Bayesian Analysis (2010) **5**, Number 1, pp. 45–64

Default priors for density estimation with mixture models

J.E. Griffin*

Abstract. The infinite mixture of normals model has become a popular method for density estimation problems. This paper proposes an alternative hierarchical model that leads to hyperparameters that can be interpreted as the location, scale and smoothness of the density. The priors on other parts of the model have little effect on the density estimates and can be given default choices. Automatic Bayesian density estimation can be implemented by using uninformative priors for location and scale and default priors for the smoothness. The performance of these methods for density estimation are compared to previously proposed default priors for four data sets.

Keywords: Density Estimation, Dirichlet process mixture models, Mixtures of normals, Normalized Generalized Gamma processes

1 Introduction

Infinite mixture of normals models, and particularly Dirichlet process mixture of normals models, have become the preferred method for density estimation in the Bayesian nonparametric literature (alternative methods are described by Walker et al. (1999) and Mueller and Quintana (2004)). The most widely used infinite mixture of normals model writes the unknown density as

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

where $\mathrm{N}(y|\mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2 and G is a discrete distribution with an infinite number of atoms. This mixing distribution G can also be written as

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

where δ_x is the Dirac measure that places mass 1 on the point x, $\sum_{i=1}^{\infty} w_i = 1$ a.s., $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots$ are i.i.d and the component weights vector $\mathbf{w} = (w_1, w_2, \dots)$ are independent of $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots$ (the assumption of independence is common but dependence is discussed by James et al. (2008)). This is a flexible model which can represent any continuous distribution on the real line. The Bayesian estimation of these models was initially investigated by Ferguson (1983) and Lo (1984). To fit the model a

*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK, mailto:J.E.Griffin-28@kent.ac.uk

 DOI:10.1214/10-BA502

prior needs to be placed on G, the mixing distribution, or alternatively $\mathbf{w}$ and μ_i, σ_i^2. A standard choice is a Dirichlet process prior. Recently, there has been interest in alternative forms of prior for G. Suggestions have included: Poisson-Dirichlet processes (Ishwaran and James 2001), Normalized independent increment processes (Nieto-Barajas et al. 2004) and normalized generalized gamma process (Lijoi et al. 2007). This paper considers the prior on (μ, σ^2). We will see that the prior on $\mathbf{w}$ often has little effect on the estimate of the density, f, with the hierarchical model developed in this paper.

Inference in mixture models can be seen as a model selection problem. Therefore, the prior distribution of μ and σ^2 plays an important role in a Bayesian analysis. However, the prior distribution of $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \ldots$ has received little attention and conjugate choices such as $\mu_i \sim \mathrm{N}(\mu_0, \sigma^2/n_0)$ and $\sigma_i^{-2} \sim \mathrm{Ga}(\alpha, \beta)$, which was suggested by Escobar and West (1995), or the non-conjugate choice $\mu_i \sim \mathrm{N}(\mu_0, \sigma_0^2)$ and $\sigma_i^{-2} \sim \mathrm{Ga}(\alpha, \beta)$ have predominated. It is well-known that the prior distributions cannot be chosen to be uninformative and so the prior introduces a scale through α, β and n_0 or σ_0^2. Escobar and West (1995) suggest interpreting n_0 as a smoothness parameter and this idea will be developed in this paper. Alternatively, several authors have proposed data-based priors which define some of the hyperparameters as a function of the data to allow automatic scaling of the prior. An alternative prior for finite mixture models was proposed by Robert and Titterington (1998) who assume a correlated prior for the means $\mu_1, \mu_2, \mu_3, \ldots, \mu_k$ (where the mixture model has k components) by defining the conditional distribution of μ_j given $\mu_1, \ldots, \mu_{j-1}$. This allows them to place noninformative priors on location and scale of the first component.

The original work of Ferguson (1983) and Lo (1984) assumes that the component variances σ_i^2 in equation (2) share a common value σ^2

$$f(y|\sigma^2) = \int \mathrm{N}(y|\mu, \sigma^2)\, dG(\mu) \tag{3}$$

$$G = \sum_{i=1}^{\infty} w_i \delta_{\mu_i}. \tag{4}$$

This prior has more recently been studied by Ishwaran and James (2002). Typically a hyperprior would be assumed for σ^2 which can be given a vague, proper prior. A drawback with this model is the single variance hyperparameter σ^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. However, the extra modelling introduced by allowing different variances may only be necessary for good predictive performance for certain types of distribution. One aspect of this paper is to consider when the simpler model gives a similar performance to the more complicated model.

This paper proposes a new prior structure for univariate density using infinite mixtures of normals models which use noninformative prior distributions to define a default prior for Bayesian density estimation. Matlab code to implement the methods are available from http://www.kent.ac.uk/ims/personal/jeg28/index.htm

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 briefly discusses the computational methods to fit these models (which are fully described in the appendix), section 4 applies these methods to four previously analysed univariate data sets with different levels of non-normality, section 5 discuss these ideas and some areas for further research and an appendix describes all samplers needed to fit the models and the propriety of the posterior of the proposed model.

2 A hierarchical infinite mixture model

The model in (3) and (4) will be used as a starting point and the form of further levels of the hierarchy will be considered. This seems like a small change to the model but it can have a large effect on the performance of the infinite mixture model for density estimation. The main feature of the proposed hierarchical model is that prior information can be directly placed on the unknown density f and a parameter controlling its smoothness, which contrasts with the standard Bayesian approach which places prior information directly onto the mixing distribution. The structure allows the construction of hierarchical models with little prior information which encourage good predictive performance (as illustrated in section 4) and which can be used to define automatic Bayesian density estimation procedures.

Initially it is assumed that all component variances are equal and a Common Component Variance (CCV) model is defined by

$$y_i|\mu_i \sim \mathrm{N}(\mu_i, a\sigma^2)$$

$$\mu_i \sim G$$

$$G \sim \mathrm{DP}(MH) \tag{5}$$

where $0 < a < 1$ and $\mathrm{DP}(MH)$ represents a Dirichlet process (Ferguson 1973) with mass parameter $M > 0$ and centring distribution $H = \mathrm{N}(\mu_0, (1-a)\sigma^2)$. The parameters μ_0 and σ^2 can be interpreted as the location and scale of the marginal distribution since the prior predictive distribution of y_i is normal with mean μ_0 and variance σ^2.

