Bayesian multivariate density estimation for observables and random effects

J.E. Griffin*
School of Mathematics, Statistics and Actuarial Science,
University of Kent, Canterbury, CT2 7NF, U.K.

June 9, 2011

Abstract

Multivariate density estimation is approached using Bayesian nonparametric mixture of normals models. Two models are developed which are both centred over a multivariate normal distribution but make different prior assumptions about how the unknown distribution departs from a normal distribution. The priors are applied to density estimation of both observables and random effects (or other unobservable random quantities). Markov chain Monte Carlo methods are described for estimation of all models. The models are applied to density estimation for observables and the application of a nonparametric linear mixed model to repeated cholesterol measurements from the Framingham study.

Keywords: Dirichlet process mixture models, Mixtures of normals, Adaptive Monte Carlo, Centring.

1 Introduction

The use of Dirichlet process mixture (DPM) models in density estimation has become increasingly popular in Bayesian statistics, following the seminal work of Lo (1984)

*Corresponding author: Jim Griffin, School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K. Tel.: +44-1227-; Fax: +44-1227-; Email: J.E.Griffin-28@kent.ac.uk.

and Ferguson (1983). The mixture of normals has become the main example and is often combined with a conjugate prior as described by Escobar and West (1995). Suppose that $y_1, y_2, \ldots, y_n$ are a sample of p-dimensional vectors. The model has the following hierarchical form for multivariate observations $y_1, y_2, \ldots, y_n$

$$y_i \sim \mathrm{N}(\theta_i, \Sigma_i),$$

$$(\theta_i, \Sigma_i) \sim G,$$

$$G \sim \mathrm{DP}(M, H)$$

where $\mathrm{N}(\theta, \Sigma)$ represents a multivariate normal distribution with mean θ and covariance Σ, $\mathrm{DP}(M, H)$ represents a Dirichlet process (Ferguson, 1973) with mass parameter M and centring distribution H. In this case, the centring distribution H is chosen to be $\mathrm{N}(\theta|\theta_0, n_0\Sigma)\mathrm{IW}(\Sigma|\alpha_\Sigma, \beta_\Sigma)$ where IW represents an inverse Wishart distribution with shape parameter α_Σ and mean $\beta_\Sigma/(\alpha_\Sigma - p - 1)$. The marginal distribution of y_i is then an infinite mixture of normals distribution. The hyperparameter θ_0 is straightforward to choose since it is the prior mean of y_i. However, it is less clear how α_Σ, β_Σ and n_0 should be chosen. In practice, their value has an important effect on the form of the density estimate is often chosen by a process of "trial-and-error".

In univariate density estimation with mixtures of normals, Griffin (2010) suggests a simplified version of the conjugate DPM model where

$$y_i \sim \mathrm{N}(\theta_i, a\sigma^2), \qquad 0 < a < 1, \tag{1}$$

$$\theta_i \sim G,$$

$$G \sim \mathrm{DP}(M, H)$$

and now $H = \mathrm{N}(\theta|\mu, (1-a)\sigma^2)$. This model implies that the prior predictive distribution of y_i is normal with mean μ and variance σ^2. The parameter a can be interpreted as a smoothness parameter with larger values of a leading to a prior which places increasing mass on unimodal distributions which look increasingly normally shaped. The model does not allow different variance for each normal component, unlike the model of Escobar and West (1995), but the model approximate any distribution of y_i since it has full support. Griffin (2010) shows that this model leads to better performance, in terms of out-of-sample prediction, than previously proposed "default" priors with mixing on both the mean and variance.

This paper is concerned with extending the model of Griffin (2010) to problems of multivariate density estimation. A useful property of the univariate model is that μ and σ^2 can be given standard priors (including the standard non-informative prior for the normal model) so that the nonparametric analysis will tend towards a parametric analysis, assuming a normal distribution for y_i, as a tends to 1. This allows us to centre semiparametric models, such as linear mixed models, defined using nonparametric priors for some distributions directly over their parametric counterparts. This theme is developed in this paper.

The paper is organized as follows: Section 2 discusses two extensions of the univariate model to multivariate problems, Section 3 describes the Markov chain Monte Carlo algorithms needed to fit the models, Section 4 presents applications of the models to density estimation and semiparametric linear mixed effect models, Section 5 is a discussion.

2 Models

Suppose that the density of y_i is $f(\cdot)$ then this density is our parameter of interest in density estimation. The model in (1) implies that

$$f(y_i) = \sum_{i=1}^{\infty} w_i \mathrm{N}(y_i|\theta_i, a\sigma^2)$$

where w_i are the sizes of the jumps of a Dirichlet process and $\theta_i \sim \mathrm{N}(\mu, (1-a)\sigma^2)$, which defines a prior on $f(\cdot)$. The prior mean of $f(\cdot)$ is

$$\mathrm{E}[f(y_i)|a, \mu, \sigma^2] = \mathrm{N}(y_i|\mu, \sigma^2).$$

Therefore, the prior mean does not depend on a and the prior for $f(\cdot)$ is centred over a normal distribution with mean μ and variance σ^2. The parameter a controls the proportion of the overall variation, σ^2, allocated to within-component variation (given by $a\sigma^2$) and between-component variation (given by $(1-a)\sigma^2$). Larger values of a lead to more variation placed into the within-component variation and lead to realized densities $f(\cdot)$ that look increasingly like a normal distribution. Therefore, the parameter can be interpret as a measure of the departure of the density of y_i from a normal distribution with larger a representing a density closer to a normal density.

The parameterisation of the model in terms of a smoothness parameter, a, and a fixed prior mean, μ, for $f(\cdot)$ which does not depend on the smoothness parameter defines a useful model for univariate density estimation. Extending these ideas to a multivariate model, with p-dimensional observations y_i, would imply that the prior should have the property that

$$\mathrm{E}[f(y_i)|a, \mu, \Sigma] = \mathrm{N}(y_i|\mu, \Sigma) \tag{2}$$

where μ is a p-dimensional mean vector, Σ is a $p \times p$-dimensional covariance matrix and a is a vector of remaining smoothness parameters. A direct extension of the model in (1) would define $y_i \sim \mathrm{N}(\theta_i, a\Sigma)$ and $H = \mathrm{N}(\theta|\mu, (1-a)\Sigma)$ leading to the required centring property in (2) but would imply that the prior for the marginal distribution of y_{ij} is the model in (1) with σ^2 replaced by Σ_{jj}, θ replaced by θ_j and smoothness parameter a. The departures from normality would be the same for all variables. This will often be unrealistically simple in practice since there may be dimensions or directions in which the marginal distributions are close to normality and other dimensions or directions in which the marginal distributions are far from normal, *e.g.* multi-modal.

Two models will be defined which allow different levels of smoothing for different dimensions of direction of y_i. Both models have the centring property that $\mathrm{E}[f(y_i)|a, \mu, \Sigma] = \mathrm{N}(y_i|\mu, \Sigma)$. The idea that there may be some dimensions of y_i whose marginal distribution is closer to normality than others will underlie Model 1 and the idea that there may be some directions of y_i in which the marginal distribution is close to normality will underlie Model 2.

Model 1 defines smoothness parameters $a_1, a_2, \ldots, a_p$ and assumes that

$$y_i \sim \mathrm{N}(y_i|\theta_i, A^{1/2}\Sigma A^{1/2}),$$

$$\theta_i \sim G,$$

$$G \sim \mathrm{DP}(M, H)$$

where

$$H = \mathrm{N}(\theta|\mu, \Sigma - A^{1/2}\Sigma A^{1/2})$$

and $A^{1/2}$ is a $p \times p$-dimensional matrix of the form

$$A^{1/2} = \begin{pmatrix} \sqrt{a_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{a_p} \end{pmatrix}. \tag{3}$$

The model is only well-defined if both covariance matrices are positive-definite which is true if $0 < a_j < 1$ for all j. The covariance matrix $A^{1/2}\Sigma A^{1/2}$ is clearly positive definite (since it is a product of positive definite matrices) and so $f(y_i|\theta_i, a, \Sigma)$ is well-defined. The positive-definiteness of $\Sigma - A^{1/2}\Sigma A^{1/2}$ can be shown in the following way. Clearly

$$\Sigma - A^{1/2}\Sigma A^{1/2} = (I - A^{1/2})\Sigma(I + A^{1/2}) - \Sigma A^{1/2} + A^{1/2}\Sigma$$

and so

$$x^T(\Sigma - A^{1/2}\Sigma A^{1/2})x = x^T(I - A^{1/2})\Sigma(I + A^{1/2})x$$

for any vector x of appropriate length. The matrix $(I - A^{1/2})\Sigma(I + A^{1/2})$ is clearly positive-definite since it is the product of positive-definite matrices and so $\Sigma - A^{1/2}\Sigma A^{1/2}$ is also positive-definite.

