relates to an alternative interpretation of the model where we assume that the observed data come from several separate subpopulations, which underlies the use of mixture models for cluster analysis. In this case we express the model in terms of latent allocation variables $s_1, \ldots, s_n$ which link each observation to a subpopulation represented by a component of the mixture where

$$y_i|s_i, \theta \sim \mathrm{N}(\mu_{s_i}, \sigma^2_{s_i})$$

$$p(s_i = j) = p_j.$$

The purpose of this paper is to suggest a simple prior structure when our goal is density estimation.

The paper is organised in the following way: section 2 discusses an alternative parameterisation of the normal mixture model and useful prior distributions for this parameterisation, section 3 describes computational methods to fit these models, section 4 applies these methods to four previously analysed univariate data sets with different levels of non-normality and a bivariate problem, section 5 discuss these ideas and some areas for further research.

2 An alternative parameristation and some prior specifications

This section introduces an alternative parameterisation of the mixture model. If we assume a model with equal component variance, $H(\mu, \sigma^2) = \mathrm{N}(\mu|\mu_0, \sigma_0^2)\delta_{\sigma^2=\sigma_k^2}$, the predictive distribution of y_i is normal with mean μ_0 and variance $\sigma_k^2 + \sigma_0^2$. The reparameterisation defines $\sigma_k^2 = a\sigma^2$ and $\sigma_0^2 = (1-a)\sigma^2$. It seems natural to define a prior distribution on the parameters of the marginal distribution of the observables, μ_0 and σ^2, rather than the centring distribution of component means, μ and σ_k^2. As Mengersen and Robert (1996) note this is linked to standardisation of the data. Transforming to $\frac{y_i-\mu_0}{\sigma}$ allows subsequent development of the model to be considered scale and location free. We now need to interpret the parameter a. A simple interpretation is in terms of the smoothness of the unknown density f. If a is large then all component means μ_i will tend to be close to μ_0 and the marginal distribution will tend to be close to the normal predictive distribution. If a is small then the components will have a small variance and the centres will be close to draws from the normal centring distribution. The Dirichlet process (Ferguson 1973) has been a standard choice of species sampling model in Bayesian nonparametric modelling since the development of computational methods by Escobar and West (1995) who also review the properties of the process. It is parametrised by a distribution H and a measure of precision M. Figure 1 shows a number of realized distributions and the distribution of the number of modes for different choices of a and M. It indicates that the number of modes is largely effected by the choice of a rather than choice M. This is not surprising since modes are determined by local

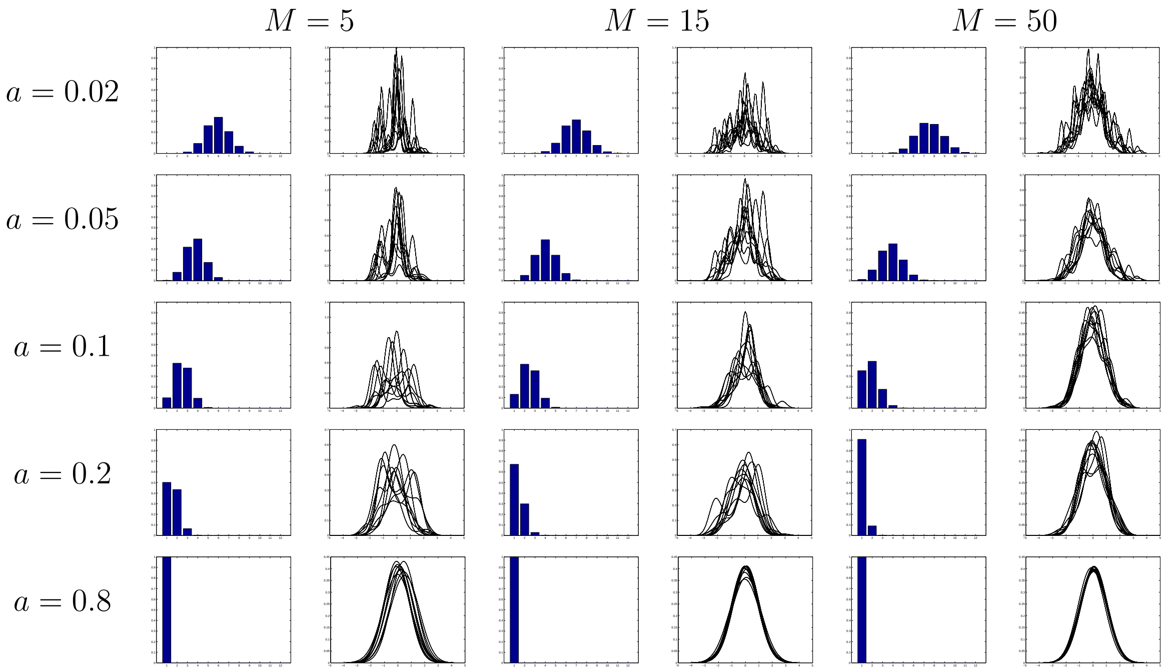

Figure 1: The prior distribution of the number of modes of F and a sample of densities under different hyperparameter choices

features of the realized distribution. The parameter a can be interpreted as a measure of local dependence (and so local variability) and the parameter M as measure of global variability. The figure also gives us an indication of the link between a and the modal number of modes. A values of a between 0.1 and 0.2 indicates a prior belief of bi- or tri-modality wheareas $a = 0.02$ indicates support to a number of modes between 3 and 9. These observations are helpful for defining a variety of prior distributions of a. The new prior distribution is a reparameterisation of the usual conjugate prior distribution where $a = \frac{n_0}{1+n_0}$, which is usually assumed fixed and small which implies unsmooth densities. The scaling is surprising since $n_0 = 0.01$ would be considered a large value but implies many modes. A notable exception is Richardson and Green (1997) who define $H(\mu, \sigma^2) = \text{N}(\zeta, \kappa^{-1})\text{Ga}(\sigma^{-2}|\alpha, \beta)$ in a finite mixture model with a Gamma hyperprior on β. Another interesting aspect of the prior is the importance of the role played a relative to M in the realised distributions.