The parameter a can be interpreted as a measure of the smoothness. If a is close to 1 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 close to zero then the components will have a small variance and the centres will be independent draws which are normally distributed with a variance close to σ^2. The distribution of the number of modes is directly related to the smoothness of the unknown distribution since modes are determined by local features of the realized distribution. Figure 1 shows a number of realized distributions and the distribution of the number of modes for different choices of a and M. The prior distribution of the number of modes may not correspond well with the number of components with non-neglible weight if the mixture contains several

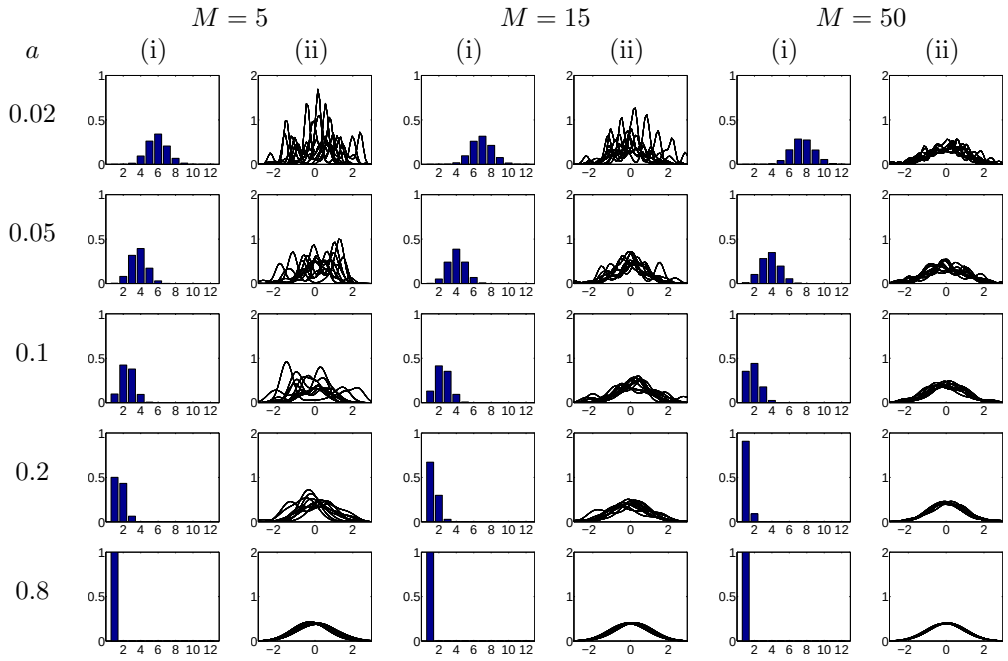

Figure 1: The prior distribution of the number of modes of f in (i) and a sample of densities in (ii) under different hyperparameter choices

components with non-neglible weight which are not well separated. The number of modes for any realized distribution is calculated by finding the number of maxima in the density function. The graphs of the prior distribution of the number of modes indicates that it is more clearly effected by the choice of a rather than M. Values of a between 0.1 and 0.2 indicate a prior belief of bi- or tri-modality whereas $a = 0.02$ indicates support to a number of modes between 3 and 9. These observations are helpful for defining a prior distributions for a. The model in equation (5) is a reparameterisation of the usual conjugate model which has the form

$$y_i|\mu_i \sim \mathrm{N}(\mu_i, \psi)$$

$$\mu_i \sim G$$

$$G \sim \mathrm{DP}(MH)$$

and $H = \mathrm{N}(\mu_0, \frac{\psi}{n_0})$. The new prior distribution is a reparameterisation where $\sigma^2 = \left(\frac{n_0+1}{n_0}\right)\psi$ and $a = \frac{n_0}{n_0+1}$. As noted by Escobar and West (1995) n_0 plays a key role as a smoothing parameter. However, in many application of these methods the parameter n_0 is assumed fixed with a fairly small value implying a prior preference for unsmooth densities. If n_0 is assumed unknown then a prior is usually placed on (ψ, n_0) rather than $\left(\left(\frac{n_0+1}{n_0}\right)\psi, \frac{n_0}{n_0+1}\right)$. An alternative, non-conjugate prior is suggested by Richardson and

Green (1997) who define $H(\mu, \sigma^2) = \mathrm{N}(\zeta, \kappa^{-1})\mathrm{Ga}(\sigma^{-2}|\alpha, \beta)$ in a finite mixture model with a Gamma hyperprior on β.

The Constant Component Variance model can represent any distribution on the real line but if the distribution has several modes with different variability around them then the model will tend to use a few normals to represent the components with larger variances. Consequently, it is useful to have a model which allows different variances to capture these distributions. The extended model is called the Different Component Variance (DCV) model and assumes that

$$y_i|\mu_i \sim \mathrm{N}\left(\mu_i, a\frac{\zeta_i}{\mu_\zeta}\sigma^2\right)$$

$$\mu_i \sim G$$

$$G \sim \mathrm{DP}(MH)$$

where $H = \mathrm{N}(\mu_0, (1-a)\sigma^2)$ and $\mu_\zeta = \mathrm{E}[\zeta_i]$, which is assumed to be finite. An inverse gamma distribution with shape parameter $\phi > 1$ and scale parameter 1 would be a standard choice for the distribution of ζ_i, which is the conditionally conjugate form. It is no longer true that $y_i \sim \mathrm{N}(\mu_0, \sigma^2)$. However, $\mathrm{E}[y_i] = \mu_0$ and $\mathrm{V}[y_i] = \sigma^2$ and the distribution will be close to normal if ϕ is large.