The model implies that the components of the mixture have the same correlation matrix as the correlation matrix of the data since the (j, k)-th term of $A^{1/2}\Sigma A^{1/2}$ is $\sqrt{a_j a_k}\Sigma_{jk}$ and so the correlation between y_{ij} and y_{ik} conditional on θ_i is $\frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}$. The prior for the marginal distribution of y_{ij} follows the model in (1) with μ replaced by μ_j, σ^2 replaced by Σ_{jj} and a replaced by a_j. Therefore, the departure from normality in dimension j is controlled by a_j.

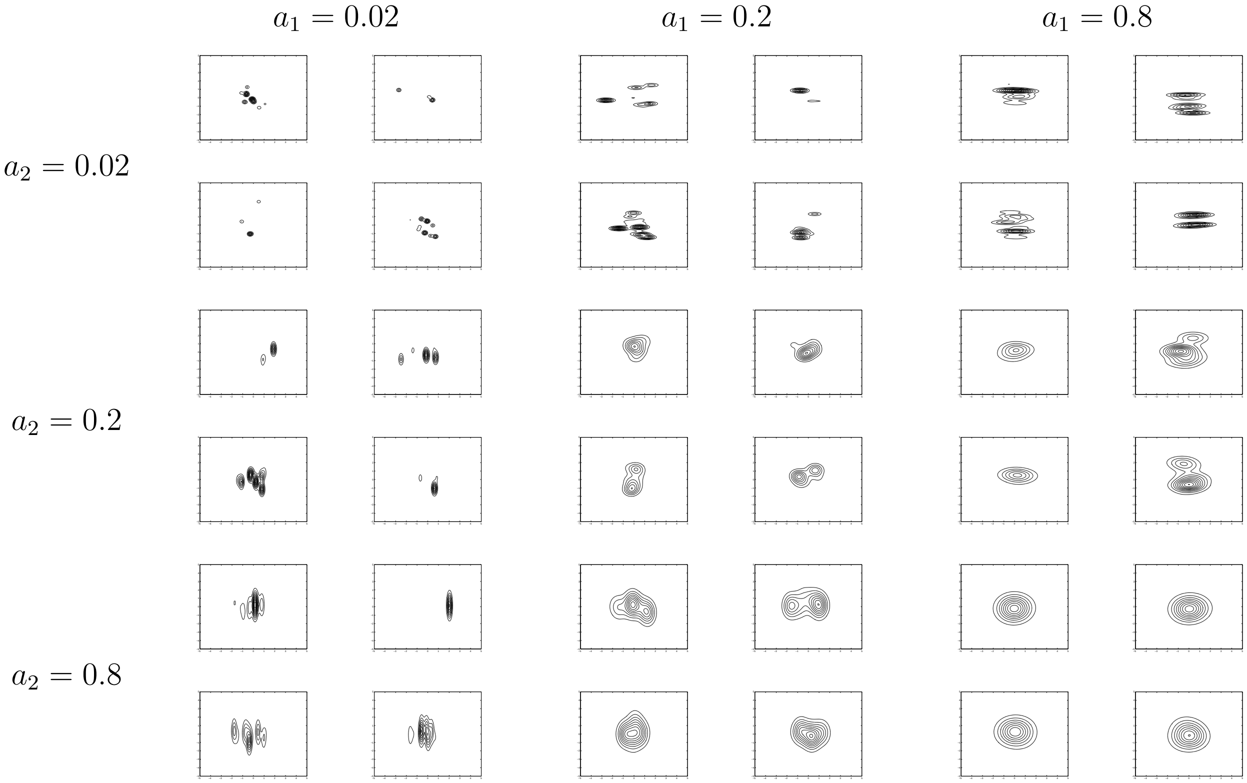

Figure 1: Four realisations of the density under Model 1 for different value of a_1 and a_2 with $M = 5$ and zero correlation.

Figure 1 shows realizations from the model when $p = 2$ with different values of a_1 and a_2. Clearly, the interpretation of the smoothness parameters is the same as the univariate models: larger values of a_k imply that the distribution of y_{ik} is increasingly normal and, small values of a_k tend to lead to multi-modal marginal distributions of y_{ik}. The multi-modality of the joint distribution depends on the value of both a_1 and a_2. In the Figure, small values for both a_1 and a_2 lead to a distribution with many modes whereas the combination of a small value for a_1 and a large value for a_2 leads to a few modes (and similarly for large a_1 and small a_2). The effect of introducing correlation between a_1 and a_2 is shown in Figure 2. The distribution is

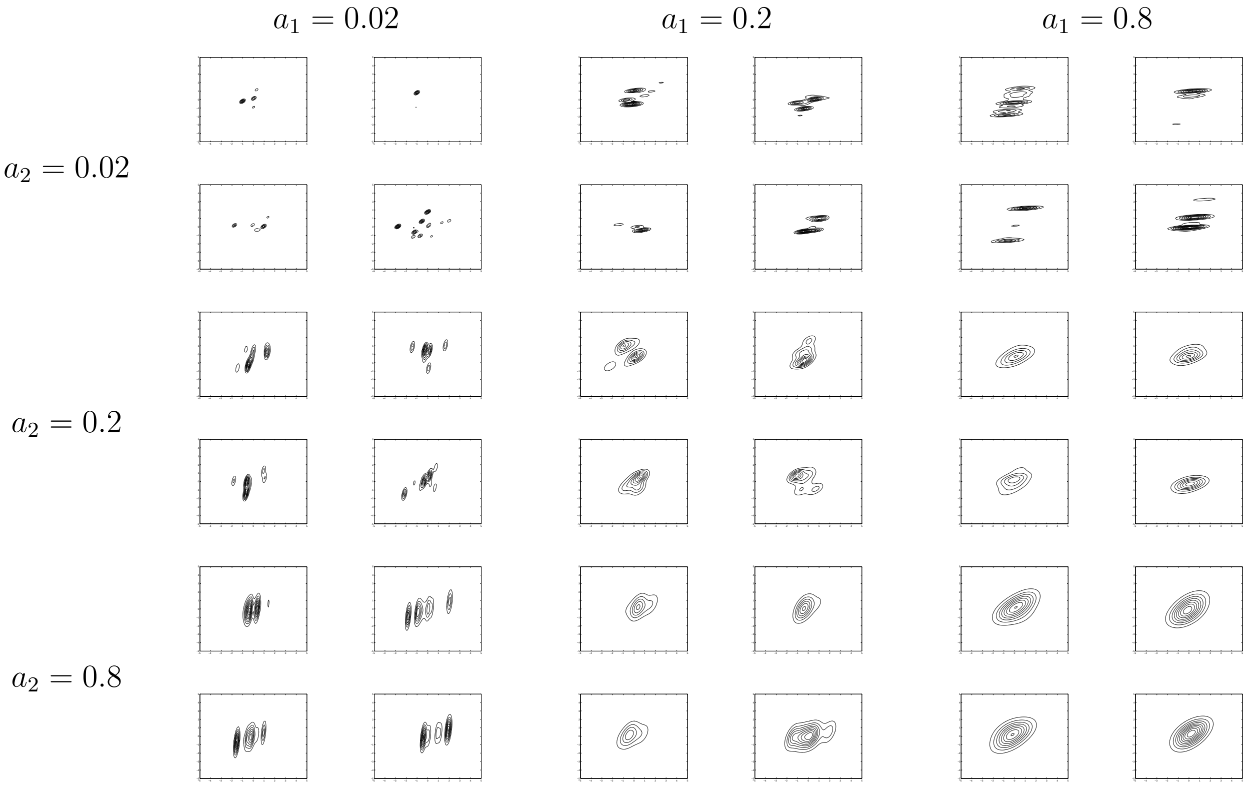

Figure 2: Four realisations of the density under Model 1 for different value of a_1 and a_2 with $M = 5$ and correlation of 0.5.

close to a normal distribution if both a_1 and a_2 are large. Small values of a_1 and a_2 leads to distributions with many modes where each component distribution have the same correlation. The effect of a small a_1 and a large a_2 is interesting and illustrates an important consequence of choosing this model. The marginal distribution of y_1 is multi-modal and the marginal distribution of y_2 is close to normality. Consequently, the joint distribution is constructed from several components with small variances in the first dimension and large variances in the second dimension.