I will use various moments of the observables and the unknown distribution f to clarify, and quantify, the roles of the parameters. The constant component-specific variance can be generalized to $\sigma_k^2\zeta_i$, where $\text{E}[\zeta_i] = 1$ to allow greater flexibility. A standard choice is an inverse gamma distribution for ζ_i with shape parameter α and scale parameter 1, which is the conditionally conjugate form. The advantage of the single σ^2 is the smaller number of parameters to be estimated in the model which we hope will lead to more tighly fitting model but a random effects specification for the variance can lead to a smaller number of components with certain data. A mixture distribution for ζ_i would define a compromise prior

$$p(\zeta_i) = w\,\delta_{\zeta=1} + (1-w)\,(\alpha-1)\mathrm{IG}(\alpha,1).$$

If we consider a more general form of mixture density for f

$$f(x) = \sum_{i=1}^{\infty} p_i k(x|\mu_i, \sigma_k^2, \phi)$$

where k is a symmetric probability density function with mean μ_i, variance $\sigma_k^2\zeta_i$ and any other parameters of the density function denoted by ϕ. Let the mean of the centring distribution H be μ_0 then the first two predictive central moments have the form

$$\mathrm{E}[y_i] = \mathrm{E}[\mu_i] = \mu_0, \quad \mathrm{V}[y_i] = \mathrm{V}[\mu_i] + \sigma_k^2\mathrm{E}[\zeta_i] = \sigma^2$$

and the overall predictive variability can be divided into a component due to the variability within components and between components so that

$$\mathrm{V}[\mu_i] = (1-a)\sigma^2, \quad \sigma_k^2 = \frac{a\sigma^2}{\mathrm{E}[\zeta_i]}.$$

The predictive skewness have the form

$$\begin{aligned}\mathrm{E}[(y_i-\mu_0)^3] &= \mathrm{E}[\mathrm{E}[(y_i-\mu_i+\mu_i-\mu_0)^3|\mu_i,\sigma_{ki}^2]]\\ &= \mathrm{E}[(y_i-\mu_i)^3|\mu_i] + \mathrm{E}[(\mu_i-\mu_0)^3],\end{aligned}$$

the sum of the within-component and between-component skewness, and the kurtosis can be expressed as

$$\begin{aligned}\mathrm{E}[(y_i-\mu_0)^4] &= \mathrm{E}[\mathrm{E}[(y_i-\mu_i+\mu_i-\mu_0)^4|\mu_i]]\\ &= \mathrm{E}[\mathrm{E}[(y_i-\mu_i)^4|\mu_i]] + 6a(1-a)\sigma^4 + \mathrm{E}[(\mu_i-\mu_0)^4].\end{aligned}$$

If both distributions are chosen to be normal then this expression equals $3\sigma^4$. However heavier tailed predictive distribution will arise through either changes to the component distribution or, perhaps more usefully, the distribution of the component means. These properties are unaffected by the choice of by the choice of the species sampling model. Of course, the species sampling model will effect the variability in the moments of realized distribution. To consider the effect of the species sampling model and the parameter a, we look at the following quantity

$$\mathrm{Cov}[f(x_1), f(x_2)] = C(x_1,x_2)\sum_{i=1}^{\infty}\mathrm{E}[p_i^2]$$

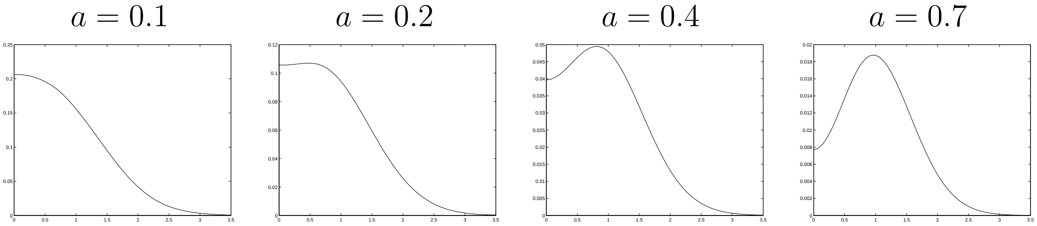
$a = 0.1$ $a = 0.2$ $a = 0.4$ $a = 0.7$

Figure 2: $C(x_1, x_1)$ with a standard normal predictive distribution and various values of a

where

$$C(x_1, x_2) = \mathrm{E}[k(x_1|\mu, a\sigma^2\zeta_i, \phi)k(x_2|\mu, a\sigma^2\zeta_i, \phi)] - \mathrm{E}[k(x_1|\mu_i, a\sigma^2, \zeta_i, \phi)]\mathrm{E}[k(x_2|\mu_i, a\sigma^2, \zeta_i, \phi)].$$

The variability of $f(x_1)$ is then

$$\mathrm{V}[f(x_1)] = C(x_1, x_1) \sum_{i=1}^{\infty} \mathrm{E}[p_i^2].$$

The first part of the product, $C(x_,x_1)$, on the right-hand side is related to the choice of k and a and the second part is related to the choice of species sampling model. If we use a Dirichlet process mixture $\sum_{i=1}^{\infty} \mathrm{E}[p_i^2] = \frac{1}{M+1}$. Figure 2 shows $C(x_1, x_1)$ when we assume a standard normal predicive distribution in the mixture of normals model with various values of a. The variability decreases as the value of a increases but a second effect is also clear: the variability will only be monotone decreasing in x for small values of a. Consequently large a represents a confidence in the density at the mean but less confidence in the density in the region around one standard deviation. An alternative measure, which underlies our understanding of the species sampling models themselves is the variability in the probability measure on a set B which can be expressed as

$$\mathrm{V}[F(B)] = \int_B \int_B C(x, y)\, dx\, dy.$$

The correlation between the density values at two points can be expressed as

$$\mathrm{Corr}[f(x_1), f(x_2)] = \frac{C(x_1, x_2)}{\sqrt{C(x_1, x_1)C(x_2, x_2)}}.$$

The correlation structure of $f(x)$ is independent of the choice of the species sampling model. Therefore we can consider a and the form of the component density as correlation parameters (although in this paper will restrict attention to the standard normal choice). Figure 3 shows the autocorrelation structure for various values of a. For small a the dark area is almost contained by two parallel lines which suggests that the correlation is a function of the distance between two points only. As a increases this pattern disappears and larger absolute values of