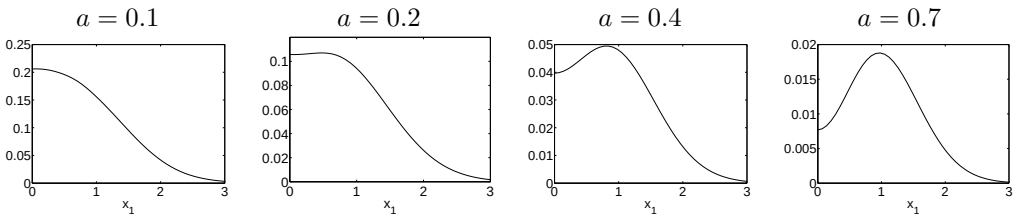

Figure 2: $C(x_1, x_1)$ with a standard normal predictive distribution and various values of a

To look at the effect of the prior for $\mathbf{w}$ and the parameter a, the following quantity is considered

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

where

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

The covariance, $\mathrm{Cov}[f(x_1), f(x_2)]$ can be expressed as the product of two parts. The first part, $C(x_1, x_2)$, is determined by a and the second part, $\sum_{i=1}^{\infty}\mathrm{E}[w_i^2]$, is determined by the choice of prior for $\mathbf{w}$. It follows that $\sum_{i=1}^{\infty}\mathrm{E}[w_i^2] = \frac{1}{M+1}$ if a Dirichlet process mixture is chosen. Figure 2 shows $C(x_1, x_1)$ with various values of a. The variability decreases as the value of a increases but a second effect is also clear: the variability

is monotonic 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.

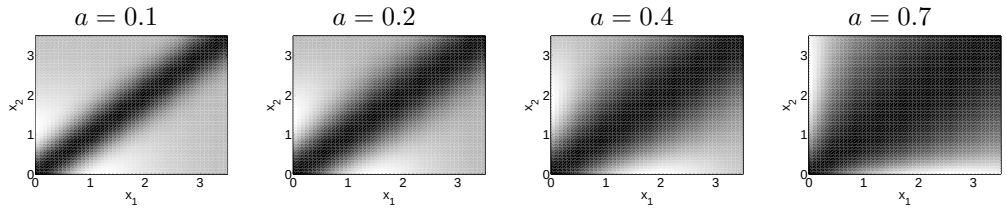

Figure 3: Prior correlation between the density values at two points x_1 and x_2 for a model with a standard normal predictive distribution and various values of a where darker colours represent larger correlations

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

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

The correlation structure of $f(x)$ is independent of the choice of prior for $\mathbf{w}$. Therefore we can interpret a as a correlation parameter. Figure 3 shows the autocorrelation structure for various values of a with darker regions representing stronger correlation. 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 x are associated with much larger ranges (the distance at which the autocorrelation is equal to some small prespecified value). 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 the Dirichlet process mass parameter 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 nonparametric priors. 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 Specification of hyperparameters

The parameterization discussed in the previous section suggests placing independent priors on the location, scale and the smoothness of the unknown distribution. As Mengersen and Robert (1996) noted 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. There are two standard choices of prior for the location μ_0 and the scale σ^2: the improper Jeffreys' prior $p(\mu_0, \sigma^2) \propto \sigma^{-2}$ and the conjugate choice $p(\mu_0, \sigma^{-2}) = \text{N}(\mu|\mu_{00}, n_{00}\sigma^2)\text{Ga}(\sigma^{-2}|\alpha, \beta)$. The second choice is proper and so leads to a proper posterior and a proof of the propriety of the posterior with improper prior is

given in Appendix B.

The hyperparameter a is restricted to $(0, 1)$ and we consider a Beta prior distribution. This represents beliefs about the smoothness of the unknown density. Priors which places a lot of mass on small values of a would correspond to a strong prior belief that the unknown density is unsmooth. The mass parameter of the Dirichlet process, M, is given a prior suggested by Griffin and Steel (2004),

$$p(M) = \theta^{\eta} \frac{\Gamma(2\eta)}{(\Gamma(\eta))^2} \frac{M^{\eta-1}}{(M+\theta)^{2\eta}}.$$

where θ can be interpret as a prior sample size and η is a variance parameter for which the prior becomes more concentrated with larger η. They suggest using $\eta = 3$ as a default value.

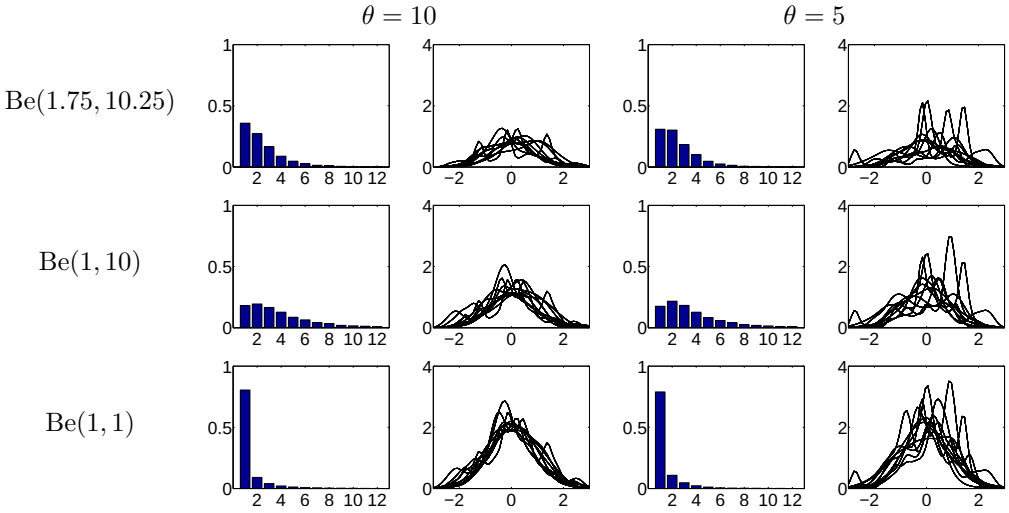

Figure 4: The prior distribution of the number of modes of F and a sample of densities under different hyperparameter choices

Figure 4 shows realisations of the density and the distribution of the number of modes for various choice of the prior distribution of a and M. If we choose a to follow a uniform distribution (Be$(1, 1)$) 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. 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. A compromise between these priors is given by a Be$(1.75, 10.25)$ prior which implies a modal number of modes of 1 but with a wider spread than the Be$(1, 1)$.

3 Computational methods

The fitting of Dirichlet process mixture models have been greatly helped by the development of efficient MCMC methods. The model with a Common Component Variance model is a conjugate Dirichlet process which can be fitted using standard methods (MacEachern 1998; Neal 2000) and the Different Component Variance model is non-conjugate and can be fitted using the methods described in Neal (2000) (his algorithm 8 was used to implement the examples in this paper). Inference about the unknown mixing distribution and the density f is possible using the method of Gelfand and Kottas (2002). These methods are specific to Dirichlet process priors and computational approaches for more general priors are developed by Papaspiliopoulos and Roberts (2008), Kalli et al. (2010) and Griffin and Walker (2010). All methods make use of the Gibbs sampler and the full conditional distribution for each parameter are fully described in each paper. A full description of the algorithms needed to fit the models is given in the appendix.