The idea underlying Model 2 is that there are linear transforms of the data for which the marginal distribution is relatively normal and other linear transformations for which the marginal distribution is far from normal. This is achieved by defining $z_i = B^{-1}(y_i - \mu)$ and assuming that $z_{i1}, z_{i2}, \dots, z_{ip}$ are independent and follow the model in (1) with $\mu = 0$, $\sigma^2 = 1$ and smoothness parameter a_j. This implies that z_i is centred over a multivariate normal distribution with mean $(0, 0, \dots, 0)$ and the identity matrix I_p as its covariance matrix. Therefore, y_i is centred over $\mathrm{N}(\mu, BB^T)$ and a natural choice for B, which will be followed in this paper, is the Cholesky

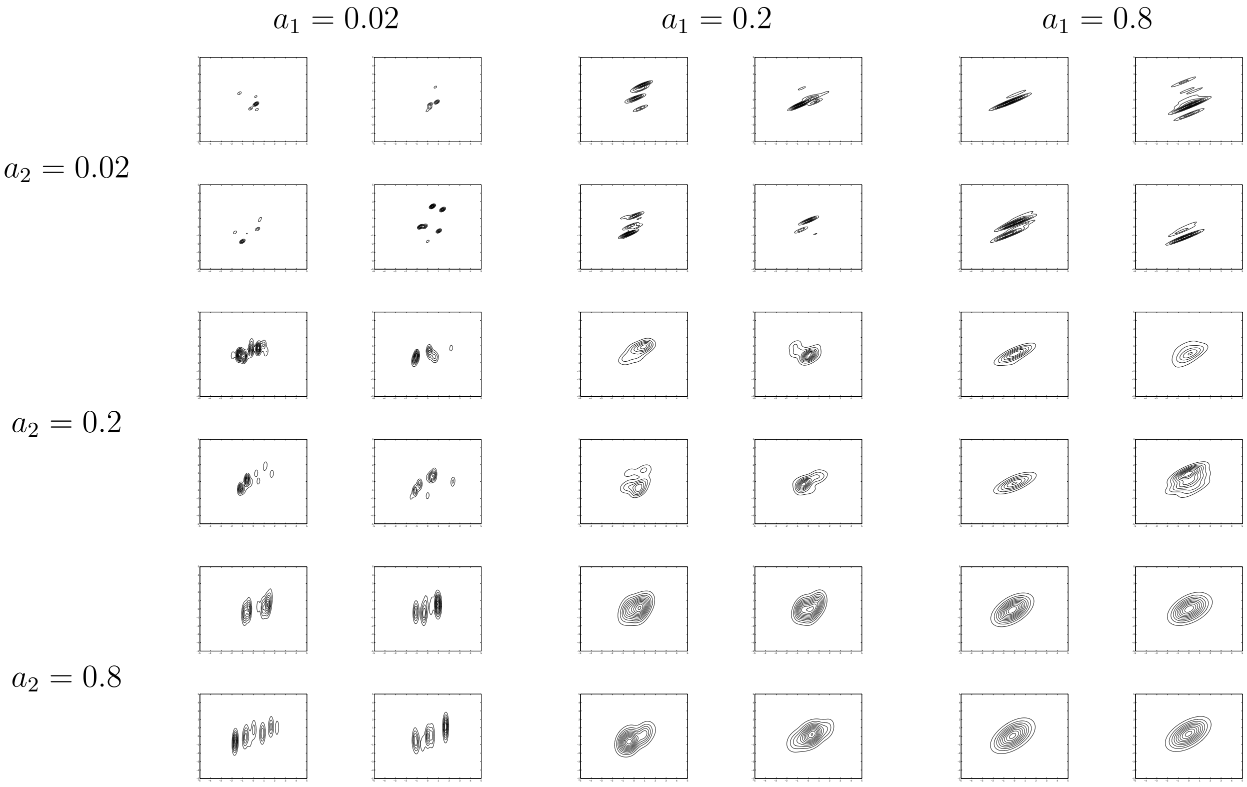

Figure 3: Four realisations of the density under Model 2 for different value of a_1 and a_2 with $M = 5$ and correlation of 0.5.

decomposition of the covariance matrix Σ. The model for y_i is then

$$y_i \sim \text{N}(\theta_i, BA^\star B^T)$$

$$\theta_i \sim \text{DP}(MH)$$

where $H = \text{N}(\theta|\mu, BB^T - BA^\star B^T)$ and A^* is a diagonal matrix with non-zero elements $a_1, a_2, \ldots, a_p$.

Model 1 and Model 2 will be the same if the correlation between y_{i1} and y_{i2} is zero. However, they differ in terms of how correlation is introduced. Unlike Model 1, Model 2 rotates the distribution if the correlation is non-zero. Realisations of this prior with a correlation of 0.5 are shown in Figure 3. Clearly small values of a_1 and a_2 lead to distributions 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.

3 Density estimation in random effect models

The distribution of unobserved quantities, such as random effects, will often have an unknown form. Bayesian approaches to density estimation can be naturally extended to this problem by assuming a nonparametric prior for the unknown distribution and so defines a semiparametric model. The Linear Mixed Model (LMM) will be the main focus of interest in this section although the ideas extend naturally to other models with random effects. The usual Normal LMM assumes that responses $y_{i1}, \ldots, y_{iT}$ are observed for the i-th individual with regressors $X_{i1}, \ldots, X_{iT}$ and $Z_{i1}, \ldots, Z_{iT}$ where

$$y_{it} = \alpha + X_{it}\beta + Z_{it}\gamma_i + \epsilon_{it},$$

$$\gamma_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \Sigma)$$

and ϵ_{it} are mutually independent and $\epsilon_{it} \sim \mathrm{N}(0, \sigma^2)$. The model makes the distinction that β are the same for all individuals and so X_{it} has the same effect for all individuals whereas γ_i varies from individual-to-individual and so allows for different effects of Z_{it} across individuals. The parameters of the model are then given priors which often have the form: $\beta \sim \mathrm{N}(0, \Sigma_\beta)$, $\alpha \sim \mathrm{N}(0, \sigma_\alpha)$, $\mu \sim \mathrm{N}(0, \Sigma_\mu)$ and $\Sigma \sim \mathrm{IW}(\eta, \Upsilon)$. The main focus of interest is usually the regression parameters β and the hyperparameters μ. A nonparametric specification for the distribution of γ_i defines a semiparametric model and leads to robust estimates of the regression coefficients.Initial work on Bayesian nonparametric estimation in these models is described by Bush and MacEachern (1996) and Kleinman and Ibrahim (1998). These assume that the distribution of γ_i follows a Dirichlet process mixture model. In this paper, specific forms of Dirichlet process mixture model (Model 1 and Model 2) are used to define LMM-Model 1 and LMM-Model 2. LMM-Model 1 assumes

$$\gamma_i \sim \mathrm{N}(\theta_i, A^{1/2}\Sigma A^{1/2})$$

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

where $H = \mathrm{N}(\theta|\mu, \Sigma - A^{1/2}\Sigma A^{1/2})$, A is defined in (3) and β, α, μ and Σ are given the priors for the Normal LMM. The model conditional on β, α, μ and Σ is centred over the normal model and the nonparametric model is the same as the parametric model model if $a_1, a_2, \ldots, a_p$ are all equal to one. Similarly, LMM-Model 2 can be defined by assuming that

$$\gamma_i \sim \mathrm{N}(\theta_i, BA^\star B^T)$$

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

where $H = \mathrm{N}(\theta|\mu, BB^T - BA^\star B^T)$, $A^\star$ is a diagonal matrix with non-zero elements $a_1, a_2, \ldots, a_p$ and β, α, μ and Σ are given the priors for the Normal LMM. Once again the model conditional on β, α, μ and Σ is centred over the normal model with mean μ and variance-covariance matrix $\Sigma = BB^T$ and the nonparametric model is the same as the parametric model model if $a_1, a_2, \ldots, a_p$ are all equal to one.