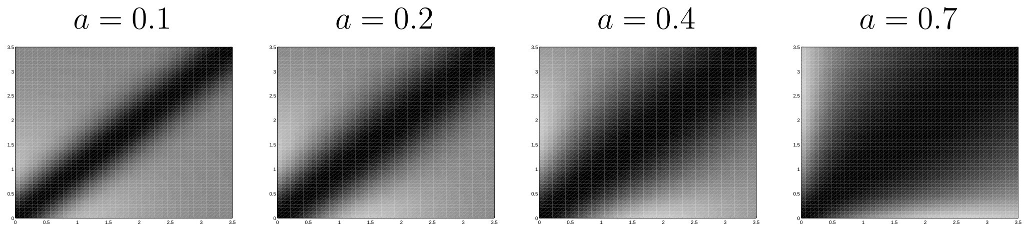
$a = 0.1$ $a = 0.2$ $a = 0.4$ $a = 0.7$

Figure 3: Prior correlation between the density values at two points x_1 and x_2 for a model with a standard normal predictive distribution and various values of a where darker colours represent larger correlations

x are associated with much larger ranges (the distance at which the autocorrelation is equal to some small prespecified value). The autocorrelation between two sets B_1 and B_2 can be expressed as

$$\begin{aligned}\text{Corr}(F(B_1), F(B_2)] &= \frac{\int_{B_1}\int_{B_2} C(x,y)\,dx\,dy}{\sqrt{\int_{B_1}\int_{B_1} C(x,y)\,dx\,dy \int_{B_2}\int_{B_2} C(x,y)\,dx\,dy}} \\ &= \int_{B_1}\int_{B_2} w(x,y)\text{Corr}(f(x), f(y))\,dx\,dy\end{aligned}$$

where

$$w(x,y) = \sqrt{\frac{C(x,x)C(y,y)}{\int_{B_1}\int_{B_1} C(x,y)\,dx\,dy \int_{B_2}\int_{B_2} C(x,y)\,dx\,dy}}.$$

The measures considered in this section quantify the relationships that are evident from the figure 1. The parameter a controls the local prior behaviour of the density function and, at least in the Dirichlet process case, the parameter of the species sampling model controls the general variability. It seems reasonable given the results on the variance and correlation of the density function to assume that these relationship will largely carry over to other species sampling models. The following section uses these ideas to develop prior distribution for a and the location and scale parameters μ_0 and σ^2.

2.1 Prior distributions for the parameters of the model

One purpose of this paper is to suggest forms of prior for the mixture model that allow us to replace a parametric distribution, in this case the normal distribution, by a nonparametric alternative. In particular it would useful to maintain the same prior structure across these two possible specifications. There are two standard choices of prior for μ_0 and σ^2: the improper choice of Jeffreys' prior $p(\mu_0, \sigma^{-2}) \propto \sigma^2$ and the conjugate choice $p(\mu_0, \sigma^{-2}) = \text{N}(\mu|\mu_{00}, \phi\sigma^2)\text{Ga}(\sigma^{-2}|\alpha, \beta)$. The second choice always leads to a proper posterior distribution. However the Jeffreys'prior can lead to an improper posterior distribution. The following result shows that the posterior distribution will always be proper. Robert and Titterington (1998) have previously considered a similar approach for a different prior in finite mixture models. They place Jeffreys' prior distribution on the parameters of the first component and then allow the location and scale of the k-th cluster to depend on the locations and scales of the previous $k-1$ components. This seems more suited to a finite mixture case for cluster analysis rather than density estimation problems where centring the predictive distribution over a particular parametric form seems a useful starting for prior specification. They observe that dependence between the priors on the parameters of each component is key to the use of improper priors for location and scale and the same is true for the prior proposed in this paper. It is simple to show posterior existence for the prior structure in this paper. In particular the posterior will exist if

$$p(y|\mu_0,\sigma^2,a)=\int \sigma^{-2}\prod_{i=1}^{l}\int k(y|\mu_i,\zeta_i,a,\sigma^2)h(\mu_i|\mu_0,a,\sigma^2)p(\zeta_i)\,d\mu_i\,d\zeta_i\,d\mu_0\,d\sigma^2$$

is finite, which is true for the the mixture of normals models considered in this paper.

Finally, the prior specification for the smoothness parameter a and the parameters of the species sampling model is considered. In this paper, the choice of species sampling model will be restricted to the Dirichlet process and a prior distribution for the mass parameter M is proposed. The form of the prior distribution of a is restricted to follow a Beta distribution and several possible parameter choices are considered. The prior distribution of the nonparametric part is defined through a prior distribution for $\zeta=\sum_{i=1}^{\infty}\mathrm{E}[p_i^2]$ with the density

$$p(\zeta)=n_0^{\eta}\frac{\Gamma(2\eta)}{(\Gamma(\eta))^2}\frac{[\zeta(1-\zeta)]^{\eta-1}}{[(n_0-1)\zeta+1]^{2\eta}}.$$

In the Dirichlet process case, where $\zeta=\frac{1}{M+1}$, the properties of this prior distribution are discussed by Griffin and Steel (2004).

Figure 4 shows realisations of the density and the distribution of the number of modes for various choice of the parameters of the prior distribution of a and the M. If we choose a to follow a uniform distribution $(0,1)$ then the distribution of the number of modes is peaked around 1. This prior is giving strong support to a unimodal density with a broadly normal shape. This could be a sensible prior distribution if our goal is to replace a parametric distribution with a nonparametric alternative. The choice of a Be$(1,10)$ places substantial mass on values of a less than 0.2 implying less smooth distributions. It gives a prior modal value of 2 for the number of modes and relatively flat shape supporting a large range of modes. This could represent a more sensible prior distribution in density estimation where we might expect to have a large departure from normality with several modes. This choices are just suggestion and other choices may be more appropriate in other situation. For example a Be$(1.75,10.25)$ acts like a compromise between the two previous choices and implies a modal value of 1 but with a wider spread than the Be$(1,1)$.

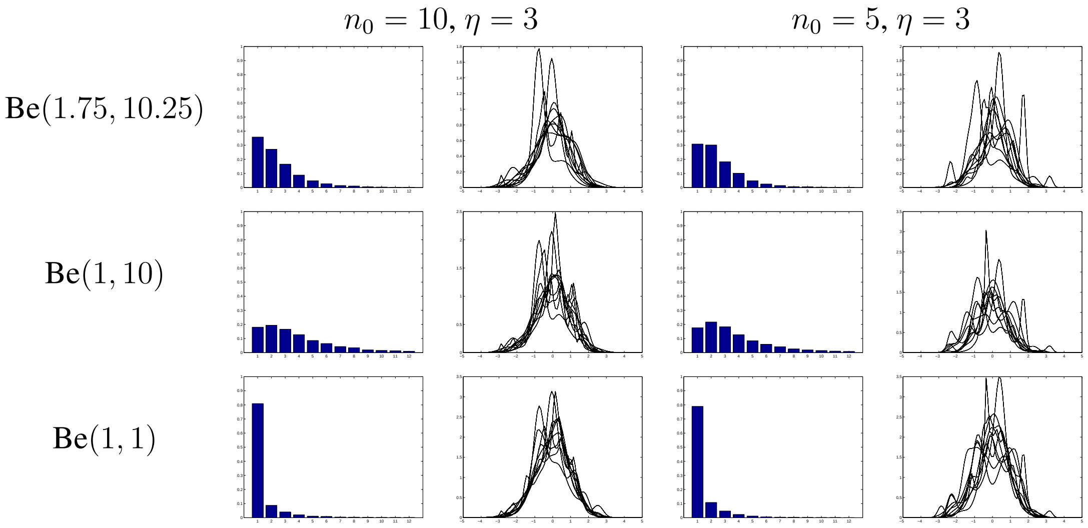

Figure 4: The prior distribution of the number of modes of F and a sample of densities under different hyperparameter choices