4 Examples

The Bayesian model developed in this paper will be illustrated on a series of data sets previously analysed in the literature: the galaxy data, acidity data, enzyme data and sodium lithium data. The "galaxy" data was initially analysed by Roeder (1990) and introduced into the Bayesian literature by Roeder and Wasserman (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. The "acidity" data refers to a sample of 155 acidity index measurement made on links in north-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). The "sodium lithium" data was previously analysed by Roeder (1994) and measures the cell sodium-lithium countertransport (SLC) in six large English kindreds. 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 $\theta = 5$ 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. All MCMC samplers were run 50000 iterations with the first 5000 iterations used as a burn-in period, which seemed sufficient for convergence of the number of clusters.

4.1 Results

Figure 5 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 distributions described in

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

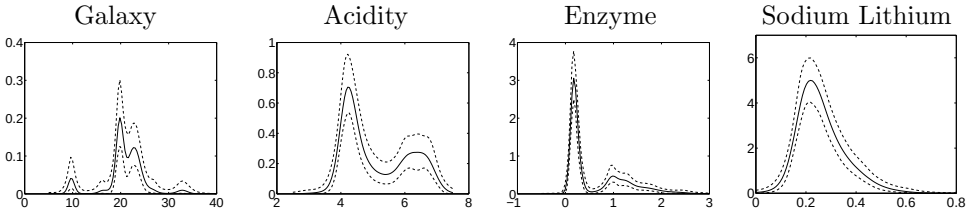

Figure 5: Posterior predictive densities for the four data sets with a pointwise 95% HPD interval

section 2.1). These results are very similar to previous analyses, although the galaxy data results do differ largely from analyses described in Marin et al. (2006) and Roeder and Wasserman (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 a number of other analyses *e.g.* Richardson and Green (1997).

	a		M	
	Be$(1,1)$	Be$(1,10)$	Be$(1,1)$	Be$(1,10)$
Ga	0.04 *(0.01, 0.12)*	0.03 *(0.01, 0.10)*	3.73 *(1.14, 10.80)*	3.93 *(1.31, 10.06)*
Ac	0.16 *(0.04, 0.46)*	0.10 *(0.03, 0.27)*	3.47 *(0.95, 10.66)*	3.23 *(0.83, 9.48)*
En	0.06 *(0.01, 0.23)*	0.05 *(0.01, 0.16)*	2.40 *(0.75, 6.40)*	2.39 *(0.69, 7.31)*
S L	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 (Ga is Galaxy, Ac is Acidity, En is Enzyme and S L is Sodium Lithium) and two priors for a

Table 2 shows summaries of the posterior distributions of a and M under the two prior distributions of a. The results show that the distributions 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. This supports the interpretation of a as a smoothness parameter. The posterior distribution is robust to the choice between the two prior distribution when the densities 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 all 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.

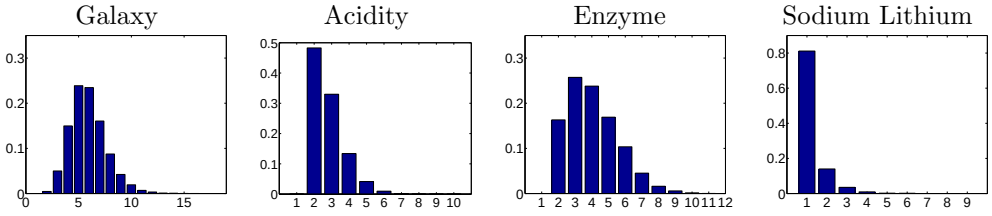

Figure 6: Posterior distribution of the number of modes for the four data sets

The inferences about the number of modes is shown in figure 6. 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 the 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 shows most support for 3 modes with good support for up to 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 distributions. 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.

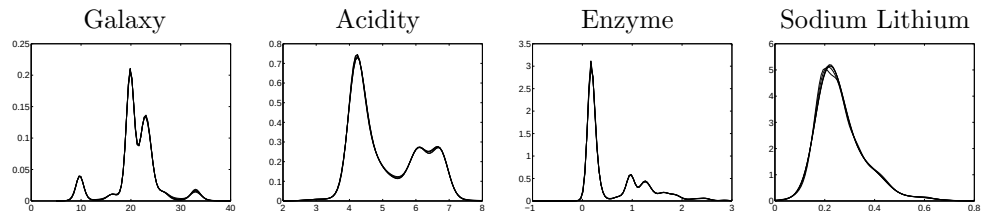

Figure 7: Posterior predictive densities using the CCV model for the four data sets using a Normalized Generalized Gamma prior with $\kappa = 0.1, 0.2, 0.3, 0.4, 0.5$

The priors considered so far are based on the Dirichlet process. Other priors could

be considered for the mixing distribution and we would hope that the results are relatively robust to the choice of this prior. To investigate the robustness of the proposed model, the normalized generalized gamma process is used as an alternative class (Lijoi et al. 2007) of prior for the nonparametric mixing distribution G. The prior has two parameters $0 < \kappa < 1$ and $M > 0$. The Dirichlet process with mass parameter M arises as a special case as $\kappa \to 0$ and the Normalized Inverse Gaussian process (Lijoi et al. 2005) arises when $\kappa = 0.5$. The CCV model was fitted with M given a Gamma prior distribution. The predictive distribution for the four data sets are shown in Figure 7 with Normalized Generalized Gamma process for the $\kappa = 0.1, 0.2, 0.3, 0.4, 0.5$. Clearly, the density estimate is robust to the choice of prior for $\mathbf{w}$.

