The use of mixture distributions in a Bayesian linear mixed effects model

Anirudh TOMER

Supervisor: Prof. Emmanuel Lesaffre
L-BioStat, KU Leuven

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics

Academic year 2015-2016

Preface

The following thesis work was conducted as part of the credits completion requirements of the MSc. Statistics programme at KU Leuven. The purpose of the thesis was to find out the efficacy of Bayesian model selection criteria, for choosing the right number of components in a mixture distribution of random effects, in a longitudinal model. When I began working on this project in August 2015, I had little idea that I would be able to go as far as I have been now. There were many significant obstacles on the way, such as derivations of the various definitions of Deviance information criteria and marginal likelihood, and choice of posterior predictive checks. While implementing them I realized that the simulation study for this thesis required much more computational power than I had. Although the pace of execution was slow, yet every time I had a result I was excited to see it. Looking backwards, I think it was perhaps the most interesting project I did in last 3 years. Through and through, I enjoyed every bit of this project. The entire work for this thesis has been done using R and JAGS (Just Another Gibbs Sampler). The source code, results of simulations and an electronic draft of this thesis can be downloaded from: `https://github.com/anirudhtomer/MScThesis`

In chapter 1 an introduction to mixture distributions and their central role in the formulation of the problem statement for this thesis is presented. Since the project was done using Bayesian methods it became essential to give an introduction to the Bayesian paradigm in Chapter 2. Further in Chapter 3 the definition of a Bayesian heterogeneity model and issues related to parameter estimation are presented. Chapter 4 constitutes the analytical calculations I did for various classes of Deviance information criteria, marginal likelihood and posterior predictive checks. Chapter 5 includes the results of the simulation study that was performed to check the efficacy of the aforementioned Bayesian model selection methods. The results from Chapter 5 are used further in Chapter 6 to apply the right model selection criteria for modeling the Blood donor data set (Nasserinejad et al., 2015) using a Bayesian heterogeneity model.

I am grateful to my supervisor Professor Emmanuel Lesaffre for keeping faith in my capabilities and for guiding me in the right direction. I enjoyed the fact that he never spoon-fed me, yet was always approachable to discuss the statistical problems. He set very clear goals at the beginning of the year and continually monitored my progress thereafter. My interest in Bayesian statistics has grown by magnitudes under his supervision and I am looking forward to contribute more in this area. I would also like to extend my gratitude to Professor Molenberghs and Professor Verbeke for the captivating lectures on longitudinal data analysis, without which the analysis of the blood donor data set would have been difficult. I am thankful to Kazem Nasserinejad from ErasmusMC for resolving many of my queries regarding the blood donor data set, and to Igor Milhorança for providing the much needed inputs at crucial times. I would also like to thank Jorne Biccler for teaching me that statistics is very much a forensic science. Lastly, I am grateful to my parents for the innumerable sacrifices they made to make sure that I had as less obstacles as possible during my studies and I dedicate this work to them.

Anirudh Tomer
Leuven, Belgium

Summary

In this master thesis we fitted a finite mixture distribution for the random effects in a Bayesian linear mixed model. Our goal was to evaluate the efficacy of Bayesian model selection methods, namely Deviance information criteria, marginal likelihood and posterior predictive checks for selection of the model with the right number of components in the aforementioned mixture. For this we generated multiple artificial data sets with different types of Gaussian mixture of random effects and applied the model selection criteria on them. Since mixture models are missing data models, we implemented various definitions of DIC as given by Celeux et al., (2006). We found that conditional data DIC's which are usually reported by software such as JAGS are not reliable for selecting the number of mixture components. DIC 4 (section 4.8) which was based on complete data likelihood performed the best among all of the DIC's. The next best performing DIC was DIC 3(section 4.4) based on marginal data likelihood. We recommend using these DIC's along with posterior predictive checks (PPC) which also worked very well to detect overfitting. We found that if inverse gamma priors were used for variance components, and uniform distribution for correlation in the distribution of random effects, then PPC's based on such models gave more extreme results in presence of overfitting the number of mixture components. We also calculated marginal likelihood for the various models using the approximation given by Chib, (1995) and found that it was not reliable for deciding the number of components required in the mixture of random effects.

While doing MCMC simulations we found some other interesting results as well, which although were not of primary interest, but are still worth mentioning here. We found that a Wishart prior for precision matrix(inverse of covariance matrix) of mixture components lead to posteriors which overestimated the precision when within subject variance was greater than between subject variance. Further we also found that if mixture components are not well separated and the number of subjects are small, then using a Dirichlet prior with small values of hyperparameter (≤ 1) for the weight distribution of mixture components can lead to the choice of underfitted models.

Contents

List of Tables

List of Figures

x

Chapter 1

Introduction

In this chapter we introduce the concept of a mixture distribution and the challenges involved in estimation of parameters of a mixture distribution. We also highlight the benefits of using a Bayesian approach for parameter estimation. Lastly we present the goal of this master thesis, in which a mixture distribution plays the central role.

1.1 Mixture distribution

A mixture distribution is a probability distribution of a random variable formed from a group of other random variables. The formation of a mixture distribution can be seen as a two step process. In the first step a particular random variable is selected from a collection of random variables based on a certain probability of selection. In the second step a value is sampled for the selected random variable from its probability distribution. For e.g. the following random variable Y has a mixture density formed from 3 normally distributed random variables.

$$Y \sim \frac{1}{6}N(-10,3) + \frac{1}{2}N(0,1) + \frac{1}{3}N(4,2)$$

Figure 1.1 shows the density function for Y. The density is trimodal with each mode corresponding to one of the components in the mixture. Mixtures such as Y which are formed from a finite sum of components are called finite mixtures. The components are also known as mixture components and their densities are called component densities. The constants multiplying the corresponding densities are called mixture weights. The mixture weights also represent the probability of selection of the component densities. Each mixture weight should be positive and the sum of all mixture weights should be equal to 1. While in our example all the mixture components belonged to the same parametric family i.e. normal distribution, it is also possible to have mixture components from different parametric families. A mixture model where it is assumed that all data points are generated from a mixture of normally distributed component densities is called Gaussian mixture model (GMM). It is important to note that the idea of a mixture distribution is rather hypothetical, as it was shown in an example that a GMM of two components could be indistinguishable from a log-normal distribution, by Titterington, Smith, and Makov, (1986).

1.1.1 Formal definition for finite mixture distribution

Given a finite set of K probability density functions $p_1(y), p_2(y), \ldots, p_K(y)$ and weights $\eta_1, \eta_2, \ldots, \eta_K$, a random variable Y is said to have a finite mixture distribution if

$$p(y) = \sum_{k=1}^{K} \eta_k p_k(y)$$
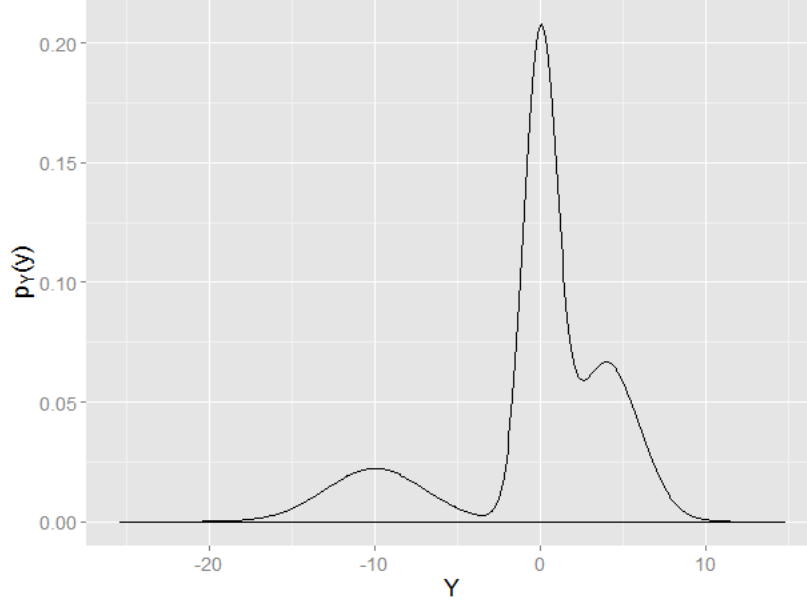
Figure 1.1: Mixture density of $Y \sim \frac{1}{6}N(-10, 3) + \frac{1}{2}N(0, 1) + \frac{1}{3}N(4, 2)$

The vector of the weights $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)^T$ is called the weight distribution. The kth weight η_k corresponds to selection probability of the kth component while sampling for Y. η_k can only take values from the K dimensional positive real coordinate space $\mathbb{R}^{+K}$ with an additional constraint, $\sum_{k=1}^{K} \eta_k = 1$.

1.1.2 Challenges

The primary challenge while modeling a mixture density for a random variable is that the number of mixture components (K), weight distribution $(\boldsymbol{\eta})$ and the corresponding parameters for component densities are rarely known in advance. Secondly, from a sample of N observations $y_1, y_2, \ldots, y_N$ sampled from the mixture density $p(y)$ one may not know which observation belongs to which component density. Formally, an allocation vector $\boldsymbol{S} = (S_1, S_2, \ldots, S_N)^T$ represents the allocation of observations to mixture components. i.e. $S_i = k$ represents that i^{th} observation belongs to k^{th} component density. Estimating the allocation vector is in fact solving the clustering problem, albeit using parametric methods in our case.