2.2 Multivariate versions

The ideas described up to this point relate to univariate density estimation. However, the multivariate problem is important and has received particularly attention in the Bayesian literature on random effects model where the distribution of the random effects is to be estimated (see *e.g.* Müller and Rosner 1997). The smoothness parameter a in the univariate case defines the proportion of the overall variance assigned to within component variation. There is no single natural extension to the multivariate case but there are two natural starting points: the orientation of the observed vectors has some meaning or the orientation of the observed vectors is essentially arbitary (in which case we would be happy to rotate axis without affecting the analysis). In both case the univariate model is extended by assuming that the mean of data is μ_0 and the covariance matrix is Σ. In the first case, we want to respect the dimension of the variables and to have different smoothness parameter (values of a) for each dimension. The choice of within-component covariance matrix Σ_k such that

$$\Sigma_{kij} = \sqrt{a_i a_j}\Sigma_{ij}$$

implies that the correlation between the i-th and j-th variable is $\frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$. This define Model I which allows different levels of departure from the centring model in different dimension and the prior for the marginal distibution of the i-th variable will the univariate model with smoothing parameter a_i. The prior covariance matrix of the μ_i will then have the form $\Sigma - \Sigma_k$. In the second case, it seems more natural to first transform the data vector y_i to

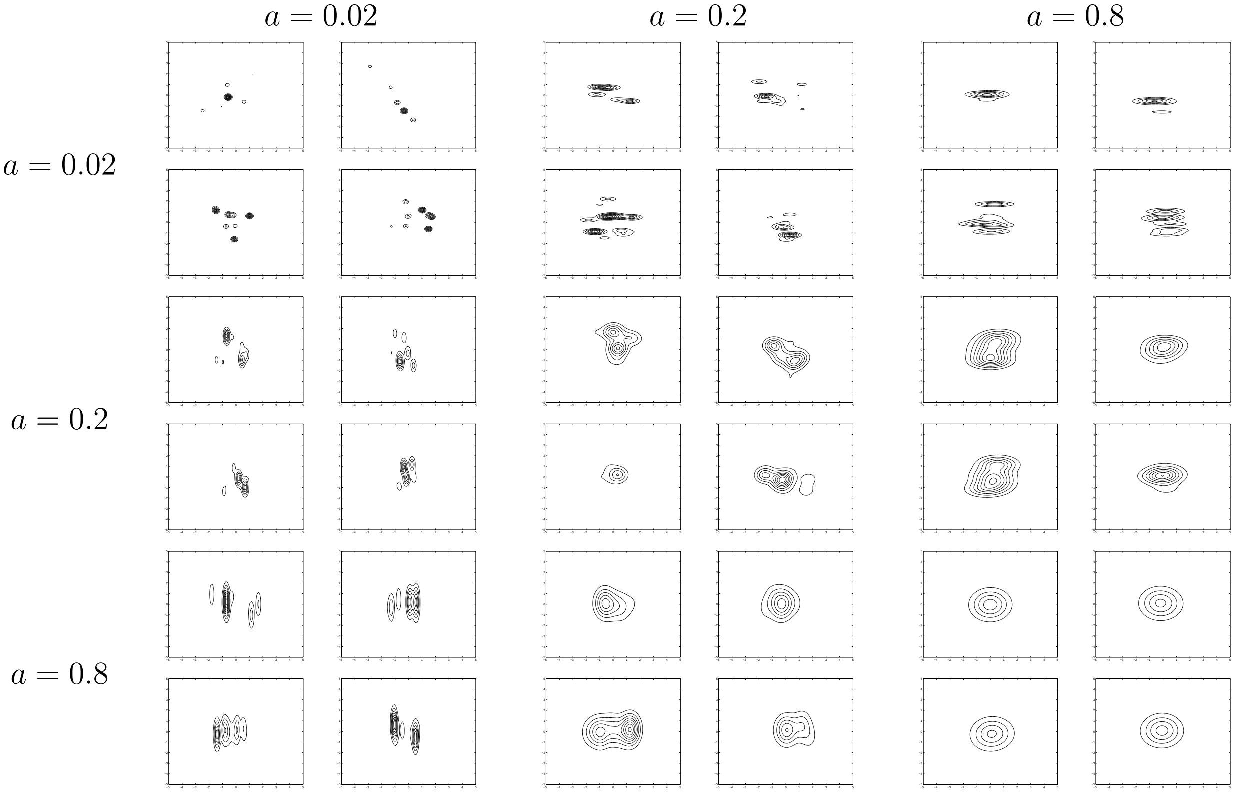

Figure 5: Four realisations of the multivariate model 1 with different value of a correlation 0 with $M = 5$

$z_i = A^{-1}(y_i - \mu_0)$ where A is the Choleksy decomposition of Σ and the distribution of z_i will be centred over a multivariate standard normal. The within component covariance matrix is assumed to have the form

$$\Sigma_k = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_p \end{pmatrix}.$$

and the between component covariance is assumed to be

$$\Sigma_0 = \begin{pmatrix} 1 - a_1 & & \\ & \ddots & \\ & & 1 - a_p \end{pmatrix}.$$

This defines Model II. Some realisations of the processes for Model I with various choices of the parameters are shown in figures 5 and 6. Clearly small values of a lead to distibutions with many modes and typically well-seperated components. The value of $a = 0.2$ and larger give rise to distribution with less modes and a more cohesive distribution. As in the univariate case, it is also possible to define a version where each cluster variance is different. Let

$$\mathrm{E}(y_i|\mu_i, \sigma^2) = \mu_i, \qquad \mathrm{V}(y_i|\mu_i, \sigma^2) = \Sigma_k \zeta_i$$

where ζ_i is a distribution with the indentity as the mean. Standard choices such a the inverted Wishart distribution can fit into this structure.