4.2 Comparison to other default priors

Several authors have proposed default priors for mixture models which use summaries of the data (such as mean or variance) to define certain hyperparameters. They work with general model

$$y_i|\mu_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$$
$$(\mu_i, \sigma_i^2) \sim G$$
$$G \sim \mathrm{DP}(MH)$$

where the centring measure H for μ and σ^2 is to be specified. Richardson and Green (1997) suggest using the choice $H(\mu, \sigma^{-2}) = \mathrm{N}\left(\frac{a+b}{2}, R^2/\epsilon\right) \mathrm{Ga}(\alpha, \beta)$ where $b = \max\{y_j\}$, $a = \min\{y_j\}$ and R is the range of the data. The hyperparameter ϵ is chosen small (they suggest the default value $\epsilon = 1$ in their paper) and $\beta \sim \mathrm{Ga}(g, \epsilon_2/R^2)$. They suggest that $g < 1 < \alpha$. This prior on β allows the model to adapt to the scale of the data by sharing information between different component variances σ_j^2. They suggest $\alpha = 2$, $g = 0.2$, $\epsilon_2 = 10$ as default values in their examples.

Alternatively, Ishwaran and James (2002) propose $H(\mu, \sigma^2) = \sim \mathrm{N}(\mu|\theta, \sigma_\mu^2)\mathrm{U}(\sigma^2|0, T)$ where $\theta \sim \mathrm{N}(0, A)$. The hyperparameter A is chosen large to represent a vague prior (the authors use $A = 1000$). The parameters T and σ_μ are scale parameters are they suggest taking σ_μ to equal 4 times the standard deviation of the data and taking T to equal the variance of the data as default values.

These two default priors are compared to the CCV and DCV models proposed in this paper. The different default priors are compared using a random fold cross-validation method with a posterior predictive criteria. The score for prior M is calculated as

$$S(A) = \frac{1}{Km} \sum_{i=1}^{K} \sum_{j=1}^{m} \log p_M(y_{\gamma_{ij}}|y_{-\gamma_i})$$

where p_M represents the posterior predictive distribution under prior M, γ_i is the first m elements of a random permutation of $\{1, 2, \ldots, n\}$ and $y_{-\gamma_i}$ represents the vector y with the elements γ_i removed. The use of cross-validation methods to compare predictive performance is discussed by Gneiting and Raftery (2007). Larger values of the score show better predictive performance.

	CCV	DCV	RG	IJ
Galaxy	-2.50	-2.49	-2.57	-2.65
Enzyme	-0.29	-0.27	-0.46	-0.69
Log Acidity	-1.13	-1.13	-1.22	-1.22
Sodium Lithium	1.01	1.01	0.78	0.92

Table 3: Log predictive scores for the four data sets with four default priors: Common Component Variances (CCV) model, Different Component Variances (DCV) model, Richardson and Green (RG) and Ishwaran and James (IJ)

The scores for the four different data sets are shown in Table 3. The difference between the CCV model and DCV model are small for all each data set except the enzyme data where the DCV model outperforms the CCV model. Clearly, the distribution of the data, which has two modes where the density around one mode is much more spread than the other. This is a case where we would expect the DCV model to perform better. Both priors outperform the other default priors across all four data sets. In fact, the cross-validation performance for the galaxy data is slightly better than the informative prior of Escobar and West (1995) who carefully choose hyperparameters for their problem. This indicates that the CCV and DCV represent effective priors for Bayesian density estimation.

5 Discussion

This paper presents an alternative parameterisation of the infinite mixtures of normals model often used for Bayesian density estimation. The unknown density, f, is treated as the main parameter of interest and prior information is placed directly onto this object. This naturally leads to an alternative parameterisation and prior distributions that are, in certain situations, much easier to specify than previously defined models. In univariate problem, the model can be fitted using a non-informative prior distribution for the scale and location. A range of default prior specification are discussed that define "automatic" Bayesian density estimator to be chosen. These specifications have good properties over a range of data sets compared to other default schemes.

There are several directions for future research. All examples involve estimating densities of observables but an appealing aspect of a Bayesian approach is that these method can be applied to density estimation of unobservable quantities, such as random effects (Mueller and Rosner 1997). Previously proposed default methods use the observed values of the data to set some hyperparameter which becomes more challenging when modelling unobversable quantities. The approach taken in this paper allows the nonparametric prior to be centred directly on the standard parametric distribution. The parameter a then becomes are measure of the departure of the density from the parametric choice. A main motivation for this alternative approach is placing prior information on the unknown distribution F rather than the centring distribution H. A prior is placed directly on the mean and variance of F rather than the more common

choice of directly placing a prior on H. This approach could be extended to higher moments of F to allow skewness or kurtosis to be modelled, which would follow from allowing H to be skewed or heavy-tailed itself. The prior would then be placed on the centred moments of F rather than H.

References

Escobar, M. D. and West, M. (1995). "Bayesian density-estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. 46, 48, 56

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics*, 1: 209–230. 47

— (1983). "Bayesian Density Estimation by Mixtures of Normal Distribution." In Rizvi, M. H., Rustagi, J., and Siegmund, D. (eds.), *Recent Advances In Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*. Academic Press: New York. 45, 46

Gelfand, A. E. and Kottas, A. (2002). "A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 289–305. 52

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction and estimation." *Journal of the American Statistical Association*, 102: 359–378. 55

Griffin, J. E. and Steel, M. F. J. (2004). "Semiparametric Bayesian Inference for Stochastic Frontier Models." *Journal of Econometrics*, 123: 121–152. 51

Griffin, J. E. and Walker, S. G. (2010). "Posterior Simulation of Normalized Random Measure Mixtures." *Journal of Computational and Graphical Statistics*, forthcoming. 52

Ishwaran, H. and James, L. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96: 161–73. 46

Ishwaran, H. and James, L. J. (2002). "Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information." *Journal of Computational and Graphical Statistics*, 11: 508–532. 46, 55

James, L. F., Lijoi, A., and Pruenster, I. (2008). "Posterior Analysis for Normalized Random Measures with Independent Increments." *Scandinavian Journal of Statistics*, 36: 76–97. 45

Kalli, M., Griffin, J. E., and Walker, S. G. (2010). "Slice Sampling Mixture Models." *Statistics and Computing*, forthcoming. 52

Lijoi, A., Mean, R. H., and Pruenster, I. (2007). "Controlling the reinforcement in Bayesian non-parametric mixture models." *Journal of the Royal Statistical Society B*, 69: 715–740. 46, 55

Lijoi, A., Mena, R. H., and Pruenster, I. (2005). "Hierarchical mixture modelling with normalized inverse-Gaussian priors." *Journal of the American Statistical Association*, 100: 1278–1291. 55

Lo, A. Y. (1984). "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates." *The Annals of Statistics*, 12: 351–357. 45, 46