In both models inference about $a_1, a_2, \ldots, a_p$ provides information about the dimensions in which the distribution departs from normality. This is indicated by small values of a_i. Larger values of a_i will lead to an analysis which is closer to the parametric analysis and so allows us to take advantage of the normal assumption if it is supported by the data.

4 Computational Methods

Both Model 1 and Model 2 are conjugate Dirichlet process mixture models and so can be fitted using standard Pólya urn scheme methods described in MacEachern (1998). The choice of a Dirichlet process prior implies there are K distinct values of $\theta_1, \theta_2, \ldots, \theta_n$ which will be denoted $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$. Latent variables $s_1, s_2, \ldots, s_n$ are introduced to link these values so that $\theta_i = \theta_{(s_i)}$.

4.1 Model 1

The parameter μ and Σ are assumed to have the priors $\mu \sim \mathrm{N}(\mu_0, \Sigma_\mu)$ and $\Sigma \sim \mathrm{IW}(\eta, \Upsilon)$. Let $\Sigma_j = A^{1/2}\Sigma A^{1/2}$, $\Sigma_0 = \Sigma - A^{1/2}\Sigma A^{1/2}$ and $n_j = \#\{i | s_i = j\}$. The full conditional distributions are as follows.

Updating s

Let K^- be the number of distinct values for $\theta_1, \ldots \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n$ and $n_j^- = \#\{k | s_k = j, k \neq i\}$. The full conditional distribution of s_i is

$$p(s_i = j) \propto n_j \mathrm{N}\left(y_i \left| \theta_j^\star, \Sigma_j^\star\right.\right), \qquad j = 1, 2, \ldots, K^-$$

and

$$p(s_i = K^- + 1) \propto M\mathrm{N}\left(y_i \left| \mu, \Sigma\right.\right)$$

where

$$\theta_j^\star = \left(n_j^- \Sigma_k^{-1} + \Sigma_0^{-1}\right)^{-1} \left(\Sigma_k^{-1} \sum_{\{k | s_k = j, k \neq i\}} y_k + \Sigma_0^{-1} \mu_0\right)$$

and

$$\Sigma_j^\star = \left(n_j^- \Sigma_k^{-1} + \Sigma_0^{-1}\right)^{-1}.$$

After updating $s_1, s_2, \ldots, s_n$, the values of $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(K)}$ are sampled from their full conditional distribution $\theta_{(j)} \sim \mathrm{N}(\theta_j^\star, \Sigma_j^\star)$ where

$$\theta_j^\star = \left(n_j \Sigma_k^{-1} + \Sigma_0^{-1}\right)^{-1} \left(\Sigma_k^{-1} \sum_{\{k|s_k=j\}} y_k + \Sigma_0^{-1}\mu_0\right)$$

and

$$\Sigma_j^\star = \left(n_j \Sigma_k^{-1} + \Sigma_0^{-1}\right)^{-1}.$$

Updating μ

The full conditional distribution of μ is $\mathrm{N}(\mu|\mu^\dagger, \Sigma^\dagger)$ where

$$\mu^\dagger = \left(K\Sigma_0^{-1} + \Sigma_\mu^{-1}\right)^{-1} \left(\Sigma_0^{-1} \sum_{i=1}^K \theta_{(i)} + \Sigma_\mu^{-1}\mu_0\right),$$

$$\Sigma^\dagger = \left(K\Sigma_0^{-1} + \Sigma_\mu^{-1}\right)^{-1}.$$

Updating Σ and a

The covariance matrix Σ and a can be updated in the following way. We reparametrise Σ to $(P, \sigma_1^2, \ldots, \sigma_p^2)$ where P is the $(p \times p)$-dimensional correlation matrix with terms

$$P_{ij} = \frac{\Sigma_{ij}}{\Sigma_{11}^{1/2}\Sigma_{22}^{1/2}}$$

and $\sigma_j^2 = \Sigma_{jj}$ and update these parameters using a Metropolis-Hastings random walk. The full conditional distribution is

$$p(\Sigma) \propto \Sigma_k^{-n/2}\Sigma_0^{K/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^n (y_i - \theta_i)^T \Sigma_k^{-1} (y_i - \theta_i)^T + \sum_{j=1}^K (\theta_{(j)} - \mu)^T \Sigma_0^{-1} (\theta_{(j)} - \mu)\right]\right\}.$$

The value of σ_j^2 is updated with a_j. The new value of σ_j^2 is proposed as $\sigma_j^{2\prime} = \sigma_j^2 \exp\{\phi_j^2 \epsilon\}$ where $\epsilon \sim \mathrm{N}(0,1)$ and the new value of a_j is proposed as $a_j' = \frac{(1+a_j)/(1-a_j)\exp\{\nu_j\epsilon\}-1}{(1+a_j)/(1-a_j)\exp\{\nu_j\epsilon\}+1}$. Let Σ' be the values of the covariance matrix with σ_j^2 and a_j replaced by $\sigma_j^{2\prime}$ and a_j'. The new values is accepted with probability

$$\max\left\{1, \frac{p(a_j')p(\Sigma')|\Sigma'|^{(\eta+p+1)/2} \exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma'^{-1}\right)/2\right\} \sigma_j^{2\prime(p+1)/2} \left(\frac{1}{1+a_j'} + \frac{1}{1-a_j'}\right)}{p(a_j)p(\Sigma)|\Sigma|^{(\eta+p+1)/2} \exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma^{-1}\right)/2\right\} \sigma_j^{2(p+1)/2} \left(\frac{1}{1+a_j} + \frac{1}{1-a_j}\right)}\right\}.$$

The new value of P_{ij} is proposed as $P'_{ij} = \frac{(1+P_{ij})/(1-P_{ij})\exp\{\zeta_{ij}\epsilon\}-1}{(1+P_{ij})/(1-P_{ij})\exp\{\zeta_{ij}\epsilon\}+1}$. Let Σ' be the values of the covariance matrix with P_{ij} replaced by P'_{ij}. The new value is accepted with probability

$$\max\left\{1, \frac{p(\Sigma')|\Sigma'|^{(\eta+p+1)/2}\exp\left\{-\text{tr}\left(\Upsilon\Sigma'^{-1}\right)/2\right\}\left(\frac{1}{1+P'_{ij}}+\frac{1}{1-P'_{ij}}\right)}{p(\Sigma)|\Sigma|^{(\eta+p+1)/2}\exp\left\{-\text{tr}\left(\Upsilon\Sigma^{-1}\right)/2\right\}\left(\frac{1}{1+P_{ij}}+\frac{1}{1-P_{ij}}\right)}\right\}.$$

The variance of the increments ζ_{ij}, ν_j and ϕ_j are tuned using the adaptive scheme of Atchadé and Rosenthal (2005) with a target acceptance rate of 0.3.

Updating M

The full conditional distribution for M is

$$p(M)\frac{\Gamma(M)}{\Gamma(M+n)}M^K$$

and can be updated using a Metropolis-Hastings random walk step. The variance of the increments is tuned using the adaptive scheme of Atchadé and Rosenthal (2005).

4.2 Model 2

We define $\Sigma = BB^T$ and assume the priors $\mu \sim \text{N}(\mu_0, \Sigma_\mu)$ and $\Sigma \sim \text{IW}(\eta, \Sigma_0)$. Model 2 can be updated in the same way as Model 1 with $\Sigma_k = BA^{1/2}A^{1/2}B^T$ and $\Sigma_0 = BB^T - BA^{1/2}A1/2B^T$.

4.3 LMM-Model 1

The model in Section 3 can be fitted using a Gibbs sampler with the updating scheme as follows. The parameters s and μ are updated marginalizing over γ which lead to much better mixing of the sampler than the Gibbs sampler including γ at all steps.