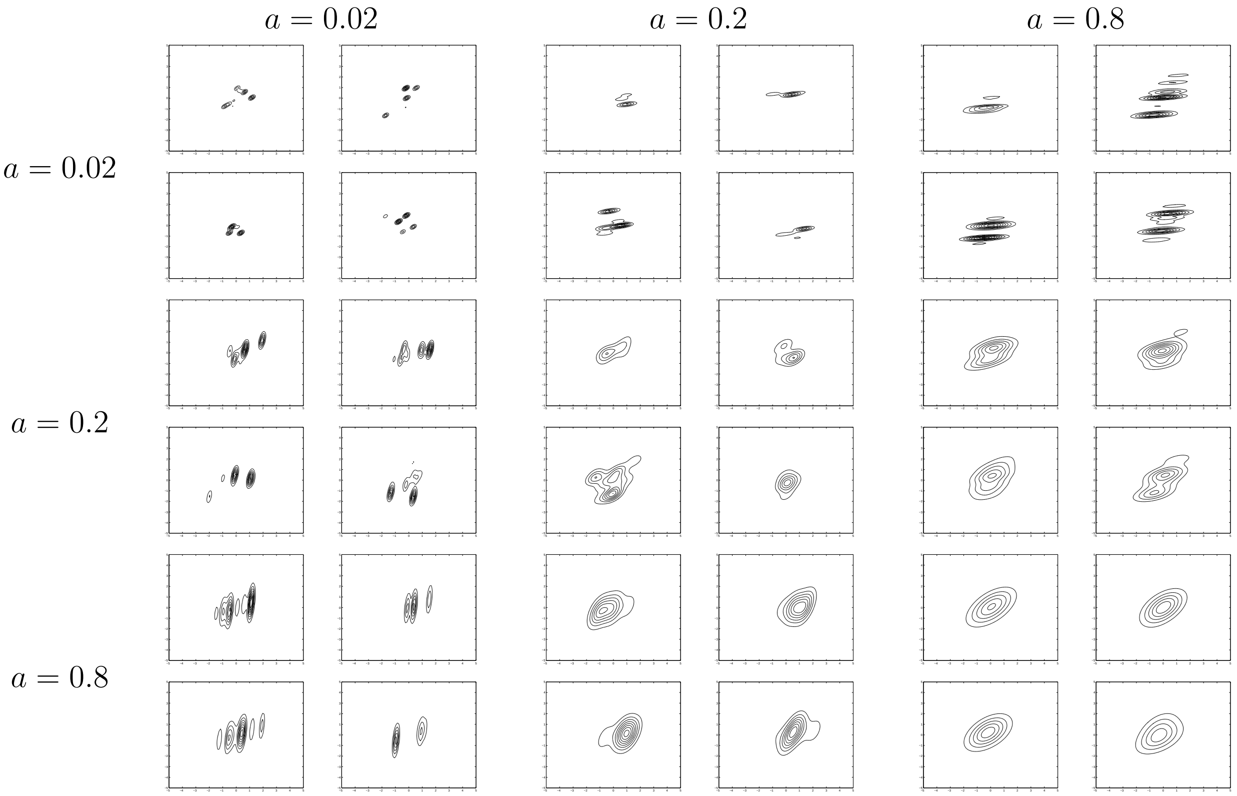

Figure 6: Four realisations of the multivariate model 1 with different value of a correlation 0.5 with $M = 5$

3 Computational methods

The fitting of Dirichlet process mixture models have been greatly helped by the development of efficient MCMC methods. The usual methods make use of the Polya urn scheme representation (Blackwell and MacQueen, 1973) to avoid the infinite number of elements in G, which are often called marginal methods. The models developed in this paper are non-conjugate and methods for this case are described in MacEachern and Müller (1998) and Neal (2000). However, all the examples in this paper use the Retrospetive Sampling scheme for stick-breaking mixture models described in Papaspiliopoulos and Roberts (2004), which uses a finite truncation of G whilst avoiding truncation error. This allows direct posterior inference for f and G. Alternatively, Gelfand and Kottas (2002) describe methods for making inference about these objects using marginal methods. Discussion of computational methods is not the purpose of this paper and the reader is referred to Papaspiliopoulos and Roberts (2004) for comparison of the various methods. All methods make use of the Gibbs sampler and the full conditional distribution for each parameter are fully described in each paper. This section describes methods for sampling any unusual full conditional distributions. Before describing these steps it is important to note that in all the methods, the i-th observations is allocated to a component value $(\mu_{s_i}, \sigma^2_{s_i})$, which leads to simple forms for many full conditional distributions in a Gibbs sampling scheme.

3.1 Updating M

M can be updated using an independence Metropolis-Hastings sampler. The Newton-Raphson method is used to find the mode of the full conditional distribution, then the proposal distribution is a t-distribution centred at the mode, with α degrees of freedom and precision parameter $\lambda = \frac{\alpha}{\alpha+1} \times$ -Hessian. A default choice of α would be 3.

3.2 Updating σ^2_k and σ^2_0 in the normal model

To update a, σ^2, we transform back to σ^2_k and σ^2_0 where $\sigma^2 = \sigma^2_k + \sigma^2_0$ and $a = \frac{\sigma^2_k}{\sigma^2_k+\sigma^2_0}$. The jacobian of the transformation is $\frac{1}{\sigma^2_k+\sigma^2_0}$. The transformed prior is

$$p(\sigma^2_k, \sigma^2_0) = \frac{1}{\sigma^2_k + \sigma^2_0}\, p_a\left(\frac{\sigma^2_k}{\sigma^2_k + \sigma^2_0}\right)\, p_{\sigma^2}(\sigma^2_k + \sigma^2_0).$$

If σ^2 has an improper prior, we use a rejection sampler with the envelope

$$\sigma^2_k \sim \text{IG}\left(n/2 + \beta\hat{a} - (1-\hat{a})\alpha, \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_{s_i})^2\right)$$

$$\sigma^2_0 \sim \text{IG}\left(k/2 + \alpha(1-\hat{a}) - \beta\hat{a}, \frac{1}{2}\sum_{i=1}^{k}(\theta_i - \mu_0)^2\right)$$

where $\hat{a}$ is the current value of $\frac{\sigma^2_k}{\sigma^2_0+\sigma^2_k}$. The acceptance probability is

$$\frac{1}{(\sigma^2_k + \sigma^2_0)^{\alpha+\beta}} \left(\frac{\sigma^2_k}{\hat{a}}\right)^{(\alpha+\beta)\hat{a}} \left(\frac{\sigma^2_0}{1-\hat{a}}\right)^{(\alpha+\beta)(1-\hat{a})}.$$