MacEachern, S. N. (1998). "Computational Methods for Mixture of Dirichlet Process Models." In Dey, D., Mueller, P., and Sinha, D. (eds.), *Practical Nonparametric and Semiparametric Bayesian Statistics*, 23–44. Springer-Verlag. 52

Marin, J.-M., Mengersen, K., and Robert, C. P. (2006). "Bayesian Modelling and Inference on Mixtures of Distributions." In Dey, D. and Rao, C. R. (eds.), *Handbook of Statistics 25*. Elsevier. 53

Mengersen, K. and Robert, C. (1996). "Testing for mixtures: A Bayesian entropic approach (with discussion)." In Berger, J., Bernardo, J., Dawid, A., Lindley, D., and Smith, A. (eds.), *Bayesian Statistics 5*. Oxford University Press : Oxford. 50

Mueller, P. and Quintana, F. (2004). "Nonparametric Bayesian Data Analysis." *Statistical Science*, 19: 95–110. 45

Mueller, 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. 56

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9: 249–265. 52, 61

Nieto-Barajas, L. E., Pruenster, I., and Walker, S. G. (2004). "Normalized random measures driven by increasing additive processes." *Annals of Statistics*, 32: 2343–2360. 46

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95: 169–186. 52

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with unknown number of components (with discussion)." *Journal of the Royal Statistical Society B*, 59: 731–792. 48, 52, 53, 55

Robert, C. and Titterington, M. (1998). "Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation." *Statistics and Computing*, 4: 327–355. 46

Roeder, K. (1990). "Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies." *Journal of the American Statistical Assocation*, 85: 617–624. 52

— (1994). "A Graphical Technique for Determining the Number of Components in a Mixture of Normals." *Journal of the American Statistical Assocation*, 89: 487–495. 52

Roeder, K. and Wasserman, L. (1997). "Practical Bayesian Density Estimation Using Mixtures of Normals." *Journal of the American Statistical Association*, 92: 894–902. 52, 53

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. 45

Appendix A: MCMC algorithms

A.1: Common Component Variances model

The full model is

$$y_i \sim \mathrm{N}(\mu_i, a\sigma^2), \qquad i = 1, \ldots, n$$

$$\mu_i \sim G,$$

$$G \sim \mathrm{DP}(MH)$$

where $H(\mu) = \mathrm{N}(\mu|\mu_0, (1-a)\sigma^2)$,

$$\mu_0 \sim \mathrm{N}(\mu_{00}, \lambda_0^{-1}), \qquad \sigma^{-2} \sim \mathrm{Ga}(s_0, s_1), \qquad a \sim \mathrm{Be}(a_0, a_1)$$

The full conditionals for the uninformative choice $p(\mu_0, \sigma^2) \sim \sigma^{-2}$ arises by taking $s_0 = 0$, $s_1 = 0$ and $\lambda_0 = 0$ in the following formulae. As with any mixture model, latent variables $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ are introduced to help implement the MCMC samplers. This model can be sampled using a standard Gibbs sampler for a conjugate Dirichlet process mixture model. Let $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(K)}$ be the distinct values of $\mu_1, \mu_2, \ldots, \mu_n$, $n_k = \sum_{j=1}^{n} \mathrm{I}(s_j = k)$, $n_k^{-i} = \sum_{i=1;j\neq i}^{n} \mathrm{I}(s_i = k)$ and K^{-i} be the number of distinct values excluding μ_i.

Updating s

The elements of $\mathbf{s}$ are updated from their full conditional which is discrete

$$p(s_i = k) \propto \begin{cases} \frac{n_k^{-i}}{\sqrt{a}} \exp\left\{-\frac{1}{2a\sigma^2}(y_i - \mu_{(k)})^2\right\} & 1 \leq k \leq K^{-i} \\ M \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu_0)^2\right\} & k = K^{-i} + 1 \end{cases}$$

and can be sampled using inversion sampling.

Updating $\mu_{(i)}$

The full conditional distribution of $\mu_{(i)}$ is $\mathrm{N}(\mu_i^\star, \sigma_i^{2\star})$ where

$$\mu_i^\star = \frac{\frac{\sum_{\{j|s_j=i\}} y_j}{a} + \frac{\mu_0}{1-a}}{\frac{n_i}{a} + \frac{1}{1-a}} \text{ and } \sigma_i^{2\star} = \frac{\sigma^2}{\frac{n_i}{a} + \frac{1}{1-a}}.$$

Updating μ_0

The full conditional distribution of μ_0 is $\mathrm{N}(\mu^\star, \sigma^{2\star})$ where

$$\mu^\star = \left(\frac{\sigma^{-2}\sum_{i=1}^K \mu_{(i)}}{1-a} + \lambda_0\mu_{00}\right) \Big/ \left(\frac{\sigma^{-2}K}{1-a} + \lambda_0\right) \text{ and } \sigma^{2\star} = 1 \Big/ \left(\frac{\sigma^{-2}K}{1-a} + \lambda_0\right).$$

Updating σ^2

The full conditional distribution of σ^{-2} is

$$\mathrm{Ga}\left(s_0 + (n+K)/2, s_1 + \frac{1}{2}\left[\frac{\sum_{i=1}^n (y_i - \mu_{(s_i)})^2}{a} + \frac{\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2}{1-a}\right]\right).$$

Updating a

The full conditional distribution of a is proportional to

$$a^{-n/2}(1-a)^{-k/2} \exp\left\{-\frac{1}{2}\sigma^{-2}\left[\frac{\sum_{i=1}^n (y_i - \mu_{(s_i)})^2}{a} + \frac{\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2}{1-a}\right]\right\}.$$

Let $z = \frac{a}{1-a}$ then the full conditional of z is proportional to

$$(1+z)^{(n+K)/2} z^{-n/2} \exp\left\{-\frac{1}{2}\sigma^{-2}\left[\frac{1}{z}\sum_{i=1}^n (y_i - \mu_{(s_i)})^2 + z\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2\right]\right\}.$$