Updating β

The full conditional distribution of β is

$$\text{N}\left(\left(\sigma^{-2}\sum_{i=1}^{n}\sum_{t=1}^{T}X'_{it}X_{it}+\Lambda_\beta\right)^{-1}\sigma^{-2}\sum_{i=1}^{n}\sum_{t=1}^{T}X'_{it}\left(y_{it}-Z_{it}\gamma_i\right), \left(\sigma^{-2}\sum_{i=1}^{n}\sum_{t=1}^{T}X'_{it}X_{it}+\Lambda_\beta\right)^{-1}\right)$$

Updating σ^2

The full conditional distribution of σ^2 is

$$\text{Ga}\left(\alpha + nT, \beta + \sum_{i=1}^{n}\sum_{t=1}^{T}(y_i - X_{it}\beta - Z_{it}\gamma_i)^2\right)$$

Updating s

The full conditional distribution of s_i is

$$p(s_i = j) \propto n_j \text{N}\left(y_i^\star \,\middle|\, Z_i\theta_{(j)}, Z_i A^{1/2}\Sigma A^{1/2} Z_i^T + \sigma^2\right), \qquad j = 1, 2, \ldots, K^-$$

and

$$p(s_i = K^- + 1) \propto M\text{N}\left(y_i^\star \,\middle|\, Z_i\mu, Z_i\Sigma Z_i^T + \sigma^2\right)$$

where

$$y_{ik}^\star = y_{ik} - X_{ik}\beta$$

If $s_i = K^- + 1$ then $\theta_{(K^- + 1)}$ is sampled from $\text{N}(\theta^\star, \Sigma^\star)$ where

$$\theta^\star = (Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}Z_i + \Sigma_0^{-1})^{-1}(Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}y_i + \Sigma_0^{-1}\mu)$$

and

$$\Sigma^\star = (Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}Z_i + \Sigma_0^{-1})^{-1}$$

where $\Sigma_k = A^{1/2}\Sigma A^{1/2}$ and $\Sigma_0 = \Sigma - A^{1/2}\Sigma A^{1/2}$

Updating θ

The full conditional distribution of $\theta_{(k)}$ is $\text{N}(\theta^\star, \Sigma^\star)$ where

$$\theta^\star = \left(\sum_{i|s_i=k} Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}Z_i + \Sigma_0^{-1}\right)^{-1}\left(\sum_{i|s_i=k} Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}y_i + \Sigma_0^{-1}\mu\right)$$

and

$$\Sigma^\star = \left(\sum_{i|s_i=k} Z_i'(Z_i\Sigma_k Z_k^T + \sigma^{-2}I_T)^{-1}Z_i + \Sigma_0^{-1}\right)^{-1}$$

Updating γ

The full conditional distribution for γ_i is

$$\text{N}\left(\left(\sigma^{-2}\sum_{t=1}^{T} Z_{it}'Z_{it} + \Sigma_k^{-1}\right)^{-1}\left(\sigma^{-2}\sum_{t=1}^{T} Z_{it}(y_{it} - Xit\beta) + \Sigma_k^{-1}\mu_{s_i}\right), \left(\sigma^{-2}\sum_{t=1}^{T} Z_{it}'Z_{it} + \Sigma_k^{-1}\right)^{-1}\right)$$

Updating Σ and a

The covariance matrix Σ and a can be updated in the following way. We reparametrise Σ to $(P, \sigma_1^2, \ldots, \sigma_p^2)$ where P is the $(p \times p)$-dimensional correlation matrix with terms

$$P_{ij} = \frac{\Sigma_{ij}}{\Sigma_{11}^{1/2}\Sigma_{22}^{1/2}}$$

and $\sigma_j^2 = \Sigma_{jj}$ and update these parameters using a Metropolis-Hastings random walk. The full conditional distribution is

$$p(\Sigma) \propto \Sigma_k^{-n/2}\Sigma_0^{K/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(\gamma_i - \theta_i)^T\Sigma_k^{-1}(\gamma_i - \theta_i)^T + \sum_{j=1}^{K}(\theta_{(j)} - \mu)^T\Sigma_0^{-1}(\theta_{(j)} - \mu)\right]\right\}.$$

The value of σ_j^2 is updated with a_j. The new value of σ_j^2 is proposed as $\sigma_j^{2\prime} = \sigma_j^2 \exp\{\phi_j^2\epsilon\}$ where $\epsilon \sim \mathrm{N}(0, 1)$ and the new value of a_j is proposed as $a_j' = \frac{(1+a_j)/(1-a_j)\exp\{\nu_j\epsilon\}-1}{(1+a_j)/(1-a_j)\exp\{\nu_j\epsilon\}+1}$. Let Σ' be the values of the covariance matrix with σ_j^2 and a_j replaced by $\sigma_j^{2\prime}$ and a_j'. The new values is accepted with probability

$$\max\left\{1, \frac{p(a_j')p(\Sigma')|\Sigma'|^{(\eta+p+1)/2}\exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma'^{-1}\right)/2\right\}\sigma_j^{2\prime(p+1)/2}\left(\frac{1}{1+a_j'} + \frac{1}{1-a_j'}\right)}{p(a_j)p(\Sigma)|\Sigma|^{(\eta+p+1)/2}\exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma^{-1}\right)/2\right\}\sigma_j^{2(p+1)/2}\left(\frac{1}{1+a_j} + \frac{1}{1-a_j}\right)}\right\}.$$

The new value of P_{ij} is proposed as $P_{ij}' = \frac{(1+P_{ij})/(1-P_{ij})\exp\{\zeta_{ij}\epsilon\}-1}{(1+P_{ij})/(1-P_{ij})\exp\{\zeta_{ij}\epsilon\}+1}$. Let Σ' be the values of the covariance matrix with P_{ij} replaced by P_{ij}'. The new value is accepted with probability

$$\max\left\{1, \frac{p(\Sigma')|\Sigma'|^{(\eta+p+1)/2}\exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma'^{-1}\right)/2\right\}\left(\frac{1}{1+P_{ij}'} + \frac{1}{1-P_{ij}'}\right)}{p(\Sigma)|\Sigma|^{(\eta+p+1)/2}\exp\left\{-\mathrm{tr}\left(\Upsilon\Sigma^{-1}\right)/2\right\}\left(\frac{1}{1+P_{ij}} + \frac{1}{1-P_{ij}}\right)}\right\}.$$

The variance of the increments ζ_{ij}, ν_j and ϕ_j are tuned using the adaptive scheme of Atchadé and Rosenthal (2005) with a target acceptance rate of 0.3.

Updating M

The full conditional distribution for M is

$$p(M)\frac{\Gamma(M)}{\Gamma(M+n)}M^K$$

and can be updated using a Metropolis-Hastings random walk step. The variance of the increments is tuned using the adaptive scheme of Atchadé and Rosenthal (2005).

Updating μ

The full conditional distribution for μ is

$$\mathrm{N}\left(\left(K\Sigma_0^{-1}+\Sigma_\mu^{-1}\right)^{-1}\left(\Sigma_0^{-1}\sum_{k=1}^{K}\theta_{(k)}+\Sigma_\mu^{-1}\mu_0\right),\left(K\Sigma_0^{-1}+\Sigma_\mu^{-1}\right)^{-1}\right).$$

4.4 LMM-Model 2

The sampler for LMM-Model 1 can be used with $\Sigma_k = BA^{1/2}A^{1/2}B^T$ and $\Sigma_0 = BB^T - BA^{1/2}A^{1/2}B^T$.

5 Examples

The models for Bayesian density estimation are applied to two problems in bivariate density estimation for observables, and a simulated and a real data example for unobservable random effects in a linear mixed model.

5.1 Bivariate density estimation: Old faithfull data

Data measuring the waiting time between eruptions and the duration of the eruption of the Old Faithful geyser in Yellowstone National Park are available in the R package ‘datasets’ and have become a popular data set for bivariate density estimation. Model 1 and Model 2 were fitted with two prior settings for a_i. The first prior setting assumes that a_i follows a uniform distribution, which is a natural default choice for parameters on $(0, 1)$, and will be written as $\mathrm{Be}(1, 1)$, a Beta distribution with both parameters equal to ones. The second choice is $a_i \sim \mathrm{Be}(1, 10)$, a Beta distribution with parameters 1 and 10, which places more mass on small values of a_i and so more mass is placed on distributions which are far from a normal distribution (Griffin, 2010). The parameters μ and Σ are given the priors $\mu \sim \mathrm{N}(\bar{y}, 10^4)$ and $\Sigma \sim \mathrm{IW}(p + 2, \hat{\Sigma})$ where $\bar{y}$ and $\hat{\Sigma}$ are the sample mean and variance-covariance matrix of the data. This choice ensures that the prior mean of Σ is $\hat{\Sigma}$.