In the proper case, where σ^2 follows an inverted Gamma distribution with shape c and scale d, we define the joint distribution

$$\sigma_k^2 \sim \text{IG}\left(n/2 + (\beta + c)\hat{a} - (1 - \hat{a})\alpha, \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_{s_i})^2 + d\hat{a}^2\right)$$

$$\sigma_0^2 \sim \text{IG}\left(k/2 + (\alpha + c)(1 - \hat{a}) - \beta\hat{a}, \frac{1}{2}\sum_{i=1}^{k}(\theta_i - \mu_0)^2 + d(1 - \hat{a})^2\right)$$

which can be used as a proposal distribution in a Metropolis-Hasting sampler which has acceptance probability

$$\min\left\{1, \frac{\left(\frac{{\sigma'}^{2\,\hat{a}}_{k}{\sigma'}^{2\,1-\hat{a}}_{0}}{{\sigma'}^2_k+{\sigma'}^2_0}\right)^{\alpha+\beta+c}\exp\left\{-d\left[\frac{1}{{\sigma'}^2_k+{\sigma'}^2_0}-\frac{\hat{a}}{{\sigma'}^2_k}-\frac{1-\hat{a}}{{\sigma'}^2_0}\right]\right\}}{\left(\frac{\sigma^{2\,\hat{a}}_{k}\sigma^{2\,1-\hat{a}}_{0}}{\sigma^2_k+\sigma^2_0}\right)^{\alpha+\beta+c}\exp\left\{-d\left[\frac{1}{\sigma^2_k+\sigma^2_0}-\frac{\hat{a}}{\sigma^2_k}-\frac{1-\hat{a}}{\sigma^2_0}\right]\right\}}\right\}$$

where ${\sigma'}^2_k$ and ${\sigma'}^2_0$ represent the proposed values of σ_k^2 and σ_0^2 respectively.

3.2.1 Updating ζ_i

If $\zeta_i \sim \text{IG}(\alpha, \beta)$ then

$$\zeta_i \sim \text{IG}\left(\alpha + 0.5n_i, 1 + 0.5\frac{\sum_{j|s_j=i}(x_j - \mu_i)^2}{(\alpha - 1)a\sigma^2}\right)$$

Updating a and σ^2 use the rejection sampler from above replacing $\sum_{j|s_j=i}(x_j - \mu_i)^2$ by $\frac{1}{(\alpha-1)\zeta_i}\sum_{j|s_j=i}(x_j - \mu_i)^2$.

3.3 Multivariate extensions

3.3.1 Updating Σ and a

For both models described in section 2.2, these parameters can be updated using a random walk Metropolis-Hastings with normal proposals whose variance have been tuned to achieve an acceptance rate close to 0.234.

4 Examples

4.1 Univariate density estimation

The Bayesian model developed in this paper will be illustrated on a series of data sets previously analysed in the literature. The "galaxy data" was initially analysed by Roeder (1992) and introduced into the Bayesian literature by Roeder and Wassermann (1997). It has become a standard data set for the comparison of Bayesian density estimation models and their related computational algorithms. The data records the estimated velocity ($\times 10^{-2}$) at which 82 galaxies are moving away from our galaxy. Some galaxies are thought to be moving at similar speeds whilst other move much faster or slower. Inferring the clusters of galaxy is the main inferential problem. Of course, this rather contradict the basis of this paper and the problem here is treated as density estimation (in common with much of the subsequent literature). However, if the clusters are not assumed normal then modality of the data may give some clue to the various groupings. The "acidity" data refers to a sample of 155 acidity index measurement made on linkes in noth-central Wisconsin which are analysed on the log scale, the "enzyme" data measures the enzymatic activity in the blood of 245 unrelated individuals. It is hypothesised that there are groups of slow and fast metabolizers. These three data sets were previously analysed in Richardson and Green (1997). A final data records the red blood cell sodium-lithium countertransport (SLC) in six large English kindreds. The data was previously analysed by Roeder (1994) who wants to distinguish between a two and three component finite mixture, which she postulates will have the same variance. Further background to the genetic implications of different types of multi-modality are explained in the reference. Some summary statistics for the four data sets are shown in table 1. In all analyses the prior for M is set to have hyperparameters $n_0 = 5$ and $\eta = 3$ and $\zeta_i \sim \text{IG}(2, 1)$. Two prior choices for a were chosen: $\text{Be}(1, 10)$ and $\text{Be}(1, 1)$ which represent a prior distribution with substantial prior mass on a wide range of modes and prior distribution that places a lot of a mass on a single mode.

Data set	sample size	mean	standard deviation
Galaxy	82	20.8	4.6
Log Acidity	155	5.11	1.04
Enzyme	245	0.62	0.62
Sodium Lithium	190	0.26	0.099

Table 1: Summary statistics for the 4 data sets

Figure 7 shows the predictive distribution (solid line) and a 95% highest probability density region of $f(x)$ for each of the four data sets when the prior distribution is $\text{Be}(1, 1)$. The results are largely unchanged by the alternative prior distribution (although some features do change which will be discussed subsequently). These results are extremely similar to previous analyses, although the galaxy data results do differ largely from analyses described in Marin *et al* (2006) and Wasserman and Roeder (1997) who find a single mode between 20 and 24 rather than the two modes inferred in this analysis. The extra mode has been found in

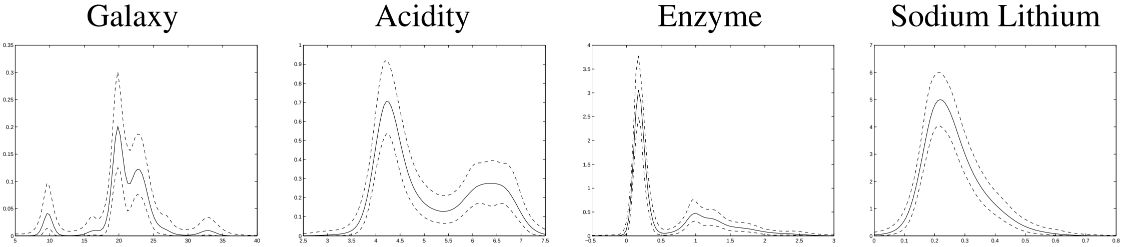

Figure 7: Posterior predictive densities for the four data sets with a pointwise 95% HPD interval

a number of other analyses *e.g.* Richardson and Green (1997).