If $n+K$ is even then the distribution can be expressed as a mixture of generalized inverse Gaussian distributions which is represented as $\mathrm{GIG}(\lambda, \chi, \psi)$ which has density

$$g(x) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\chi\psi})} x^{\lambda-1} \exp\left\{-\frac{1}{2}\left(\frac{\chi}{x} + \psi x\right)\right\}, \qquad x > 0$$

where $K_\lambda(y)$ is the modified Bessel function of the third kind. The mixture distribution has the form

$$\sum_{j=1}^{(n+K)/2} w_j \,\mathrm{GIG}\left(x \,\middle|\, j - n/2 + 1, \frac{\sum_{i=1}^n (y_i - \mu_{(s_i)})^2}{\sigma^2}, \frac{\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2}{\sigma^2}\right)$$

where

$$w_j = \left(\begin{array}{c} (n+K)/2 \\ j \end{array} \right) \frac{2K_{j-n/2+1}\left(\sqrt{\left(\sum_{i=1}^n (y_i - \mu_{(s_i)})^2\right)\left(\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2\right)\Big/ \sigma^4}\right)}{\left(\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2 \Big/ \sum_{i=1}^n (y_i - \mu_{(s_i)})^2\right)^{(j-n/2+1)/2}}.$$

If $n+K$ is odd we use the rejection envelope

$$\sum_{j=1}^{(n+K)/2+1} w_j \, \text{GIG}\left(x \left| j - n/2 + 1, \frac{\sum_{i=1}^n (y_i - \mu_{(s_i)})^2}{\sigma^2}, \frac{\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2}{\sigma^2}\right.\right)$$

where

$$w_j = \left(\begin{array}{c} (n+K)/2 \\ j \end{array} \right) \frac{2K_{j-n/2+1}\left(\sqrt{\left(\sum_{i=1}^n (y_i - \mu_{(s_i)})^2\right)\left(\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2\right)\Big/ \sigma^4}\right)}{\left(\sum_{i=1}^K (\mu_{(i)} - \mu_0)^2 \Big/ \sum_{i=1}^n (y_i - \mu_{(s_i)})^2\right)^{(j-n/2+1)/2}}$$

and a point, z, simulated from this mixture is accepted with probability $(1+z)^{-1/2}$. Then $a = \frac{z}{1+z}$ in either case.

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.

A.2: Different Component Variances model

The full model is

$$y_i \sim \text{N}\left(\mu_i, a(\phi-1)\zeta_i\sigma^2\right), \qquad i = 1, \ldots, n$$
$$(\mu_i, \zeta_i) \sim G,$$
$$G \sim \text{DP}(MH)$$

where $H(\mu, \zeta^{-1}) = \text{N}(\mu|\mu_0, (1-a)\sigma^2)\text{Ga}(\zeta^{-1}|\phi, 1)$,

$$\mu_0 \sim \text{N}(\mu_{00}, \lambda_0^{-1}), \qquad \sigma^{-2} \sim \text{Ga}(s_0, s_1), \qquad a \sim \text{Be}(a_0, a_1)$$

The full conditionals for the uninformative choice $p(\mu_0, \sigma^2) \sim \sigma^{-2}$ arises by taking $s_0 = 0$, $s_1 = 0$ and $\lambda_0 = 0$ in the following formulae. As with any mixture model, latent variables $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ are introduced to help implement the MCMC samplers. This model is a nonconjugate Dirichlet process mixture model and can be sampled using algorithm 8 of Neal (2000). The updating of M is the same as the updating in

the common component variance model. Let $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(K)}$ be the distinct values of $\mu_1, \mu_2, \ldots, \mu_n$ and $\zeta_{(1)}, \zeta_{(2)}, \ldots, \zeta_{(K)}$ be the distinct values of $\zeta_1, \zeta_2, \ldots, \zeta_n$. Let $n_k = \sum_{j=1}^{n} \mathrm{I}(s_j = k)$, $n_k^{-i} = \sum_{i=1;j\neq i}^{n} \mathrm{I}(s_i = k)$.

Updating s

To update s_i we choose a value of m (usually 3 or 4) which is fixed for the whole algorithm and uses the following step.

- If there is a j such that $s_j = s_i$ for $j \neq i$ then

  $$w_j = n_j^{-i} \text{ for } 1 \leq j \leq K \text{ and } w_j = M/m \text{ for } K+1 \leq j \leq K+m$$

  and simulate $\mu_{(i)} \sim \mathrm{N}(\mu_0, (1-a)\sigma^2)$ and $\zeta_{(i)}^{-1} \sim \mathrm{Ga}(\phi, 1)$ for $K+1 \leq i \leq K+m$.

- Otherwise,

  $$w_j = n_j^{-i} \text{ for } 1 \leq j \leq K \text{ and } j \neq s_i, w_{s_i} = M/m \text{ and } K+1 \leq j \leq K+m-1$$

  and simulate $\mu_{(i)} \sim \mathrm{N}(\mu_0, (1-a)\sigma^2)$ and $\zeta_{(i)}^{-1} \sim \mathrm{Ga}(\phi, 1)$ for $K+1 \leq i \leq K+m-1$.

Then the elements of $\mathbf{s}$ are updated from their full conditional which is discrete

$$p(s_i = j) \propto w_j \sqrt{\frac{1}{\zeta_{(j)}}} \exp\left\{ -\frac{(y_i - \mu_{(j)})^2}{2a\sigma^2(\phi-1)\zeta_{(j)}} \right\}$$

and can be sampled using inversion sampling. Finally, any cluster which does not have a point allocated to it is deleted.

Updating $\mu_{(i)}$

The full conditional distribution of $\mu_{(i)}$ is $\mathrm{N}(\mu_i^\star, \sigma_i^{2\star})$ where

$$\mu_i^\star = \frac{\frac{\sum_{\{j|s_j=i\}} y_j}{(\phi-1)\zeta_{(i)}a} + \frac{\mu_0}{1-a}}{\frac{n_i}{(\phi-1)\zeta_{(i)}a} + \frac{1}{1-a}} \text{ and } \sigma_i^{2\star} = \frac{\sigma^2}{\frac{n_i}{(\phi-1)\zeta_{(i)}a} + \frac{1}{1-a}}.$$

Updating μ_0

The full conditional distribution of μ_0 is $\mathrm{N}(\mu^\star, \sigma^{2\star})$ where

$$\mu^\star = \left(\frac{\sigma^{-2}\sum_{i=1}^{K} \mu_{(i)}}{1-a} + \lambda_0\mu_{00} \right) \Big/ \left(\frac{\sigma^{-2}K}{1-a} + \lambda_0 \right) \text{ and } \sigma^{2\star} = 1 \Big/ \left(\frac{\sigma^{-2}K}{1-a} + \lambda_0 \right).$$