The posterior mean density of the observations and the data are shown in Figure 4 For all models and all priors, the estimated density is bimodal with one mode around (2, 55) and another mode around (4.5, 80) and follows the data. The density around (4.5, 80) is clearly skewed. The posterior distribution of a_1 and a_2 for Model 1 are summarized in Table 1. The posterior median of a_1 is much smaller than those of a_2 for both priors. The results indicate that the marginal density of duration is close to normal but the marginal density of waiting time is far from normal. Results for a_1 and

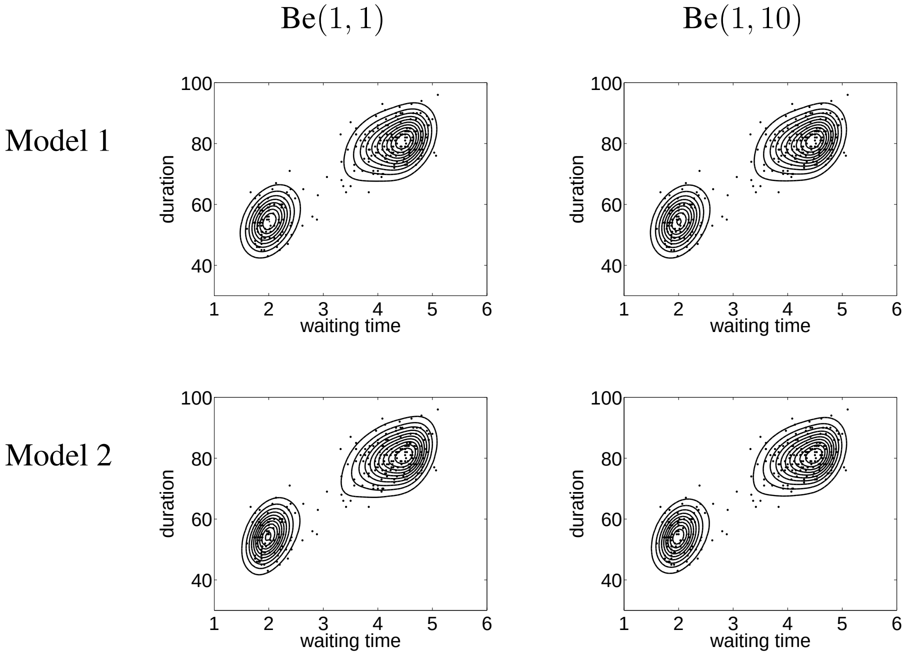

Figure 4: Contour plots of the posterior mean density under Models 1 and 2 with two different priors for a_1 and a_2 for the Old Faithfull data (observations are shown as dots).

	a_1	a_2
Be$(1, 1)$	0.12 *(0.04, 0.24)*	0.31 *(0.14, 0.51)*
Be$(1, 10)$	0.09 *(0.03, 0.19)*	0.20 *(0.08, 0.37)*

Table 1: The posterior distribution of a_1 and a_2 summarized as the posterior median and 95% credible interval for the Old Faithfull data using Model 1 with two different priors for a_1 and a_2.

a_2 are harder to interpret in Model 2 since the prior and likelihood is invariant to the permutation of a_1 and a_2 and no results are presented.

5.2 Bivariate density estimation: Aircraft data

Bowman and Azzalini (1997) describe bivariate density estimation for data related to a study of the development of aircraft technology originally analysed by Saviotti and Bowman (1984). This will be re-analyzed using the methods described in this paper. The data 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. The priors for a_i, μ and Σ were chosen in the same way as in the example in Section 5.1.

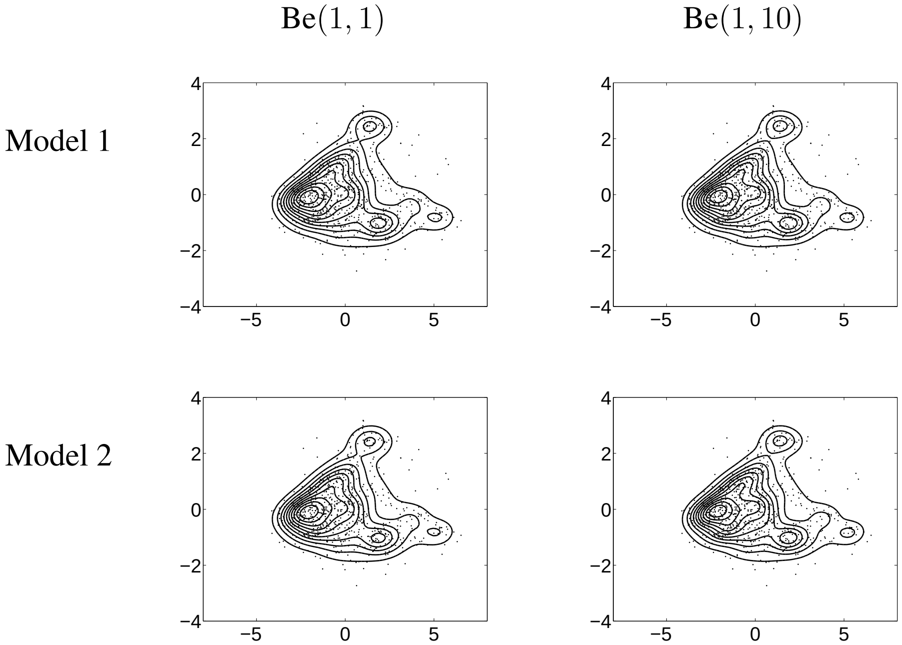

Figure 5: Contour plots of the posterior mean density under Models 1 and 2 with two different priors for a_1 and a_2 for the aircraft data (observations are shown as dots).

Figure 5 shows a scatterplot of the data and the posterior mean density of the observations across both models and prior settings for a_i. The results are robust to these choices. The posterior mean gives a good description of the data. In particular, the higher density of points around -4 for the first dimension x_1 is well captured. The posterior distributions of a_1 and a_2 for Model 1 under the two priors for a_i are summarised in Table 2. The posterior medians for both a_1 and a_2 are close to 0.1

	a_1	a_2
$\text{Be}(1,1)$	0.11 *(0.06, 0.24)*	0.10 *(0.05, 0.20)*
$\text{Be}(1,10)$	0.09 *(0.04, 0.19)*	0.08 *(0.04, 0.15)*

Table 2: The posterior distribution of a_1 and a_2 summarized as the posterior median and 95% credible interval for the aircraft data using Model 1 with two different priors for a_1 and a_2.

indicating that there are similar levels of non-normality in both variables, as illustrated by the posterior mean densities.

5.3 Nonparametric linear mixed model: Simulated Data

The nonparametric linear mixed model of Section 3 was applied to a simulated data set. The design of the simulated data sets follows from Ghidey et al. (2004). It is assumed that observations, $y_{i1}, y_{i2}, \dots, y_{i6}$ are made at 6 time points $t_i = (0, 2, 4, 6, 8, 10)$ and that

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \gamma_{0i} + \gamma_{1i} t_{ij} + \epsilon_{ij}$$

where $\epsilon_{ij} \sim \mathrm{N}(0, 0.04)$. This linear growth model where the intercept and slope are assumed to be different for each individual. Data were generated from this model with $\beta_0 = 2.35$ and $\beta_1 = 0.28$. Two distributions were assumed for the random effect $\gamma = (\gamma_0, \gamma_1)$:

- Example 1

$$\gamma \sim \mathrm{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \left(\begin{array}{cc} 0.15 & 0.02 \\ 0.02 & 0.04 \end{array}\right)\right)$$

- Example 2

$$\gamma \sim 0.5 \times \mathrm{N}\left(\left(\begin{array}{c} -1 \\ 0 \end{array}\right), \left(\begin{array}{cc} 0.15 & 0.02 \\ 0.02 & 0.04 \end{array}\right)\right) + 0.5 \times \mathrm{N}\left(\left(\begin{array}{c} 1 \\ 0 \end{array}\right), \left(\begin{array}{cc} 0.15 & 0.02 \\ 0.02 & 0.04 \end{array}\right)\right).$$

Example 1 assumes a normal distribution for the random effects whereas Example 2 assumes that they are drawn from a mixture of two normal distributions. The marginal distribution of b_0 in Example 2 is bi-modal whereas the marginal distribution of b_1 is normal. A sample was generated from the two examples with $n = 200$.