While Maximum likelihood (ML) based methods such as the EM algorithm can be used to deal with the challenges mentioned above, there are certain downsides to them. Firstly it is well known that 95% confidence intervals of ML estimates are based on asymptotic normality of the estimators. Thus in case of small sample size, or small mixture weights the results will not be correct (Frühwirth-Schnatter, 2013, pg. 35). A Bayesian approach however is immune to these issues as the posterior distribution of parameters is allowed to be non-normal. For univariate and multivariate GMM the likelihood function is given by,

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\sigma^2}, \boldsymbol{\eta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} f_N(y_i; \mu_k, \sigma_k^2)\eta_k$$

where $f_N(y_i; \mu_k, \sigma_k^2)$ denotes the density of the normal $N(\mu_k, \sigma_k^2)$ distribution. This likelihood function is unbounded and has many spurious nodes near the boundary of the parameter

space for variance(σ_k^2) of the components (Day, 1969; Kiefer and Wolfowitz, 1956). A Bayesian approach however, handles this problem elegantly using priors for parameters of the component densities. For e.g. Frühwirth-Schnatter, (2013, pg. 176) combined the likelihood with the prior $p(\mu_k, \sigma_k^2) \propto p(\sigma_k^2), \sigma_k^2 \sim$ Inv-Gamma$(1, 4)$, and showed that it leads to a joint posterior density $p(\mu, \sigma^2|\boldsymbol{y})$ of parameters in which σ_k^2 was bounded away from 0. Thus all the spurious nodes near the boundary of the parameter space for σ_k^2 were cut out, whereas they were apparent from the likelihood function.

1.1.3 Applications of mixture distributions

Mixture models have found usage in a variety of domains. Some of the examples are:

- Spike sorting of neural data: Both GMM and mixture of multivariate t-distributions have been used.(Lewicki, 1994; Shoham, Fellows, and Normann, 2003).

- Speaker recognition as well as speech to text conversion algorithms have used mixture models (Povey et al., 2011; Simancas-Acevedo et al., 2001; Xiang and Berger, 2003).

- Image processing: GMM have been used to find features in an image, such as objects, boundaries etc. (Fu and Wang, 2012). For e.g. Yang, (1998) have used GMM to model the distribution of skin color pixels. Many authors have also proposed using GMM for face recognition. i.e. as a biometric identification mechanism.

- Finance: Brigo and Mercurio, (2002) proposed to use a log-normal mixture distribution for pricing of financial assets.

- Biology: Mixture models have found usage in genetics and cell biology.(Gianola et al., 2007; Sim et al., 2012)

The example applications we cited involved usage of mixture models to adjust for a hidden attribute in the data which could not be collected or to approximate a density which was not of a known form. However mixtures have also been used as supplementary method in various models, a list of which can be found in Frühwirth-Schnatter, (2013, pg. 238). One such usage in linear mixed models was proposed by Verbeke and Lesaffre, (1996) and it also formed the problem statement of this thesis.

1.2 Goal of master thesis

Verbeke and Lesaffre, (1996) proposed to use a finite mixture distribution of normally distributed components for the prior distribution of random effects in a linear mixed effects model (LMM). This particular LMM is also known as heterogeneity model. In this thesis our focus was on the Bayesian version of the heterogeneity model, where all parameters involved are assigned a probability distribution. Needless to say, the issues described in section 1.1.2 are also applicable for the Bayesian heterogeneity model. The aim of this master thesis was to evaluate existing Bayesian approaches for model selection, namely Deviance Information Criterion (DIC), marginal likelihood and posterior predictive checks(PPC) for selecting the right number of mixture components for the distribution of random effects. Since we have worked within the Bayesian framework, we used MCMC methods instead of the frequentist point estimation methods. We also generated artificial data sets to check efficacy of each of the aforementioned model selection criteria and then used the most effective ones to decide the number of mixture components for the random effects distribution in blood donor longitudinal data set (Nasserinejad et al., 2015).

Chapter 2

Bayesian paradigm

In this chapter we introduce the basics of the Bayesian paradigm. i.e. Bayes rule and Bayesian summary measures.

2.1   The Bayesian motivation: A toy example

What primarily differentiates the Bayesian paradigm from frequentist paradigm is that the parameters θ are random variables rather than being a constant. The distribution of parameters based on the data is called the posterior distribution and can be represented as $p(\theta|y)$. Whereas the initial distribution of parameters is called the prior distribution, represented by $p(\theta)$. To further signify the ideological differences between the Bayesian paradigm and frequentist paradigm we are presenting an example next.

Suppose there are three people A, B and C, of whom A and B each are captains of a sports team and C is the referee who tosses the coin. Let us assume that based on experiences of an old friend, captain B gets to know that the referee purposefully attempts at getting a heads on the toss. However given the nature of this problem, it is hard to quantify this belief in a single number. Instead a belief that there is a 70 to 90% chance that the result will be a heads is more likely than a belief that there is exactly an 80% chance for the same. One might also have a slightly vague belief that there is more than 50% chance that the toss will result into a heads. Secondly, given the fact that not all coins are alike it is impossible for the probability of getting a heads to be constant, even if the referee tosses identically on each trial.

(a) Prior $p(\pi)$: $\beta(9,3)$ (b) Posterior $p(\pi|\boldsymbol{y})$: $\beta(15,7)$
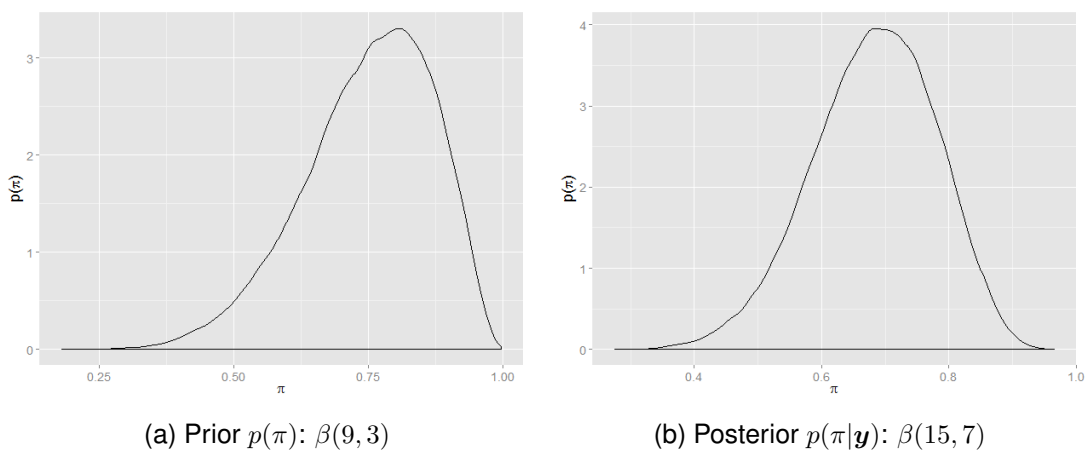
Figure 2.1: Prior and posterior PDF for π; the probability of getting heads.

While the beliefs of captain B are subjective, they represent the prior probability distribution of a random variable in Bayesian paradigm. In our example the random variable is probability (π) of getting a heads. In figure 2.1a we can see one such prior distribution corresponding to the belief that the chance of getting a heads on the toss is more than getting tails and it is more likely to be somewhere between 70 to 90%.

2.2 Bayes rule

We now present Bayes rule which is central to the Bayesian parameter estimation process. Bayes rule for the estimating the continuous parameter π is given by

$$p(\pi|\boldsymbol{y}) = \frac{L(\pi|\boldsymbol{y})p(\pi)}{p(\boldsymbol{y})} = \frac{L(\pi|\boldsymbol{y})p(\pi)}{\int L(\pi|\boldsymbol{y})p(\pi)\,\mathrm{d}\pi} \tag{2.1}$$

The result $p(\pi|\boldsymbol{y})$ is called the posterior distribution of the parameter. It can be used to make statistical inference about the parameter π. An intuitive way to get the motivation behind the Bayes rule is to imagine the denominator as marginal probability of $\boldsymbol{y}$ calculated using the law of total probability.

It is also possible to apply Bayes rule for estimating parameters in context of the current example. Suppose after 10 matches captain B observed that 6 times out 10 the toss resulted in a heads. Assuming that the tosses were independent, then given the likelihood function $L(\pi|\boldsymbol{y})$, the MLE of π will be $\hat{\pi} = 0.6$. Whereas Bayes rule gives us the entire posterior distribution of parameter π as shown in figure 2.1b. The mean value $E(p(\pi|\boldsymbol{y}))$ of the posterior distribution is 0.7, which if we compare with the MLE $\hat{\pi}$=0.6 we can see that Bayesian posterior mean is influenced by the prior as well.

2.3 The role of prior distribution

We can see in equation 2.1 that the computation of posterior involves solving the integral in the denominator. One can avoid solving the integral by choosing a prior such that the resulting posterior is from the same parametric family as the prior and thus available in closed form. Such priors are termed as conjugate priors. However it is not always feasible to choose a conjugate prior and numerical approximation for calculation of the posterior is required. The most widely used algorithms for posterior approximation are Markov chain Monte Carlo (MCMC) techniques such as Gibbs sampling, Metropolis hastings algorithm, Hamiltonian Monte Carlo and their variants etc. The priors can also be classified as informative or non-informative/vague/diffuse. The prior we chose in our example was informative, whereas a diffuse prior could have been the uniform distribution $U(0, 1)$. In absence of prior knowledge a non informative prior is advised. A more detailed overview of the priors can be found in Lesaffre and Lawson, (2012).

2.4 Bayesian inference

Given the posterior distribution of a parameter $p(\theta|y)$ one can use the point estimates such as median, mean $E_\theta(\theta|y)$, or MAP (maximum a posteriori) $\arg\max_\theta p(\theta|y)$ for inference. It is however the interval estimates where the Bayesian paradigm contrasts more with frequentist approach. Bayesian 95% interval estimates are called credible intervals. While the frequentist 95% confidence interval is interpreted as the interval in which 95 out of 100 times one can find the population parameter θ, the Bayesian 95% credible interval is interpreted as interval of the 95% most plausible theta values a posteriori. The credible interval can be equal tailed or

a highest posterior density interval (HPDI). The Bayesian paradigm also allows one to make inference on future values of the data by taking the current data into account. This is done using the posterior predictive distribution (PPD)

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\,\mathrm{d}\theta$$

The point and interval summary measures for PPD are similar to the ones for posterior distribution of parameters $p(\theta|y)$. We will also discuss Bayesian model selection in the forthcoming chapters.