Updating σ^2

The full conditional distribution of σ^{-2} is

$$\text{Ga}\left(s_0 + (n+K)/2, s_1 + \frac{1}{2}\left[\sum_{i=1}^{n}\frac{(y_i-\mu_{(s_i)})^2}{(\phi-1)a\zeta_{(s_i)}} + \frac{\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2}{1-a}\right]\right).$$

Updating a

The parameter a can be updated using mixtures of generalized inverse Gaussian distributions in a similar way to the updating in the common component variance model. If $n+K$ is even, the mixture distribution has the form

$$\sum_{j=1}^{(n+K)/2} w_j \,\text{GIG}\left(x \left| j-n/2+1, \sum_{i=1}^{n}\frac{(y_i-\mu_{(s_i)})^2}{(\phi-1)\zeta_{(s_i)}\sigma^2}, \frac{\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2}{\sigma^2}\right.\right)$$

where

$$w_j = \left(\begin{array}{c}(n+K)/2\\ j\end{array}\right) \frac{2K_{j-n/2+1}\left(\sqrt{\left(\sum_{i=1}^{n}b_i\right)\left(\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2\right)\Big/\sigma^4}\right)}{\left(\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2\Big/\sum_{i=1}^{n}b_i\right)^{(j-n/2+1)/2}}$$

and

$$b_i = (y_i-\mu_{(s_i)})^2/[(\phi-1)\zeta_{(s_i)}], \qquad 1\le i\le n.$$

If $n+K$ is odd we use the rejection envelope

$$\sum_{j=1}^{(n+K)/2+1} w_j \,\text{GIG}\left(x \left| j-n/2+1, \sum_{i=1}^{n}\frac{(y_i-\mu_{(s_i)})^2}{(\phi-1)\zeta_{(s_i)}\sigma^2}, \frac{\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2}{\sigma^2}\right.\right)$$

where

$$w_j = \left(\begin{array}{c}(n+K)/2\\ j\end{array}\right) \frac{2K_{j-n/2+1}\left(\sqrt{\left(\sum_{i=1}^{n}b_i\right)\left(\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2\right)\Big/\sigma^4}\right)}{\left(\sum_{i=1}^{K}(\mu_{(i)}-\mu_0)^2\Big/\sum_{i=1}^{n}b_i\right)^{(j-n/2+1)/2}}$$

and

$$b_i = (y_i-\mu_{(s_i)})^2/[(\phi-1)\zeta_{(s_i)}], \qquad 1\le i\le n.$$

A point, z, simulated from this mixture is accepted with probability $(1+z)^{-1/2}$. Then $a=\frac{z}{1+z}$ in either case.

Updating ζ

The full conditional distribution of $\zeta_{(i)}$ is $\text{IG}\left(\phi+\frac{1}{2}n_i, 1+\frac{1}{2}\frac{\sum_{\{j|s_j=i\}}(x_j-\mu_{(i)})^2}{(\phi-1)a\sigma^2}\right)$.

B: Proof of a proper posterior for CCV model

To show that the posterior of the CCV model is proper (a similar approach is possible for the DCV model), we need to check that

$$p(y) = \sum_s \int p(y|\theta, \mu_0, a, \sigma^2) p(\theta) p(\mu_0) p(\sigma^2) p(a)\, d\theta d\mu_0 d\sigma^2 da < \infty.$$

It suffices to check that $p(y|a, s) < \infty$ for all s and a since $p(s)$ and $p(a)$ are proper prior distributions. Suppose that the allocations $s_1, s_2, \ldots, s_n$ have K distinct values and let $S_k = \{i|s_i = k\}$ then

$$\begin{aligned} p(y, \theta, \mu_0, \sigma^2|a, s) =& (2\pi)^{-(n+k)/2} \sigma^{-(n+k+2)/2} a^{-n/2} (1-a)^{-k/2} \\ & \times \exp\left\{ -\frac{\sigma^{-2}}{2} \left[\sum_{k=1}^{K} \sum_{i \in S_k} \frac{(y_i - \theta_k)^2}{a} + \sum_{k=1}^{K} \frac{(\theta_k - \mu_0)^2}{1-a} \right] \right\} \end{aligned}$$

and integrating across $\theta_1, \theta_2, \ldots, \theta_K$ gives

$$\begin{aligned} \int p(y, \theta, \mu_0, \sigma^2|\theta, a, s) d\theta =& (2\pi)^{-n/2} \sigma^{-(n+2)/2} a^{-n/2} (1-a)^{-k/2} \prod_{k=1}^{K} b_k^{-1/2} \\ & \times \exp\left\{ -\frac{\sigma^{-2}}{2} \left[\frac{1}{a} \sum_{k=1}^{K} \sum_{i \in S_k} y_i^2 + K \frac{\mu_0^2}{1-a} - \sum_{k=1}^{K} \frac{c_k^2}{b_k} \right] \right\} \end{aligned}$$

where $b_k = \frac{n_k}{a} + \frac{1}{1-a}$ and $c_k = \frac{\mu_0}{1-a} + \frac{\sum_{i \in S_k} y_i}{a}$. Integrating across μ_0 gives

$$\begin{aligned} \int p(y, \theta, \mu_0, \sigma^2, |a, s) d\theta d\mu_0 =& (2\pi)^{-n/2} \sigma^{-(n+1)/2} a^{-n/2} (1-a)^{-k/2} d^{-1/2} \prod_{k=1}^{K} b_k^{-1/2} \\ & \times \exp\left\{ -\frac{\sigma^{-2}}{2} \left[\frac{1}{a} \sum y_i^2 - \sum_{k=1}^{K} \frac{(\sum_{i \in S_k} y_i)^2}{b_k a^2} - \frac{e^2}{d} \right] \right\} \end{aligned}$$

where $d = \sum_{k=1}^{K} (\frac{1}{1-a} - \frac{1}{(1-a)^2 b_k})$ and $e = \sum_{k=1}^{K} \frac{\sum_{i \in S_k} y_i}{b_k a (1-a)}$ and clearly

$$\int p(y|\theta, \mu_0, a, \sigma^2) p(\theta) p(\mu_0) p(\sigma^2)\, d\theta d\mu_0 d\sigma^2 < \infty.$$