Table 2 shows summaries of the posterior distributions of a and M under the two prior distributions of a. The parameter a has been interpreted as the smoothness of the realised distribution and related to the number of posterior modes. The results show that the distribution which are less smooth (in particular the multi-modal galaxy data) have smaller estimates of a, which is estimated with good precision in each case. Unsurprisingly the unimodal distribution of sodium lithium has the highest estimates of a. The posterior distribution is robust to the choice between the two prior distribution when the density are estimated to be less smooth. For distributions which have higher levels of smoothness the prior distribution is much more influential. This mostly shows a prior-likelihood mismatch since the tighter prior distribution places nearly at its mass below 0.2 and neglible mass above 0.3. Clearly under the more dispersed prior distribution the posterior distribution for the acidity and sodium lithium data sets place mass at larger values. This suggests that a dispersed prior distribution will be useful when we are unsure about the smoothness and likely modality of the data. The posterior inferences of M for each data set show only small differences between the posterior median and credibility intervals, illustrating that differences in modality will not be captured in these models by the M parameters. The results for the number of clusters (not shown) also display a lack of difference in the form of the posterior distribution across the different data sets.

	a		M	
Data set	Be$(1,1)$	Be$(1,10)$	Be$(1,1)$	Be$(1,10)$
Galaxy	0.04 (0.01, 0.12)	0.03 (0.01, 0.10)	3.73 (1.14, 10.80)	3.93 (1.31, 10.06)
Acidity	0.16 (0.04, 0.46)	0.10 (0.03, 0.27)	3.47 (0.95, 10.66)	3.23 (0.83, 9.48)
Enzyme	0.06 (0.01, 0.23)	0.05 (0.01, 0.16)	2.40 (0.75, 6.40)	2.39 (0.69, 7.31)
Sodium Lithium	0.44 (0.12, 0.82)	0.17 (0.04, 0.41)	3.71 (0.79, 15.01)	2.25 (0.49, 6.69)

Table 2: The posterior distribution of a summarised by the posterior median with 95% credibility interval in brackets for the 4 data sets

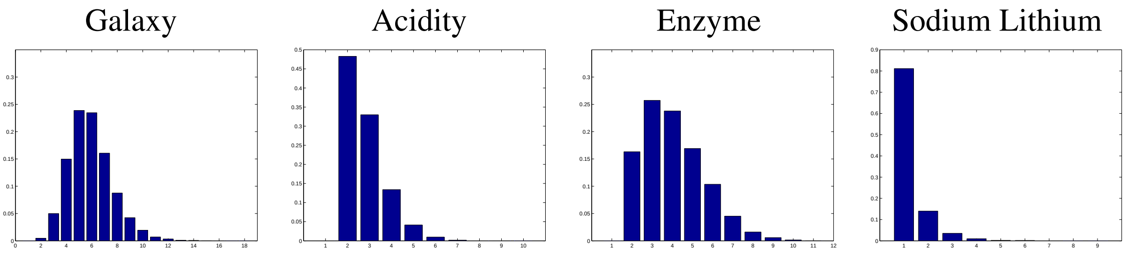

Figure 8: Posterior predictive densities for the four data sets

The inferences about the number of modes is shown in figure 8. The degree of posterior uncertainty for most of the data sets (with the exception of sodium lithium) is substantial and is obscured in the posterior predictive distributions. In all cases the results are shown for Be$(1, 1)$ prior, as with a, the results are unchanged with the second prior for the galaxy and enzyme data. The galaxy data supports a range of values between 3 and 9. The values 5 and 6 receive almost equal posterior support. The acidity data shows strongest support for 2 modes and some uncertainty about and extra 1 or 2 modes. The enzyme data also shows a large amount of posterior uncertainty about the number of modes. It show most support for 3 modes with good support for upto 7 modes. The results are rather surprising given the shape of the posterior predictive distribution. It seems reasonable to conjecture that the form of the model may lead to these results. The data can be roughly divided into two groups. The skewness of the second group can only be captured by a number of normal distribution. This may lead to rather unrealistic estimates of the number of modes. The sodium lithium data set results are shown for the Be$(1, 1)$ prior. The posterior distribution strongly supports a single mode with a posterior probability of about 0.8.

4.2 Multivariate data example

As an example, I re-analyse a data set, previously analysed by Bowman and Azzalini (1997), that relates to a study of the development of aircraft technology originally analysed by Saviotti and Bowman (1984). The data set contain six characteristics (total engine power, wing span, length, maximum take-off weight, maximum speed and range) of aircraft designs. The first two principal components are shown in figure and can be interpreted as "size" and "speed adjusted for size". Further details are given in the reference. A Beta(1,1) prior distribution was used for a_1 and a_2. The prior distribution of Σ was chosen to be an inverse Wishart distribution with 3 degrees of freedom and the prior mean fixed to the sample covariance matrix. The data is analysed using Model I to illustrate the methodology although Model II seems more appropriate in this application.

The posterior distribution of a_1 and a_2 are summarised in table 3 and show that there is a

parameter	median	95% credible interval
a_1	0.103	(0.078, 0.137)
a_2	0.095	(0.086, 0.140)

Table 3: Summary of the posterior distribution of a_1 and a_2 for the aircraft data

similar level of non-normality in both variables. Once again both parameter are estimated to a good level of certainty. Figure 9 shows a scatterplot of the data and the posterior predictive distribution for the chosen prior. The predictive distribution gives a good description of the data. In particular, the higher density of points on the for small x_1 is well captured and gives similar results to Bowman and Azzalin (1997).

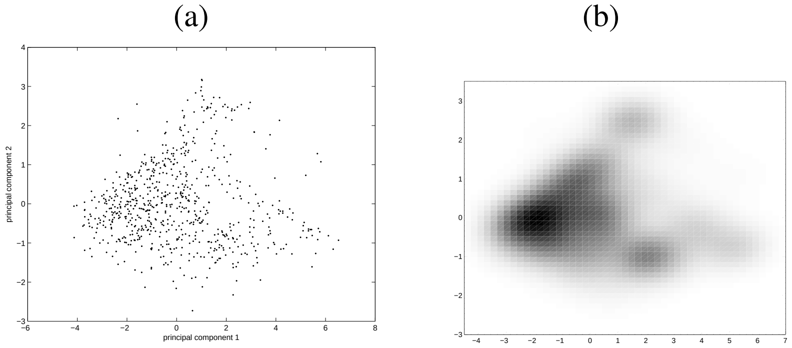

Figure 9: Aircraft data: (a) a scatterplot of the data and (b) a heatplot of the posterior predictive density function where darker colours represent higher density values