2.5   Bayesian software

Various Bayesian software tools such as BUGS, STAN, PROC MCMC in SAS etc. are used for running the MCMC procedures mentioned above. For the purpose of this thesis we used JAGS, which is strongly related to BUGS (Bayesian inference Using Gibbs Sampling) family. We also used the R package R2jags to execute JAGS code via R.

Chapter 3

Bayesian linear mixed effects model

3.1 Introduction to linear mixed model

A linear mixed effects model, also known as linear mixed model(LMM) is a statistical model for hierarchical data. For e.g. one such hierarchy could be, repeated observations taken from multiple patients and patients grouped under multiple hospitals. The specialty of these models is that apart from the fixed effects, they also model the correlation between the observations falling in the same group at a certain level in the hierarchy. The correlation is modeled using a combination of random effects and measurement error, and the response is modeled as a linear function of both fixed and random effects.

There are many synonymous terminologies for data sets which are hierarchical in nature albeit with subtle nuances differentiating them. In this thesis our focus was on longitudinal data sets. A longitudinal data set is the one where multiple observations are collected from subjects at different points in time. For e.g. measurement of hemoglobin of 20 patients with observations taken every month for a period of 24 months. The observations collected from a subject are correlated in such a case, and a linear mixed model provides a suitable framework to model that correlation.

3.1.1 LMM definition

The LMM for the observations of the ith subject among the n subjects can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \tag{3.1}$$

where $1 \le i \le n$,
$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T$ is a vector of observations for the ith subject taken at m_i time points, and $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, ..., \boldsymbol{y}_n^T)^T$ represents the entire vector of observed responses,
$\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \ldots, \boldsymbol{x}_{im_i}^T)^T$ is the $m_i \times (d+1)$ design matrix for the ith subject in the data set, and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n)^T$ is the entire design matrix for the n subjects,
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ is a $(d+1) \times 1$ vector of fixed effects with β_0 being the intercept,
$\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}^T, \boldsymbol{z}_{i2}^T, \ldots, \boldsymbol{z}_{im_i}^T)^T$ is the $m_i \times q$ design matrix of covariates multiplying the random effects,
$\boldsymbol{b}_i = (b_{0i}, b_{1i}, \ldots, b_{(q-1)i})^T$ is a $q \times 1$ vector of random effects with b_{0i} being the random intercept. The random effects $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, G)$ with G being the $q \times q$ covariance matrix,
$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{im_i})^T$ is a $m_i \times 1$ vector of measurement errors. The errors $\boldsymbol{\varepsilon}_i \sim N_{m_i}(\boldsymbol{0}, R_i)$ with R_i being the $(m_i \times m_i)$ covariance matrix of errors,

The errors $\boldsymbol{\varepsilon}_i$ and the random effects $\boldsymbol{b}_i$ are assumed to be independent. R_i is usually a diagonal matrix of the form $\sigma^2 I_{m_i}$. While one might only model the correlation between the observations

of a subject using random effects, it is also possible to model the serial correlation component in which case R_i is not a diagonal matrix.

3.2 Motivation for Bayesian linear mixed model

One of issues with the frequentist LMM is that while the parameters in matrices G and R_i are estimated using ML/REML, only a point estimate is further used in estimation of fixed effects (see Verbeke and Molenberghs, 2009, chap. 5). Hence the uncertainty in estimation of random effects is ignored. Although frequentist inference approaches try to mitigate this issue by modifying the distributional assumptions of the test statistic for fixed effects (Verbeke and Molenberghs, 2009, pg. 56), a Bayesian approach takes into account the variability in parameter estimates natively. A similar problem occurs in the estimation of random effects $\boldsymbol{b}_i$. The frequentist strategy is to use Empirical Bayes estimates, where the posterior distribution of random effects use point estimates of parameters G and R_i. Thus the uncertainty in estimation is ignored again. On the other hand the Bayesian approach averages over the entire posterior distribution of the hyperparameters to obtain the posterior $p(\boldsymbol{b}_i|\boldsymbol{y})$. In light of these reasons, we have modeled our data using Bayesian linear mixed models.

The Bayesian linear mixed model or BLMM can be obtained by assigning a distribution to all the parameters involved in a LMM. It means that the model presented in section 3.1.1 can be extended by giving a prior distribution for the following:

- $\sigma^2 \sim p(\sigma^2)$

- $\boldsymbol{\beta} \sim p(\boldsymbol{\beta})$

- $G \sim p(G)$

3.3 Motivation for mixture of random effects

As we saw above, in a LMM the random effects are assumed to be multivariate normally distributed. It could be too strong an assumption though in certain cases. An example is the longitudinal studies where at a given time point we would like to categorize subjects in groups. For e.g. group with a high risk of having a certain disease in future vs. group with a low risk. While in retrospective studies this task is easier as we know exactly which patients were diagnosed with the disease and which were not, this could be difficult in a study where we would like to categorize patients into different groups well before diagnosis. Here is a toy example: Imagine that in a longitudinal study we are measuring a response Y which is a continuous indicator of a disease. Assume that from a previous study it is known that patients which are in high risk group for the disease tend to have a higher response Y during all times. Also assume that the trend of Y over time remains the same for both groups otherwise. Figure 3.1 shows individual profiles of such subjects from a simulated dataset. Looking at this plot we can say that a random intercept component will be enough to model individual profiles. Since we do not know which patient belongs to which group, this heterogeneity can be appropriately modeled by considering that the random intercept is a mixture of two normal components. Another reason for using a mixture distribution is that the random effects distribution may not be of a known form and the mixture distribution may very well approximate it.

In a LMM is quite common to use histogram of Empirical Bayes estimates of random effects to detect groups of individuals. However Verbeke and Lesaffre, (1996) have shown that if the prior is misspecified (for e.g. if in our example we use a univariate normal distribution), then the
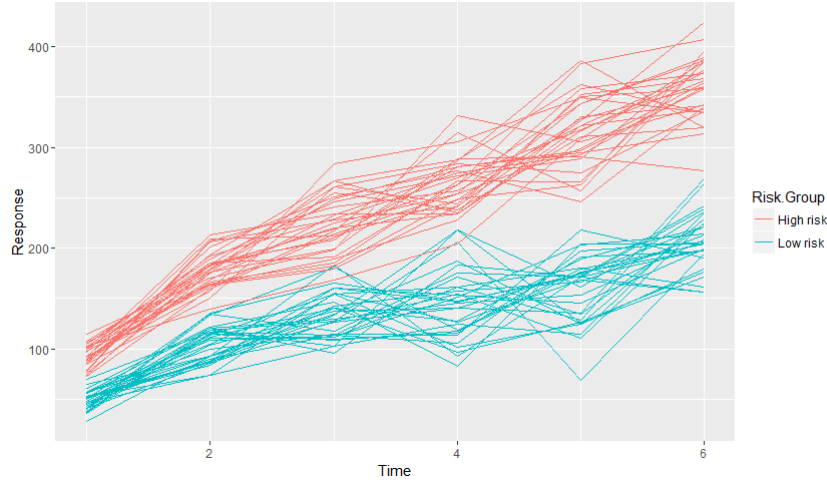
Figure 3.1: Individual profiles of 30 subjects from each group.

histogram of estimates of random effects will be shrunk towards the prior distribution. Thus it would be impossible to classify the subjects into different categories based on Empirical Bayes estimates of random effects as they are incorrect. A solution to this problem is using a mixture of Gaussian components for random effects distribution. Such a linear mixed model is termed as a heterogeneity model.

3.3.1 Bayesian heterogeneity model

The formal definition of a Bayesian heterogeneity model can be given by extending the Bayesian linear mixed model definition given in section 3.2. Since, now the random effects have a Gaussian mixture distribution we will use the following notation to express the distribution mathematically.

$$\boldsymbol{b}_i \sim \sum_{k=1}^{K} \eta_k N_q(\boldsymbol{b}_k^C, G_k)$$

where $\boldsymbol{b}_k^C$ and G_k are the mean vector and covariance matrices for the kth component in the mixture distribution respectively. The vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)^T$ is the weight distribution for the component densities. The vector $\boldsymbol{S} = (S_1, S_2, ..., S_n)^T$ represents the allocation vector for the n subjects. Since we are following the Bayesian paradigm, in addition to prior distribution for β and σ^2 we also have prior for $\boldsymbol{\nu} = (\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, \ldots, \boldsymbol{b}_K^C, G_1, G_2, \ldots, G_K, \boldsymbol{\eta})$.

3.4 Estimation of parameters in the Bayesian heterogeneity model

In this section we discuss some of the challenges in Bayesian estimation of parameters in the Bayesian heterogeneity model. We also discuss the approaches we used to deal with them in this thesis.

3.4.1 Marginal vs. hierarchical model

Suppose that in the heterogeneity model the random effect $\boldsymbol{b}_i$ are known, then following LMM equation has a hierarchical interpretation.

$$\begin{aligned} \boldsymbol{y}_i | \boldsymbol{b}_i &\sim N(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i, \boldsymbol{\varepsilon}_i) \\ \boldsymbol{\varepsilon}_i &\sim N_{m_i}(\boldsymbol{0}, R_i) \end{aligned} \tag{3.2}$$

11

It is important to note that the distribution of y_i does not depend on allocation S_i directly due to hierarchical independence. i.e. It is b_i which depends on S_i because $b_i|S_i \sim N(b_{S_i}^C, G_{S_i})$. However if one knows the random effect then knowing the allocation is not necessary in the hierarchical model. i.e. $p(y_i|b_i, S_i) = p(y_i|b_i)$. Though it is possible integrate out the random effects b_i and obtain the corresponding Bayesian heterogeneity model where y_i depend on allocation S_i.

$$y_i|S_i \sim N(X_i\beta + Z_i b_{S_i}^C, \varepsilon_i^*)$$
$$\varepsilon_i^* \sim N_{m_i}(0, Z_i G_{S_i} Z_i^T + R_i) \tag{3.3}$$

The marginal model is recommended by Frühwirth-Schnatter, Tüchler, and Otter, (2004) for good mixing of chains, and while doing the simulation study (presented in chapter 5) we found that claim to be true. However, we observed that estimating parameters using a marginal model took quite a long time for each iteration. It also did not give posterior estimates of the random effects b_i which were required for calculation of certain definitions of DIC (discussed in chapter 4). Besides we found that a model with hierarchical centering took less time for each iteration and had as much autocorrelation in the posterior density samples as with the use of the marginal model.