The posterior mean density under Models 1 and 2 for Example 1, shown in Figure 6, is a good estimate of the true random effects distribution. In Model 1, the posterior medians of a_1 and a_2 are 0.21 and 0.17 respectively indicating that both marginal distributions are close to normal. The estimates of the regression parameters β_0 and β_1, shown in Table 3, are very similar for the parametric model (which is correctly specified) and both nonparametric models. Therefore, the two nonparametric models can replicate the parametric analysis when it is supported by the data.

The posterior means of the unknown density under Models 1 and 2 for Example 2 are shown in Figure 7. Both estimates show the bi-modal random effects distribution generating the data. The posterior medians of a_1 and a_2 are 0.09 and 0.21 indicating that the marginal distribution of b_1 is less normally distributed than b_2 and that b_2 is close to normally distributed which both match the properties of the model used to

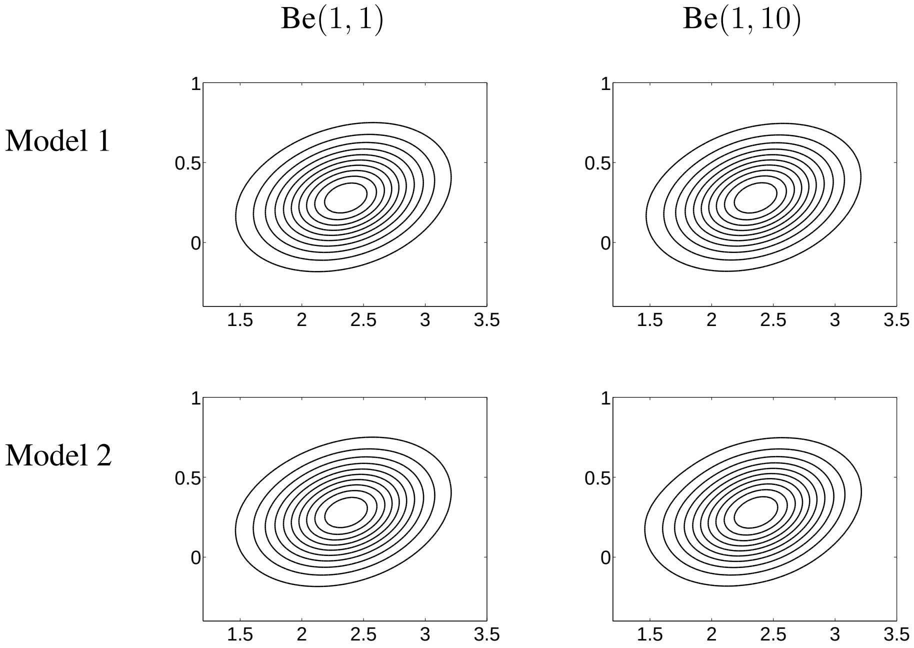

Figure 6: Contour plots of the posterior mean density of γ under Models 1 and 2 with two different priors for a_1 and a_2 for Example 1.

generate the data. The esimated regression coefficients for Example 2 are shown in Table 4. Both two nonparametric models and the parametric models show very similar results for the posterior median. However, the parametric model underestimates the uncertainty in the estimate of β_0, since it assumes a normal distribution for b_0 rather than the true bi-modal distribution.

5.4 Nonparametric linear mixed model: Cholesterol Data

The use of the model for density estimation in a linear mixed model are illustrated using repeated cholesterol measurements for a sample of 200 subjects from the Framingham study. The data was originally analysed by Zhang and Davidian (2001) and has been subsequently re-analysed by Ghidey et al. (2004) and Ho and Hu (2008). The measurements for taken every 2 years for a period of 10 years. There are two fixed effects, age and sex, and the intercept and the effect of time are assumed to be random effects drawn from a bivariate distribution. Assuming a nonparametric specification for the distribution leads to a semiparametric model. The full model can be written

$$Y_{ij} = \beta_{0i} + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_{3i} t_{ij} + \epsilon_{ij}$$

Model 1

	$\mathrm{Be}(1,10)$	$\mathrm{Be}(1,1)$
β_0	2.34 *(2.28, 2.40)*	2.34 *(2.29, 2.40)*
β_1	0.28 *(0.25, 0.31)*	0.28 *(0.25, 0.31)*
a_1	0.209 *(0.055, 0.464)*	0.706 *(0.240, 0.980)*
a_2	0.175 *(0.046, 0.423)*	0.612 *(0.158, 0.972)*

Model 2

	$\mathrm{Be}(1,10)$	$\mathrm{Be}(1,1)$
β_0	2.34 *(2.29, 2.40)*	2.34 *(2.29, 2.40)*
β_1	0.28 *(0.25, 0.31)*	0.28 *(-0.25, 0.31)*

Parametric

β_0	2.35 *(2.29, 2.40)*
β_1	0.28 *(0.25, 0.31)*

Table 3: The posterior distribution of the regression coefficients summarized as the posterior median and 95% credible interval for Example 1 using parametric Normal model and Models 1 and Model 2 with two different priors for a_1 and a_2.

$$\begin{pmatrix} \beta_{0i} \\ \beta_{3i} \end{pmatrix} \sim F$$

where Y_{ij} is the j-th observation for the i-th subject, t_{ij} is the time of the measurement, age_i and sex_i are the measured covariates for the i-th subject. It is assumed that $\epsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$ and F is assumed to follow Model 1 or Model 2. The posterior means of F under the different models and different prior settings are shown in Figure 8. The estimated distribution show a positive correlation between the intercept and the slope. The distribution is not close to normally distributed with the iso-contour not so tightly packed for larger values of the intercept. The posterior means of the marginal distributions for the intercept and slope are shown in Figure 9. The intercept has a clear positive skewness for all models and all priors whereas the slope has a negative skewness which becomes more pronounced for the $\mathrm{Be}(1,10)$ prior.

The estimated regression coefficients for the data under the two models with the two beta priors are shown in Table 5. The estimates for the slope under the two models for both prior setting are similar and are relatively robust. The posterior median of the intercept parameter varies more across the different models and differ from the value for the parametric analysis. The $\mathrm{Be}(1,1)$ prior for both models led to results closer to the value for the parametric analysis. A little surprisingly, the posterior median of

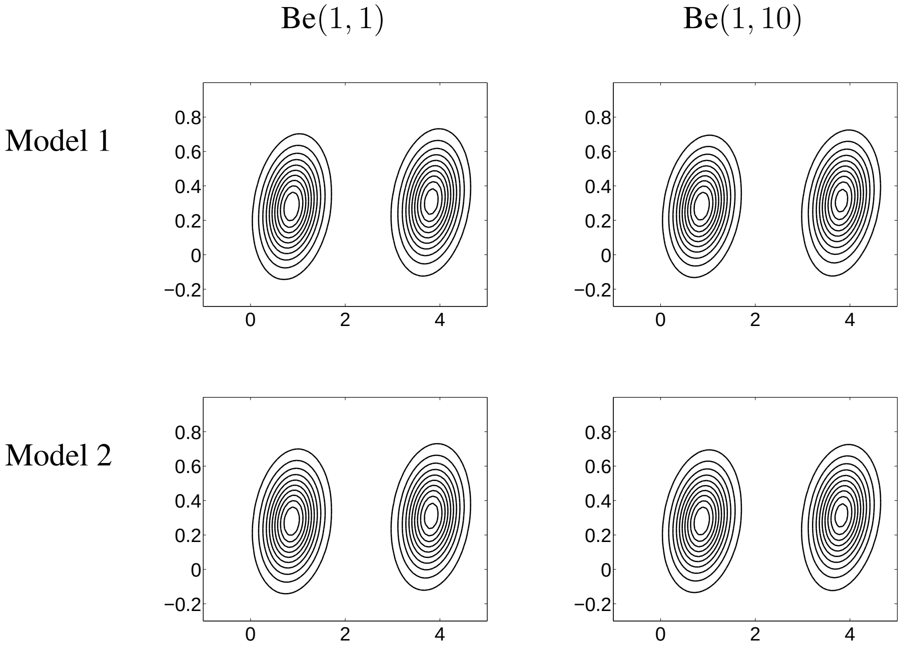

Figure 7: Contour plots of the posterior mean density of γ under Models 1 and 2 with two different priors for a_1 and a_2 for Example 2.

a_1 is larger than a_2 showing that the marginal distribution of the intercept is closer to normality than the marginal distribution of the slope. This is due to heavier than normal tails in the posterior mean distributions of a_2. The effect of age has very similar posterior median values across all models fitted to the data. However, there is difference between the estimate of the effect of sex under the nonparametric models, which have posterior medians close to -0.06, and the parametric model, which has a median of -0.041.