5 Discussion

This paper presents an alternative interpretation of the species sampling mixture of normals model often used for Bayesian density estimation which contrast with the more usual subpopulation motivation. The unknown density, f, is treated as the main parameter of interest and prior information is consequently placed directly onto this object. This naturally leads to an alternative parameterisation and prior distribution that are, in certain situations, much easier to specify than previously defined models. It is usual to fix n_0 in the standard conjugate prior distribution and define a value of σ^2 related to the overall variability in the data. In univariate problem, the model can be fitted using a non-informative prior distribution for the scale and location for which the posterior distribution exists. A range of default prior specification are discussed that allow an "automatic" Bayesian density estimator to be chosen. These specifications have good properties over a range of data sets which have a different numbers of modes. Recent developments in computational methods for non-conjugate Dirichlet process and general stick-breaking prior distribution (Neal 2000, Neal and Jain 2006, Papaspiliopoulos and Roberts 2004) make these ideas feasible. However, in common with many other Bayesian methods, non-informative prior distributions can lead to posterior distributions for σ^2 which have long tails. Use of prior information about the location and scale will lead to more concentrated posterior distributions which may be preferable in a some applications. However, the automatic nature of improper priors is often appealing.

This paper has concentrated on density estimation of distributions of observables but many Bayesian applications of nonparametric density estimation of unobservable quantities, such as random effects. The approach developed here can play a more important role in these problems where choices of scale for the component and the distribution of the component means will be hard to choose in many practical applications. The specification describe in this paper allows us to replace a parametric distribution by a nonparametric specification of f whilst retaining the other structure of the parametric model. For example, the univariate analyses presented in this paper directly generalize the standard Bayesian normal model with Jeffreys' prior for the location and scale.

This paper has been restriced mostly to the Dirichlet process mixture of normals model which has been used extensively in the practical applications of Bayesian nonparametric methods. This paper could be generalized in a number of ways. A number of alternative species sampling models have been considered and it would be interesting to see the effect of alternative species sampling models on the inference. An alternative generalisation consider changing either the component specific distribution from normal or, perhaps more usefully, the centring distribution of f. The results in section 2 suggest that the former idea will lead to different prior correlation structures for the density function. Other centring distribution are also possible. A simple method assumes that some lower order prior predictive moments of f are fixed to coincide with those of a parametric distribution. For example, we could replace at t-distribution with a mixture of normals where the the mean of the normals are also drawn from a t-distribution. The results in section 2 make the link between the skewness and kurtosis of the various distributionss explicit.

References

Blackwell, D. and MacQueen, J.B. (1973): "Ferguson distributions via Pólya urn schemes," *Annals of Statistics*, 1, 353-355.

Bowman, A. W. and A. Azzalini (1997): "Applied Smoothing Techniques for Data Analysis," Oxford: Oxford University Press.

Bush, C. A. and S. N. MacEachern (1996): “A Semiparametric Bayesian Model for Randomised Block Designs,” *Biometrika*, 83, 275-285.

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

Ferguson, T. S. (1973): “A Bayesian Analysis of Some Nonparametric Problems,” *The Annals of Statistics*, 1, 209-230.

Ferguson, T. S. (1983): “Bayesian Density Estimation by Mixtures of Normal Distribution,” in *Recent Advances In Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, eds: M. H. Rizvi, J. Rustagi and D. Siegmund, Academic Press: New York.

Gelfand, A. E. and A. Kottas (2002): “A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Proces Mixture Models,” *Journal of Computational and Graphical Statistics*, 11, 289-305.

Ishwaran, H. and James, L. (2001): “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association*, 96, 161-73.

Ishwaran, H. and James, L. F. (2002): “Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information,” *Journal of Computational and Graphical Statistics*, 11, 1-26.

Jain, S. and R. M. Neal (2005): “Splitting and merging components of a nonconjugate Dirichlet process mixture model,” Technical Report 0507, Department of Statistics, University of Toronto.

James, L. F. (2006): “Spatial neutral to the right species sampling mixture models,” prepared for “Festschrift for Kjell Doksum”.

Lijoi, A., R. H. Mena, and I. Prünster (2005): “Hierarchical mixture modelling with normalized inverse-Gaussian priors,” *Journal of the American Statistical Association*, 100, 1278-1291./par

Lo, A. Y. (1984): “On a Class of Bayesian Nonparametric Estimates: I. Density Estimates,” *The Annals of Statistics*, 12, 351-357.

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

Marin, J.-M., K. Mengersen and C. P. Robert (2006): “Bayesian Modelling and Inference on Mixtures of Distributions,” *Handbook of Statistics 25*, (eds: D. Dey and C.R. Rao).

Mengersen, K. and C. Robert (1996): “Testing for mixtures: A Bayesian entropic approach (with dicussion),” in *Bayesian Statistics 5*, eds: J. Berger, J. Bernardo, A. Dawid, D. Lindley and A. Smith, Oxford University Press : Oxford.

Müller, P. and Quintana, F. (2004): "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95-110.

Müller, P. and Rosner, G. (1997): "A Bayesian population model with hierarchical mixture priors applied to blood count data," *Journal of the American Statistical Association*, 92, 1279-1292.

Neal, R. M. (2000): "Markov chain sampling methods for Dirichlet process mixture models," *Journal of COmputational and Graphical Statistics*, 9, 249-265.

Nietro-Barajas, L. E., I Prünster and S. G. Walker (2004): "Normalized random measures driven by increasing additive processes," *Annals of Statistics*, 32, 2343-2360.

Papaspiliopoulos, O. and Roberts, G. (2004): "Retrospective MCMC for Dirichlet process hierarchical models," technical report, University of Lancaster.

Pitman, J. (1996): "Some Developments of the Blackwell-MacQueen Urn Scheme," in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, eds: T. S. Ferguson, L. S. Shapley and J. B. MacQueen, Institue of Mathematical Statistics Lecture Notes.

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

Robert, C. and M. Titterington (1998): "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation," *Statistics and Computing*, 4, 327-355.

Roeder, K. (1990): "Density Estation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Assocation*, 85, 617-624.

Roeder, K. (1994): "A Graphical Technique for Deteremining the Number of Components in a Mixture of Normals," *Journal of the American Statistical Assocation*, 89, 487-495.

Roeder, K. and L. Wasserman (1997): "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894-902.

Walker, S. G., Damien, P., Laud, P. W. and Smith, A. F. M. (1999): "Bayesian nonparametric inference for random distributions and related functions," (with discussion) *Journal of the Royal Statistical Society B*, 61, 485-527.