3.4.2 Hierarchical centering

The random effects b_i in a mixed model can be seen as random deviations from the fixed effects (β) with a mean 0. For a longitudinal data set, it means that the overall effect of a covariate such as the intercept for a subject should be the sum of both fixed and random effects. In this case matrices X and Z both share columns corresponding to the variable intercept. To enforce the mean 0 on the random effects in a mixture distribution of random effects, the following condition should be satisfied.

$$E(b_i|\nu) = \sum_{k=1}^{K} \eta_k N_q(b_k^C, G_k) = 0 \tag{3.4}$$

where ν is defined in section 3.3.1. This further means that $E(y_i|\nu) = X_i\beta$. This parametrization, which was also used in the original paper on heterogeneity model (Verbeke and Lesaffre, 1996) is called the non-centralized parametrization. The centralized parametrization assumes that the random effects are centered around a non zero mean. The choice of parametrization has an effect on the rate of convergence of the chains in MCMC process. While doing the simulation study we observed that imposing the constraint in equation 3.4 drastically slowed the convergence as well increased the autocorrelation in parameter estimates. Thus, in this thesis we have only used centralized parametrization.

3.4.3 Starting values

The choice of starting values is important in mixtures especially when the components are not well separated. In the R package bayesmix (Gruen and Plummer, 2015) the authors used $\frac{1}{k+1}, \frac{2}{k+1}, ..., \frac{2}{k+1}$ quantiles of the sample data for the starting values of the means $\mu_1, \mu_2, ..., \mu_K$ of the K components. In the Bayesian heterogeneity model we extracted the random component from the data as described in section 5.1.1 and then calculated the sample quantiles. We found out that it resulted into chains with an improved convergence. We also calculated starting values for the fixed effects β using OLS because OLS parameter estimates are unbiased and consistent [pg. 50](Verbeke and Molenberghs, 2009).

3.4.4 Choice of priors

Since we followed the Bayesian paradigm and the parameters in the Bayesian heterogeneity model are random variables and we were required to assign a prior distribution to the parameters. For the kth component in the mixture of random effects, both the mean b_k^C and covariance matrix G_k were unknown. To obtain the joint posterior of these parameters as a known density one the conditionally conjugate prior $b_k^C | G_k \sim N(\boldsymbol{\mu}_0, \frac{G_K}{g_0})$ and $G_k^{-1} \sim \mathcal{W}(n_0, \Psi)$ was suitable. Here $\boldsymbol{\mu}_0, g_0, n_0, \Psi$ are the hyperparameters for the corresponding prior distributions. During our simulations we found that choosing $g_0 = 1$ yielded very high values for variance components whereas for $g_0 > 1$ JAGS was unable to find a sampler. We thus stuck to widespread alternative practice of specifying independent priors for the mean (b_k^C) and covariance matrix (G_k) (Gelman and Hill, 2006, chap. 17). For e.g. a common non informative prior for b_k^C (say, having only random intercept and slope) is $N(\mathbf{0}, \begin{bmatrix} 10^5 & 0 \\ 0 & 10^5 \end{bmatrix})$. This prior is equivalent of specifying independent diffuse univariate normal priors for the mean of random intercept and random slope respectively.

Choice of prior for covariance matrix

The choice of prior for the covariance matrix (G_k) posed an interesting problem to us. Lesaffre and Lawson, (2012, pg. 260) suggest using an inverse Wishart prior with small diagonal elements for the scale hyperparameter and degrees of freedom hyperparameter equal to the dimension of G_k. For e.g $IW(\begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, 2)$ could be one such prior. The corresponding prior for precision matrix G_k^{-1} is the Wishart prior $W(\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}, 2)$. i.e. the scale of Wishart prior is inverse of the scale hyperparameter for inverse Wishart distribution. As we found later in our simulations, a big value for diagonal elements of scale matrix of Wishart distribution influenced the posterior more than the likelihood did.

Another option is to use independent inverse gamma priors for random intercept and random slope variances and uniform prior $U(-1, 1)$ for the correlation between the two random effects. The upside of this approach is that it may give almost the same estimates as one can get from frequentist analysis, but the downside is that MCMC iterations are slower because the posterior is not available as a known density. Another benefit of this approach, as we later found out during simulations is that when the mixture distribution is overfitted, then the extra components tend to have very high variance estimates for random intercept and random slope. We also used this property to design a posterior predictive check.

Choice of priors for β and σ^2

We assumed that the parameters β and σ^2 are independent from $b_1^C, b_2^C, ..., b_K^C, G_1, G_2, ..., G_K, \boldsymbol{\eta}$. The problem of choosing a conjugate prior for β and σ^2 is similar to what we discussed in the section above. The solution is also alike, i.e. using independent univariate normal priors such as $N(0, 10000)$ for each of the β_d and a Gamma$(0.0001, 0.0001)$ prior for $\tau = \frac{1}{\sigma^2}$ (Gelman and Hill, 2006, chap. 17).

Choice of prior for η

The conjugate prior for the weight distribution $\boldsymbol{\eta}$ is the Dirichlet prior Dir$(a_0, a_1, ..., a_K)$. Frühwirth-Schnatter, (2013, pg. 105) suggest choosing values of hyperparameters $a_0, a_1, ..., a_K$ to be

greater than 1 in data sets where one of the components is nearly empty or the components are not well separated. For such data sets we observed label switching (section 3.4.5) when we used Dirichlet prior with hyperparameter value equal to 1 or less than 1. We found out during the simulation study that choosing slightly larger values for the hyperparameter, such as 2 or 3, indeed mitigated the issue of label switching (section 5.1.3), however very large values also made the prior very informative.

3.4.5 Label Switching

We used a mixture distribution for random effects in the Bayesian heterogeneity model and thus the allocation vector S for the mixture was not known in advance. In this case the mixture likelihood for the response y is given by equation 4.1. The mixture likelihood function is symmetrical and has $K!$ modes (Frühwirth-Schnatter, 2013, pg. 44). This creates a problem called label switching while running the MCMC procedure.

The label switching problem can be explained with the following example. Suppose we have a mixture distribution $0.5N(5,1) + 0.5N(7,1)$ of two components C_1 and C_2 and we have few observations sampled from the mixture. Using the MCMC procedure we can one estimate the parameters of the two components. The MCMC procedure for missing data models like mixture models uses a technique called data augmentation. The idea of data augmentation is similar to the frequentist EM algorithm. i.e. we begin with some random allocation vector S_{initial} and estimate parameters using the complete data likelihood. An example expression of a complete data likelihood for Bayesian heterogeneity model is expression 4.6. For the MCMC sampler, labels μ_1 and μ_2 exist for the two means, however either one can correspond to μ_{C1} or μ_{C2}. i.e. labels are not associated with actual components from the beginning. Assume that the allocation vector S_{initial} is such that it assigns all observations from component C_1 under label 1 and all observations from component C_2 under label 2. Under such a scheme a posterior sample $(\mu_1, \mu_2) = (5, 7)$ is likely. However if we take a conjugate of this allocation vector then $(\mu_1, \mu_2) = (7, 5)$ is also likely to be sampled. This can be attributed to the fact that we have a mixture likelihood function which is bimodal. In cases where the components are not well separated, because of a certain scheme of allocations S, the sampler might sample μ_1 from both modal regions of the likelihood resulting into a posterior which is bimodal as well. However it can also lead to partially explored posteriors, which may not be useful for making any inference.

Dealing with label switching

One of the techniques used for dealing with label switching is imposing a formal identifiability constraint such as $\mu_1 < \mu_2$. However Frühwirth-Schnatter, Tüchler, and Otter, (2004) suggest that arbitrary identifiability constraints should not be applied as they are often ineffective. Instead in an example they chose an identifiability constraint based on pre-analysis of the various modal regions. In the case of Bayesian heterogeneity model the mean is a vector comprising of random intercept and random slope means. In this thesis we have applied identifiability constraints on the mean vector based on graphical exploration of the mixture density as shown in section 5.1.1. It is important to note that if more components than needed are chosen, then label switching is unavoidable, and could also be seen as an indicator for overfitting (Frühwirth-Schnatter, 2013, pg. 104).

Another technique to deal with label switching is post-processing of MCMC chains by relabeling the output(Richardson and Green, 1997; Stephens, 2000). We too employed this technique, albeit in a very simplified form approximating marginal likelihood using Chib's approximation (section 4.2).

Chapter 4

Model selection criteria

In practice the right number of components in a mixture distribution are not known in advance. In this chapter we describe 3 of the Bayesian model selection criteria to select the model with the right number of mixture components in a Bayesian heterogeneity model.

4.1 Deviance information criteria

The Deviance information criteria or DIC was first proposed by Spiegelhalter et al., (2002) for Bayesian model selection. The motivation for DIC is similar to frequentist AIC/BIC criteria in the sense that DIC also penalizes more elaborate models using a penalty component. The definition for DIC is given by

$$\text{DIC} = -2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}) + 2\mathsf{p}_\mathsf{D}$$

where $\bar{\boldsymbol{\theta}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}) \,\mathrm{d}\boldsymbol{\theta}$ is the posterior mean of the parameters. The penalty for model complexity is given by $\mathsf{p}_\mathsf{D} = -2\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$. It can also be written as,

$$\mathsf{p}_\mathsf{D} = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$$

where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of Bayesian deviance $D(\boldsymbol{\theta})$, and $D(\bar{\boldsymbol{\theta}})$ is the Bayesian deviance evaluated at the posterior mean of the parameters. Bayesian deviance is defined as $D(\boldsymbol{\theta}) = -2\log p(\boldsymbol{y}|\boldsymbol{\theta}) + 2\log f(\boldsymbol{y})$. We do not discuss the term $f(\boldsymbol{y})$ because it cancels out in the expression for p_D.