6 Discussion

This paper has presented methods for Bayesian nonparametric density estimation for multivariate data and a semiparametric linear mixed model with an unknown random effects distribution. The prior for the unknown distribution is centred over a normal distribution, in the sense that the prior predictive distribution of an observation y conditional on μ and Σ is a normal distribution with mean μ and variance Σ. Two specifications are considered. One assumes parameters that measure the departure from normalty in each dimensions. The second assumes that a rotated version of y is

Model 1

	$\mathrm{Be}(1, 10)$	$\mathrm{Be}(1, 1)$
β_0	2.34 *(2.13, 2.55)*	2.34 *(2.13, 2.55)*
β_1	0.29 *(0.26, 0.32)*	0.29 *(0.26, 0.32)*
a_1	0.093 *(0.025, 0.210)*	0.119 *(0.040, 0.270)*
a_2	0.208 *(0.064, 0.445)*	0.659 *(0.189, 0.960)*

Model 2

	$\mathrm{Be}(1, 10)$	$\mathrm{Be}(1, 1)$
β_0	2.34 *(2.13, 2.55)*	2.34 *(2.13, 2.55)*
β_1	0.29 *(0.26, 0.32)*	0.29 *(0.26, 0.32)*

Parametric

β_0	2.34 *(2.13, 2.55)*
β_1	0.29 *(0.26, 0.32)*

Table 4: The posterior distribution of the regression coefficients summarized as the posterior median and 95% credible interval for Example 2 using parametric Normal model and Models 1 and Model 2 with two different priors for a_1 and a_2.

centred over a normal distribution with covariance given by the identity matrix. Semiparametric linear mixed models can be defined by assuming that the distribution of the random effects is modelled using one of these nonparametric models. The semiparametric model is then centred over the semiparametric model and the inference can be "shrunk" towards the parametric model when the data supports the parametric model.

References

Atchadé, Y. F. and J. S. Rosenthal (2005). On Adaptive Markov Chain Monte Carlo Algorithms. *Bernoulli 11*, 815–828.

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

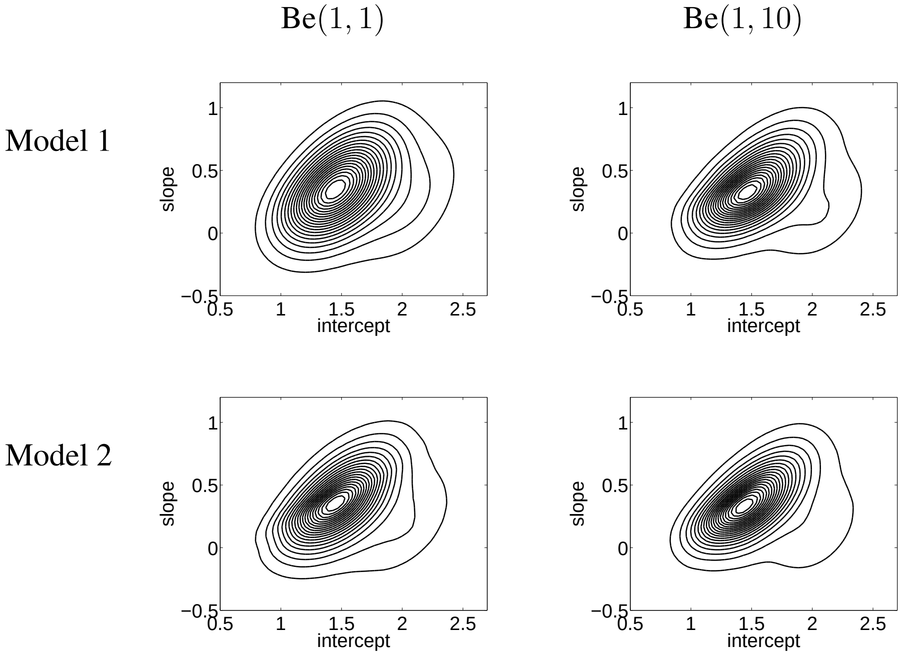

Figure 8: Contour plots of the posterior mean density under Models 1 and 2 with two different priors for a_1 and a_2 for the chlosterol data.

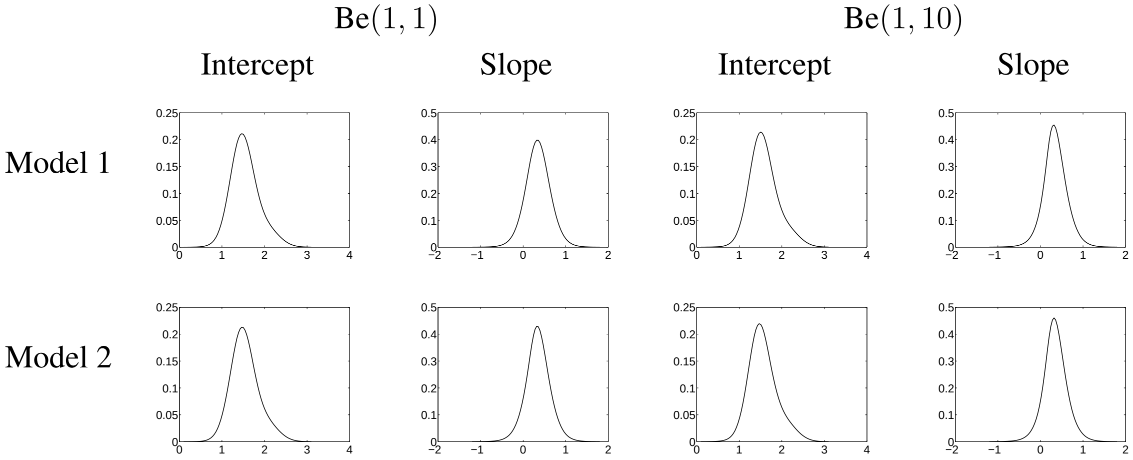

Figure 9: Plots of the posterior mean marginal densities for intercept and slope under Models 1 and 2 with two different priors for a_1 and a_2 for the chlosterol data.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Model 1

	Be$(1,10)$	Be$(1,1)$
β_1 (age)	0.015 *(0.008, 0.022)*	0.017 *(0.010, 0.022)*
β_2 (sex)	-0.064 *(-0.158, 0.031)*	-0.056 *(-0.153, 0.039)*
intercept	1.73 *(1.43, 2.04)*	1.67 *(1.42, 2.01)*
slope	0.28 *(0.23, 0.33)*	0.28 *(0.23, 0.33)*
a_1	0.140 *(0.048, 0.325)*	0.283 *(0.097, 0.635)*
a_2	0.094 *(0.013, 0.286)*	0.305 *(0.065, 0.857)*

Model 2

	Be$(1,10)$	Be$(1,1)$
β_1 (age)	0.016 *(0.010, 0.022)*	0.016 *(0.010, 0.022)*
β_2 (sex)	-0.063 *(-0.157, 0.031)*	-0.060 *(-0.154, 0.037)*
intercept	1.70 *(1.42, 1.97)*	1.68 *(1.42, 1.97)*
slope	0.28 *(0.23, 0.33)*	0.28 *(0.23, 0.33)*

Parametric

β_1 (age)	0.017 *(0.010, 0.026)*
β_2 (sex)	-0.041 *(-0.160, 0.076)*
intercept	1.64 (1.26, 1.96)
slope	0.28 (0.22, 0.34)

Table 5: The posterior distribution of the regression coefficients and the smoothness parameters summarized as the posterior median and 95% credible interval for Example 1 using parametric Normal model and Models 1 and Model 2 with two different priors for a_1 and a_2.

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

Ghidey, W., E. Lesaffre, and P. Eilers (2004). Smooth Random Effects Distribution in a Linear Mixed Model. *Biometrics 60*, 945–953.

Griffin, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis 5*(1), 45–64.

Ho, R. K. W. and I. Hu (2008). Flexible modelling of random effects in linear mixed models - a bayesian approach. *Computational Statistics and Data Analysis 52*, 1347–1361.

Kleinman, K. P. and J. G. Ibrahim (1998). A semiparametric bayesian approach to the random effects model. *Biometrics 54*, 921–938.

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

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

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*, 795–802.