4.1.1 DIC for missing data models

Mixture models and mixed models are both a member of the class of models called missing data models. The allocation vector S in a mixture model, and the matrix of random effects $b = (b_1^T, b_2^T, ..., b_n^T)^T$ in a LMM, are not observed directly. Thus one could have various incompatible definitions of DIC based on observed data likelihood, complete data likelihood and conditional data likelihood, as shown by DeIorio and Robert in a discussion on the paper of Spiegelhalter et al., (2002). Further, Celeux et al., (2006) proposed multiple definitions of DIC under each of the aforementioned likelihood classes and showed that each has a different value and thus a different impact on model selection. In this thesis we took some of those definitions and applied them on the Bayesian heterogeneity model.

Observed DIC

Let us denote the set of all the unknown parameters in the Bayesian heterogeneity model as $\boldsymbol{\theta}$. i.e. $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\nu})$, where $\boldsymbol{\nu}$ is as defined in section 3.3.1. The first category of DIC is associated

with observed data likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ which can be obtained by marginalizing over the allocation S_i and random effects $\boldsymbol{b}_i$ for the subjects. The following expression gives the observed data likelihood for the Bayesian heterogeneity model.

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{Z}_iG_k\boldsymbol{Z}_i^T + R_i)\eta_k \tag{4.1}$$

where $f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{Z}_iG_k\boldsymbol{Z}_i^T + R_i)$ denotes the density of the normal $N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{Z}_iG_k\boldsymbol{Z}_i^T + R_i)$ distribution. Based on equation 4.1 we extended the various definitions of DIC proposed by Celeux et al., (2006). The first one is the classical definition for DIC, as shown below.

$$\text{DIC}_1 = -4\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}})) \tag{4.2}$$

One of the problems with DIC_1 is that in case of label switching, one may obtain a negative p_D. In our simulation study we encountered this problem as well (section 5.1.3). The reason is that when label switching happens, the mean of the posteriors may take a value which lies somewhere in between the various modes. On the other hand, in such a scenario a posterior mode will belong to only one of the modal regions. Thus the next definition of DIC is formed by replacing posterior mean with the posterior mode in the calculation of $D(\bar{\boldsymbol{\theta}})$.

$$\text{DIC}_2 = -4\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_M)) \tag{4.3}$$

where $\hat{\boldsymbol{\theta}}_M = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior mode of the parameters. Celeux et al., (2006) further suggest that for models where non identifiability of parameters is endemic, as in the case of mixtures, one should use the following MCMC estimator for the approximation of $p(\boldsymbol{y}|\boldsymbol{\theta})$,

$$\hat{p}(\boldsymbol{y}) = \prod_{i=1}^{n} \hat{p}(\boldsymbol{y}_i) = \prod_{i=1}^{n} \frac{1}{m} \sum_{l=1}^{m} \sum_{k=1}^{K} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}^{(l)} + \boldsymbol{Z}_i\boldsymbol{b}_k^{C(l)}, \boldsymbol{Z}_iG_k^{(l)}\boldsymbol{Z}_i^T + R_i^{(l)})\eta_k^{(l)}$$

where $\boldsymbol{\beta}^{(l)}, \boldsymbol{b}_k^{C(l)}, G_k^{(l)}, R_i^{(l)}, \eta_k^{(l)}$ are the samples from the lth MCMC iteration of a chain of length m, and $\hat{p}(\boldsymbol{y})$ is the MCMC estimator for $p(\boldsymbol{y}|\boldsymbol{\theta})$. The estimator is also called the MCMC predictive density estimator, and gives us the next definition of DIC.

$$\text{DIC}_3 = -4\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) + 2\log \hat{p}(\boldsymbol{y}) \tag{4.4}$$

The benefit of using DIC_3 is that p_D always takes a positive value. While doing the simulation study we found that this claim was not only true, but p_D also never took extreme values; even when models were severely overfitted leading to label switching. Lastly, in each of the equations 4.2, 4.3, 4.4, $\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}))$ can be calculated by using the following approximation.

$$\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|\boldsymbol{\theta})) = \frac{1}{m} \sum_{l=1}^{m} \log p(\boldsymbol{y}|\boldsymbol{\theta}^{(l)}) \tag{4.5}$$

Complete DIC

The second class of the DIC is based on the complete data likelihood. The complete data for the ith subject in a Bayesian heterogeneity model is based on $(\boldsymbol{y}_i, S_i, \boldsymbol{b}_i)$. For the ith subject we will denote the complete data by $\boldsymbol{y}_i^F$ and for the entire data set we will denote it by $\boldsymbol{y}^F$. The following equation shows the complete data likelihood for the Bayesian heterogeneity model.

$$\begin{aligned} f(\boldsymbol{y}^F|\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{b}_i, S_i)f(\boldsymbol{b}_i|S_i)f(S_i) \\ &= \prod_{i=1}^{n} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, R_i)f_N(\boldsymbol{b}_i; \boldsymbol{b}_{S_i}^C, G_{S_i})f(S_i; \boldsymbol{\eta}) \end{aligned} \tag{4.6}$$

16

Based on the complete data $\boldsymbol{y}^F$, the complete data DIC is given by following expression,

$$\text{DIC} = -4\text{E}_{\boldsymbol{\theta}|\boldsymbol{y}^F}(\log p(\boldsymbol{y}^F|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}^F|\bar{\boldsymbol{\theta}})) \tag{4.7}$$

Since part of the complete data are not observed, Celeux et al., (2006) propose integrating the expression in 4.7 with respect to the missing data $(\boldsymbol{b}, \boldsymbol{S})$ to obtain the following definition of DIC.

$$\text{DIC}_4 = -4\text{E}_{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\boldsymbol{\theta})) + 2\text{E}_{\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\bar{\boldsymbol{\theta}})) \tag{4.8}$$

The first part of DIC_4 is not available in closed form for the Bayesian heterogeneity model, however it can still be approximated using the output of Gibbs sampler in the same way as in equation 4.5. The motivation for this approximation is that during each iteration of the Gibbs sampler, parameter values are sampled from the conditional distribution of the parameters. i.e. in our case conditional on the other parameters and unobserved data both. We further verified the approximation by comparing the computed DIC_4 using the closed form solution for a mixture distribution, with the aforementioned approximation, both suggested by Celeux et al., (2006). We found that the two computed DIC values differ only in the decimal places.

While the second part of DIC_4, i.e. $\text{E}_{\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\bar{\boldsymbol{\theta}}))$ can also be approximated in the same way as in equation 4.5, but the added complexity here is that the posterior mean of the parameters is given by $\bar{\boldsymbol{\theta}} = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\boldsymbol{y}^F)\,\mathrm{d}\boldsymbol{\theta}$. The MCMC sampler only returns the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$, whereas $p(\boldsymbol{\theta}|\boldsymbol{y}^F)$ is required for DIC_4 because it is based on the complete data $\boldsymbol{y}^F$. Celeux et al., (2006) suggested using the posterior estimates of $\boldsymbol{b}, \boldsymbol{S}$ from $p(\boldsymbol{b}, \boldsymbol{S}|\boldsymbol{y})$ to analytically calculate the posterior distribution of parameters $p(\boldsymbol{\theta}|\boldsymbol{y}^F)$, from which the posterior means of the various parameters can be obtained as shown below.

$$\bar{\boldsymbol{\beta}}^{(l)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{H}^{(l)})$$

$$\bar{\sigma^2}^{(l)} = \frac{(\boldsymbol{y} - \boldsymbol{H}^{(l)} - \boldsymbol{X}\bar{\boldsymbol{\beta}}^{(l)})^T(\boldsymbol{y} - \boldsymbol{H}^{(l)} - \boldsymbol{X}\bar{\boldsymbol{\beta}}^{(l)})}{(\sum_{i=1}^n m_i - p - 1) - 2}$$

$$\bar{\boldsymbol{b}}_k^{C(l)} = \frac{\sum_{i=1}^n I(S_i^{(l)} = k)\boldsymbol{b}_{\boldsymbol{i}}^{(l)}}{n_k^{(l)}}$$

$$\bar{G}_k^{(l)} = \frac{\sum_{i=1}^n I(S_i^{(l)} = k)(\boldsymbol{b}_{\boldsymbol{i}}^{(l)} - \bar{\boldsymbol{b}}_k^{C(l)})(\boldsymbol{b}_{\boldsymbol{i}}^{(l)} - \bar{\boldsymbol{b}}_k^{C(l)})^T}{(n_k^{(l)} - 1) - \text{rank}(\bar{\boldsymbol{b}}_k^{C(l)}) - 1}$$

$$\bar{\eta}_k^{(l)} = \frac{a_k + n_k^{(l)}}{\sum_{u=1}^K a_u + n}$$

where, $\boldsymbol{H}^{(l)} = ((\boldsymbol{Z}_1\boldsymbol{b}_{\boldsymbol{1}}^{(l)})^T, (\boldsymbol{Z}_2\boldsymbol{b}_{\boldsymbol{2}}^{(l)})^T, ..., (\boldsymbol{Z}_n\boldsymbol{b}_{\boldsymbol{n}}^{(l)})^T)^T$ and $\bar{\boldsymbol{\beta}}^{(l)}$, $\bar{\sigma^2}^{(l)}$ and $\bar{\boldsymbol{\nu}}^{(l)}$ are the expected values of parameters of interest at the lth iteration. These can be further used to approximate $\text{E}_{\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\bar{\boldsymbol{\theta}}))$.

The next definition of DIC under the class of complete data is motivated by the fact that the at times $\text{E}(\boldsymbol{b}, \boldsymbol{S}|\boldsymbol{y})$ takes values outside the support of the joint distribution of $\boldsymbol{b}, \boldsymbol{S}$ (Celeux et al., 2006). Thus the following definition of DIC is proposed using the posterior mode as the estimate instead.

$$\text{DIC}_5 = -4\text{E}_{\boldsymbol{\theta},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}^F|\boldsymbol{\theta})) + 2\log p(\boldsymbol{y}, \hat{\boldsymbol{b}}_M, \hat{\boldsymbol{S}}_M|\hat{\boldsymbol{\theta}}_M) \tag{4.9}$$

Conditional DIC

The third class of the DIC is based on the assumption that $\boldsymbol{b}, \boldsymbol{S}$ can be seen as an additional parameter rather than as missing data. We will represent the new posterior parameter space as $\boldsymbol{\theta}^{\text{cond}} = (\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{S})$. The following equation shows the conditional data likelihood,

$$p(\boldsymbol{y}|\boldsymbol{\theta}^{\text{cond}}) = \prod_{i=1}^{n} f_N(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, R_i) \tag{4.10}$$

Based on this conditional likelihood, Celeux et al., (2006) proposed the following DIC definition.

$$\text{DIC}_6 = -4\text{E}_{\boldsymbol{\theta}^{\text{cond}}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}^{\text{cond}})) + 2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_M^{\text{cond}})) \tag{4.11}$$

where $\hat{\boldsymbol{\theta}}_M^{\text{cond}} = \arg\max_{\boldsymbol{\theta}^{\text{cond}}} p(\boldsymbol{\theta}^{\text{cond}}|\boldsymbol{y})$, and $\text{E}_{\boldsymbol{\theta}^{\text{cond}}}(\log p(\boldsymbol{y}|\boldsymbol{\theta}))$ can be approximated as done in equation 4.5.

4.2   Marginal Likelihood

The marginal likelihood of the data represents the probability of data given the model. This can be calculated by marginalizing the likelihood over the model parameters $\boldsymbol{\theta}$.

$$p(\boldsymbol{y}|M) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)\,\mathrm{d}\boldsymbol{\theta} \tag{4.12}$$

Given two competing models for the data, M_1 and M_2, one can further use expression 4.12 to calculate the odds of model M_1 against the model M_2 given the data. These odds are called the Posterior odds and can be written as,

$$\frac{p(M_1|\boldsymbol{y})}{p(M_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|M_1)p(M_1)}{p(\boldsymbol{y}|M_2)p(M_2)}$$

where $\frac{p(M_1)}{p(M_2)}$ is called prior odds, and $\frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}$ is called the Bayes Factor. Since we have the same prior belief in each of these models, prior odds is equal to 1. Marginal likelihood is a relative measure of model fit and unless prior odds are not equal to 1, the marginal likelihood will be equal to Bayes Factor. To calculate the Bayes Factor we will use the method proposed by Chib, (1995). Chib's idea is that one can rewrite the Bayes rule in equation 2.1 to get the marginal likelihood formula as

$$m(\boldsymbol{y}) = p(\boldsymbol{y}|M) = \frac{L(\boldsymbol{\theta}|\boldsymbol{y}, M)p(\boldsymbol{\theta}|M)}{p(\boldsymbol{\theta}|\boldsymbol{y}, M)} \tag{4.13}$$

Equation 4.13 is valid for all $\boldsymbol{\theta}$, though Chib recommends using posterior mode $\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{y})$ of parameters or the maximum likelihood estimate $\arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{y})$. We decided to choose the latter of the two. Further, in context of the Bayesian heterogeneity model, we will denote the selected parameter value as $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \sigma^{2*}, \boldsymbol{\nu}^*)$. Thus Chib's approximation for marginal likelihood on log scale is given by,

$$\log \hat{m}(\boldsymbol{y}) = \log L(\boldsymbol{\theta}^*|\boldsymbol{y}) + \log p(\boldsymbol{\theta}^*) - \log p(\boldsymbol{\theta}^*|\boldsymbol{y}) \tag{4.14}$$

Note that we have dropped the model indicator M from equation 4.14 for readability. We will now show calculations for determining the marginal likelihood value using Chib's approximation.

Firstly $\log L(\boldsymbol{\theta}^*|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}^*)$ can be easily determined using the formula given in equation 4.1. The calculation of $\log p(\boldsymbol{\theta}^*)$ is also straightforward because we take independent priors for the parameters in our model, the details of which are given in section 3.4.4. Assuming that the

parameters of component densities of the mixture distribution of random effects are independent, one can use the following to calculate $\log p(\boldsymbol{\theta}^*|\boldsymbol{y})$.

$$\log p(\boldsymbol{\theta}^*|\boldsymbol{y}) = \sum_{k=1}^{K} \log p(G_k^*|\boldsymbol{y}) + \sum_{k=1}^{K} \log p(\boldsymbol{b}_k^{C*}|G_k^*, \boldsymbol{y}) + \log p(\sigma^{2*}|G_k^*, \boldsymbol{b}_k^{C*}, \boldsymbol{y})$$
$$+ \log p(\boldsymbol{\beta}^*|G_k^*, \boldsymbol{b}_k^{C*}, \sigma^{2*}, \boldsymbol{y}) + \log p(\boldsymbol{\eta}^*|G_k^*, \boldsymbol{b}_k^{C*}, \sigma^{2*}, \boldsymbol{\beta}^*, \boldsymbol{y}) \quad (4.15)$$

An interesting problem one faces in such an expansion is that the posteriors may not be available as a well known density. For e.g. in the case of covariance matrix G_k, an inverse Wishart prior for the covariance matrix G_k is a conjugate prior, conditional on the true mean $\boldsymbol{b}_k^C$ being known. To solve this problem, Chib, (1995) suggested using the Rao-Blackwellization method. For e.g. the Rao-Blackwellized estimate of $p(G_k^*|\boldsymbol{y})$ is given by,

$$\prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}) = \int \prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}, \boldsymbol{b}_k^C) p(\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, ..., \boldsymbol{b}_K^C, \boldsymbol{b}, \boldsymbol{S}|\boldsymbol{y}) \, \mathrm{d}\boldsymbol{b}_1^C \, \mathrm{d}\boldsymbol{b}_2^C ... \, \mathrm{d}\boldsymbol{b}_K^C \, \mathrm{d}\boldsymbol{b} \, \mathrm{d}\boldsymbol{S}$$

$$\approx \frac{1}{m} \sum_{l=1}^{m} \prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}, \boldsymbol{b}^{(l)}, \boldsymbol{S}^{(l)}, \boldsymbol{b}_k^{C(l)}) \quad (4.16)$$

$$\approx \frac{1}{m} \sum_{l=1}^{m} \prod_{k=1}^{K} f_{\mathcal{W}^{-1}}(G_k^*; n_k^{(l)} + n_0, \Psi + \sum_{i=1}^{n_k^{(l)}} (\boldsymbol{b}_i^{(l)} - \boldsymbol{b}_k^{C(l)})(\boldsymbol{b}_i^{(l)} - \boldsymbol{b}_k^{C(l)})^T)$$

where, $n_k^{(l)}$ are number of subjects classified under component k in iteration l, $f_{\mathcal{W}^{-1}}$ represents the density function of an inverse Wishart distribution and (n_0, Ψ) are the parameters for the inverse Wishart distribution specified as prior for the variance covariance matrix of the component densities. In 4.16 one approximates the integral with the samples obtained from the MCMC iterations. As we can see the benefit of this approach is that $p(G_k^*|\boldsymbol{y}, \boldsymbol{b}^{(l)}, \boldsymbol{S}^{(l)}, \boldsymbol{b}_k^{C(l)})$ is the well known inverse Wishart density. If instead of an inverse Wishart prior, independent inverse gamma priors for variance parameters of random effects and a uniform prior for correlation was chosen, then an interesting problem arises. The posterior distribution of correlation conditional on knowing the variance parameters, mean and complete data, is not available as a well known density in the Rao-Blackwellization procedure. One way to solve this problem is to approximate the posterior density with a well known density. However this is required to be done as many times as the number of MCMC iterations, making it practically infeasible. We thus avoided the calculation of Bayes factor in the case where we took independent inverse gamma priors for variance of random effects and uniform prior for correlation.

Proceeding further with the Rao-Blackwellization procedure we obtained the following approximations for the remaining parameters.

$$\prod_{k=1}^{K} p(\boldsymbol{b}_k^{C*}|G_k^*, \boldsymbol{y}) \approx \frac{1}{m} \sum_{l=1}^{m} \prod_{k=1}^{K} f_N(\boldsymbol{b}_k^{C*}; (G_0^{-1} + n_k^{(l)} G_k^{*-1})^{-1}(G_0^{-1}\boldsymbol{\mu}_0 + n_k^{(l)} G_k^{*-1}\overline{\boldsymbol{b}}_{ik}^{(l)}), (G_0^{-1} + n_k G_k^{*-1})^{-1})$$

$$(4.17)$$

$$p(\sigma^{2*}|G_k^*, \boldsymbol{b}_k^{C*}, \boldsymbol{y}) \approx \frac{1}{m} \sum_{l=1}^{m} f_{Inv-Gamma}(\sigma^{2*}; \alpha_0 + \frac{\sum_{i=1}^{n} m_i}{2}, \beta_0 + \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}^{(l)} - \boldsymbol{z}_{ij}\boldsymbol{b}_i^{(l)})}{2})$$

$$(4.18)$$

$$p(\boldsymbol{\beta}^*|G_k^*, \boldsymbol{b}_k^{C*}, \sigma^{2*}, \boldsymbol{y}) \approx \frac{1}{m}\sum_{l=1}^{m} f_N(\boldsymbol{\beta}^*; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{H}^{(l)}), \sigma^{2*}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \qquad (4.19)$$

$$p(\boldsymbol{\eta}^*|G_k^*, \boldsymbol{b}_k^{C*}, \sigma^{2*}, \boldsymbol{\beta}^*, \boldsymbol{y}) \approx \frac{1}{m}\sum_{l=1}^{m} f_{Dir}(\boldsymbol{\eta}^*; a_{01} + n_1^{(l)}, a_{02} + n_2^{(l)}, ..., a_{0K} + n_K^{(l)}) \qquad (4.20)$$

where, $(\boldsymbol{\mu}_0, G_0)$ are the parameters for the multivariate normal prior for the mean $\boldsymbol{b}_k^C$ of the kth component density,
$\overline{\boldsymbol{b}}_{ik}^{(l)} = \frac{\sum_{i=1}^n I(S_i^{(l)}=k)\boldsymbol{b}_i^{(l)}}{n_k^{(l)}}$ is the mean of the estimated random effects corresponding to the n_k subjects classified under the kth component in the lth MCMC iteration,
(α_0, β_0) are the parameters of the inverse gamma density specified as the prior for the within subject variance σ^2,
$a_{01}, a_{02}, ..., a_{0K}$ are the parameters of the Dirichlet density specified as the prior for component weight vector $\boldsymbol{\eta}$,
$f_{Inv-Gamma}$ and f_{Dir} are the density functions of Inverse gamma and Dirichlet distributions respectively.

These values can be substituted in equation 4.15 to obtain an estimate of $\log p(\boldsymbol{\theta}^*|\boldsymbol{y})$, which can be further substituted in equation 4.14 to obtain $\log \hat{m}(\boldsymbol{y})$. In models where marginal likelihood is known to work well as a model selection criteria, models with higher value of $\log \hat{m}(\boldsymbol{y})$ should be preferred.

4.3   Posterior predictive checks

The motivation behind a posterior predictive check is to evaluate the model fit using simulations from the posterior predictive distribution(PPD) $p(\tilde{\boldsymbol{y}}|\boldsymbol{y})$. For a simple model such as $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, one can perform an informal check by sampling 1000(say) values from the PPD 20(say) times and then comparing the histogram of the original sample with the histogram of the data sampled from PPD. If the histograms do not match then one can say that the model does not fit the data well.

In context of mixture distributions Frühwirth-Schnatter, (2013) suggest that one should first sample allocation vector $\tilde{\boldsymbol{S}}$ based on the posterior density of the weight distribution $p(\boldsymbol{\eta}|\boldsymbol{y})$ and then generate samples $\tilde{\boldsymbol{y}}$ from $p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{S}}, \boldsymbol{\theta})$. A similar approach can be followed in hierarchical models by first generating new random effects $\tilde{\boldsymbol{b}}$. This was termed as the mixed predictive approach by Marshall and Spiegelhalter, (2003). As an alternative of mixed predictive approach, one can use the posterior allocations $\boldsymbol{b}|\boldsymbol{y}$ instead of $\tilde{\boldsymbol{b}}$. However the problem with this approach is that it leads to a more conservative posterior predictive check (Congdon, 2010).

4.3.1   PPC for the Bayesian heterogeneity model

Detecting overfitting the number of components

When more than the required number of components are fitted to the mixture distribution of random effects then during the MCMC procedure some of the components may remain empty. i.e. no observation gets allocated to some of the components. In such a case, the posterior of variance covariance matrix G_k as well as of mean $\boldsymbol{b}_k^C$ is sampled from the prior distribution. Since the priors we considered were non informative, the posterior samples of parameters are arbitrarily large or small values. Thus the random effects $\tilde{\boldsymbol{b}}$ sampled from the posterior distribution

of the parameters of empty components, and the posterior predictive sample from $p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{b}})$ will also be arbitrarily small or large. We used this property to form a posterior predictive check. However this approach is only possible with mixed predictive checking. The reason is that in our simulation study we found that the posterior samples from $p(\boldsymbol{b}_i|\boldsymbol{y})$ do not show any anomaly even if the mixture distribution of random effects is overfitted/underfitted. Thus following the classical PPC one may not be able to detect over/under fitting based on the posterior samples from $p(\boldsymbol{b}_i|\boldsymbol{y})$.

We will now discuss two of the issues with the above approach and their solutions. Firstly, the weights η_k of empty components are sampled from the Dirichlet posterior which gets information from all other components as well. Thus iterations in which the components remain empty, the weights also remain small for the corresponding components. This problem can be obviated by sampling a large number of observations from the posterior predictive distribution. The second problem is that if one uses an inverse Wishart prior for the variance covariance matrix of component densities then although it is non-informative for the correlation, it imposes certain restrictions on the variances. One can instead choose independent inverse gamma priors for variances and uniform prior for correlation in such a scenario, albeit under the restriction that the sample variance covariance matrix is invertible.

The first step in the mixed predictive check approach for Bayesian heterogeneity model is to sample the new allocations $\tilde{S}$. The second step is to sample the random effects from $p(\tilde{\boldsymbol{b}}_i|\tilde{S}_i)$ for the subjects. We used the following test statistic based on these sampled random effects.

$$T(\boldsymbol{r}) = \frac{1}{\sum_{i=1}^{n} m_i} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left(r_{ij} - \bar{r}_{i.} \right)^2 \tag{4.21}$$

where, $r_{ij} = \boldsymbol{z}_{ij}\tilde{\boldsymbol{b}}_i + \varepsilon_{ij}$ for the mixed predictive distribution of the random effects and $\bar{r}_{i.} = \frac{1}{m_i} \sum_{j=1}^{m_i} r_{ij}$. $\boldsymbol{z}_{ij}$ is as defined in section 3.1.1. For the observed data set, r_{ij} can be calculated as $r_{ij} = y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is sampled from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$. When the number of components in the mixture distribution of random effects are more than required, then during some of the iterations the test statistic is expected to have arbitrarily large values.

Underfitting the number of components

We found in our simulation study that it was not possible to distinguish the distribution of the test statistic $T(\boldsymbol{r})$ with underfitted mixture of random effects from the distribution of the test statistic with rightly fitted mixture of random effects. The reason for it is that the full Bayesian estimates for random effects $\boldsymbol{b}_i$ can be almost the same for underfitted as well as rightly and overfitted models. This further results into an almost equal estimate of within subject variance σ^2 for the various models. The mean structure parameter estimates are also almost the same and thus eventually with these minor differences it is impossible to differentiate between the models using PPC. One way to circumvent this issue is to fit as many components as possible till overfitting is detected, and then the choice among the underfitted models could be done on the basis of interpretation of the clusters and the size of the clusters.

4.3.2 Posterior predictive p-values

The motivation of Posterior predictive p-values(PPP) is similar to frequentist p-values. The PPP are however averaged over the entire posterior distribution of the test statistic. Given a test statistic $T(\boldsymbol{y})$, the frequentist p-value is equal to the probability $P(T(\tilde{\boldsymbol{y}}) > T(\boldsymbol{y}))$, where $p(\tilde{\boldsymbol{y}}) = p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$. In the Bayesian paradigm the parameter $\boldsymbol{\theta}$ has a posterior distribution. Thus, although the same probability for the test statistic is calculated as before, but it is averaged over the entire posterior distribution $p(\theta|\boldsymbol{y})$. A small PPP value indicates bad model fit. In context of

the Bayesian heterogeneity model PPP values may not be as desirable for testing over/under fitted for the reasons mentioned in the previous subsection. However they can still be used to detect a poor fit to the data itself.

Chapter 5

Simulation study

In this chapter we share the results from the simulation study we performed to check the efficacy of the model selection criteria described in Chapter 4. We implemented the Bayesian heterogeneity model using the R package R2jags (Su and Yajima, 2015) and analyzed the MCMC chains using the R package ggmcmc (Marín, 2016). For the calculation of marginal likelihood we required the density function of Wishart distribution, which was available in two packages, namely MCMCpack and mixAK. There were inconsistencies in the results from the two implementations and we eventually used mixAK (Komárek, 2015) as the MCMCpack package produced density function value to be ∞ in some cases.

5.1 Data sets for simulation study

The data sets we simulated were motivated by the study on predicting Zebu cow's weights in sub Saharan Africa (Lesosky et al., 2012). We assumed our response to be the weight of the Zebu cows. The predictors we considered were hypothetical, namely gender of a cattle (Male/Female), birth year of the cattle (1996/1997), age of the cattle at the first measurement and the time at which measurement was taken. The measurements of the cows were done at 10 different equally spaced time intervals. We further added subject specific random intercept and random slope effect to each response so that the repeated measurements for a given cow were correlated. Simultaneously we made sure that these cow specific random effects were mixture distributed. We will refer to the cows as subjects here forth.

5.1.1 Description of each data set

Our aim was to create data sets differing in number of mixture components for random effects, number of subjects, statistical power to detect the fixed effects, separation of mixture components and number of subjects per component. To analyze the efficacy of model selection criteria under these different scenarios we created multiple data sets. To get a rough idea about the random effects in each of these data sets, we first did a graphical analysis. For this purpose we first regressed the response y on the 3 predictors: age, gender and birth year of cattle using OLS(section 3.4.3). It is also possible to use linear mixed model for estimating the fixed effects. We then regressed the residuals $y_{ij} - x_{ij}\beta$ on the intercept and time of measurement for every subject separately to obtain a rough estimate $\tilde{b}_i$ of the random effect of subjects. This estimator however overestimates the actual size of the random effects because the within subject variance is also included in it. It is not possible to use Empirical Bayes estimates of the random effects because they may fail to reflect the heterogeneity in the random effects population (Verbeke and Lesaffre, 1996). Lastly, it is important to note that fitting an incorrect mean structure can lead to a incorrect representation of the random effects distribution as shown in figure 5.1b.

Data set 1: No mixture distribution of random effects

The first data set we created was without a mixture of random effects. i.e. $b_i \sim N(0, G)$. In total we generated data of 80 subjects, each having 10 repetitions. Based on the approach mentioned above, a plot of the random effect values for this data set is shown in figure 5.1a.

Data set 2: 3 well separated components for the mixture of random effects

The next data set we created had 3 well separated components forming the mixture distribution of random effects. In total we generated data of 180 subjects, each having 10 repetitions. A plot of the rough estimates of random effect values for this data set is shown in figure 5.2a.

(a) Data set 1   (b) Data set 2: Missing covariate age

Figure 5.1: Rough estimate $\tilde{b}_i$ for random effects

Data set 3: 3 well separated components but less subjects

This data is similar to Data set 2 in all regards except for the number of subjects. We generated only 36 subjects in total in this data set. A plot of the rough estimates of random effect values for this data set is shown in figure 5.2b.

Data set 4: 3 fused components for the mixture of random effects

In this data set we simulated the random effects from a mixture distribution which had 3 fused components. For e.g. if one sees the plot of the rough estimates of random effect values for this data set (figure 5.3a) then it is not clear if there are more than 2 components.

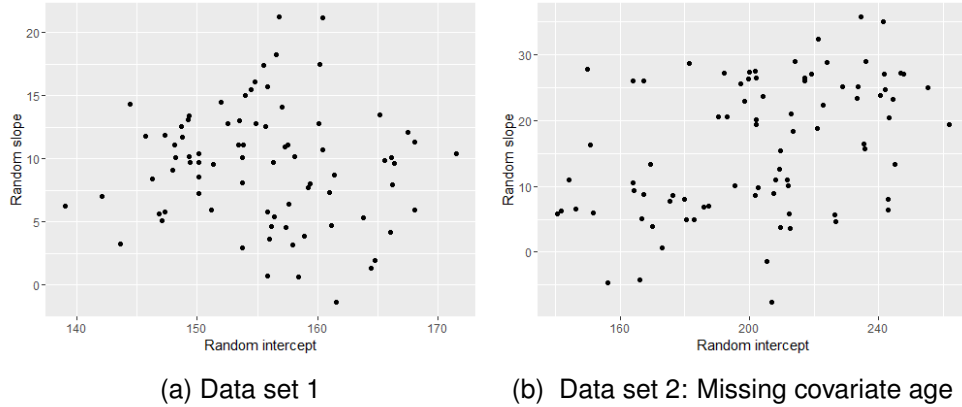(a) Data set 2   (b) Data set 3
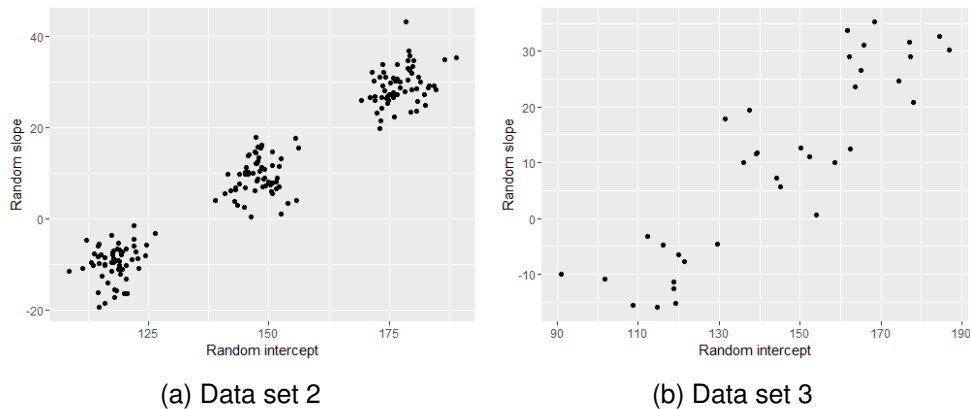
Figure 5.2: Rough estimate $\tilde{b}_i$ for random effects

24

Data set 5: 3 fused components but less subjects

This data is similar to Data set 4 in all regards except for the number of subjects. We generated only 36 subjects in total in this data set. A plot of the rough estimates of random effect values for this data set is shown in figure 5.3b.

Data set 6: 5 well separated components

In this data set we simulated the random effects from a mixture distribution which had 5 well separated components. However this time we generated unequal number of subjects for every component. The plot of the rough estimates of random effect values for this data set is shown in figure 5.4a.

(a) Data set 4   (b) Data set 5
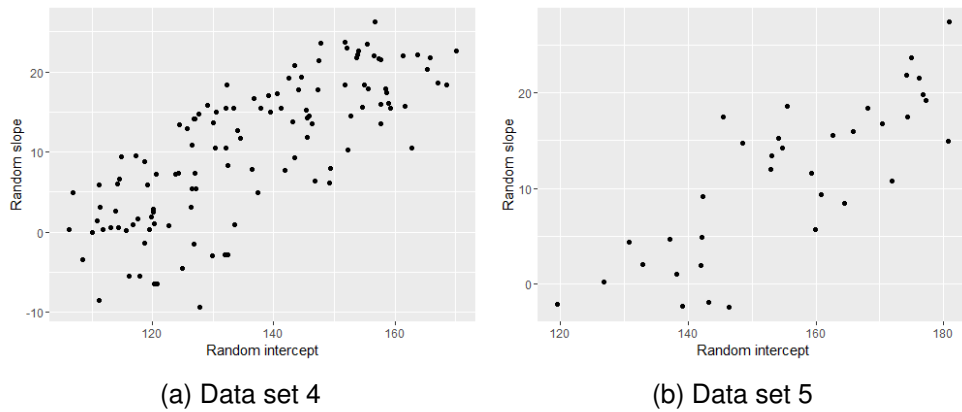
Figure 5.3: Rough estimate $\tilde{b}_i$ for random effects

Data set 7: 5 fused components

This data is similar to Data set 6 in all regards except that the number of subjects per component are less, and the components are not so well separated. The plot of the rough estimates of random effect values for this data set is shown in figure 5.4b.

(a) Data set 6   (b) Data set 7
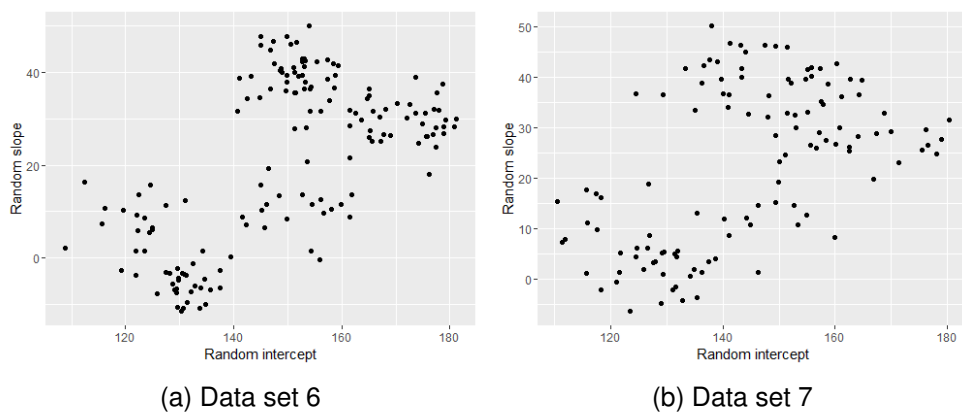
Figure 5.4: Rough estimate $\tilde{b}_i$ for random effects.

5.1.2   Running MCMC simulations

We first discuss some of the issues we encountered while doing the MCMC simulations. The first one was label switching across the chains while calculating Bayes factor using Chib's ap-

proximation; i.e. label 1 corresponding to component 1 in one chain and corresponding to some other component in another chain. This gave inconsistent and incorrect estimates for the various calculations we did. Although we dealt with it using the mechanisms given in section 3.4.5, mechanisms such as applying an identifiability constraint decreased the speed of simulations drastically. The second issue was high autocorrelation in the chains. We tried various strategies for it. For e.g. using a marginal model was helpful and so was hierarchical centering.

However despite these measures we had to employ a thinning of 1 per 100 iterations and in some cases 1 per 200 iterations to make sure the resulting chains were not autocorrelated. Since we lacked computational resources required to run longer chains we had to be content with chains of length 1300 (after thinning). Lastly, we observed that in models where mixture of random effects were fitted with more components than needed, there was very high autocorrelation in the chain which could not be reduced despite high thinning. The convergence tests for such chains showed that the chains did not converge for the parameters of some of the component densities. The parameters corresponding to fixed effects converged in such chains though. Since, for some of the component densities the parameters were not converged, it made little sense to compare DIC, Marginal likelihood and PPC based on such models with models for which the chains converged. Despite that when we tried to check the aforementioned model selection criteria for such models and observed some interesting patterns which are discussed in the next section.

5.1.3 Deviance information criteria

The Deviance information criteria results are based on a single chain and have been rounded to the nearest integer. Table 5.1 shows the values of the various deviance information criteria (section 4.1) applied to data set 1. Since the data set 1 had no mixture of random effects, models with 2 or more components are overfitting the data. As we mentioned earlier that the chains did not converge for some of the parameters in the models with more components than needed, the impacts of checking DIC for such chains are visible in the DIC and p_D values. For e.g. DIC_1 is equal to 9 for 4 components, which is misleading and unrealistic. We also obtained a negative value (-602) for p_{D_1} when we fitted 3 components. Celeux et al., (2006) noted that this may happen when posterior mean borrows from several modal regions of the posterior density and ends up with a value that is located between modes. As a remedial measure they suggested using DIC_3 instead of the other two observed data DIC measures.

Table 5.1: DIC and p_D for data set 1. True number of components = 1

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	4143	4144	4143	5160	4402	3483
2	4146	4147	4145	5161	4403	3480
3	3536	4148	4147	5163	4388	3465
4	9	4151	4149	5166	4407	3485

# Comp Fitted	p_{D_1}	p_{D_2}	p_{D_3}	p_{D_4}	p_{D_5}	p_{D_6}
1	9	9	9	903	145	125
2	9	10	9	903	145	122
3	-602	9	8	901	126	107
4	9	11	9	903	145	127

Table 5.2 shows the values of various DIC we obtained for data set 2. One of the patterns we

observed in these results is that DIC_4 decreased continuously till the right number of components were fitted, whereas for the overfitted models it either decreased by a bit or remained more or less the same. A similar pattern was observed for DIC_5 and DIC_3. However they do not seem to be as discerning as DIC_4. DIC_1 and DIC_2 have the lowest DIC for the model with 3 components, among all the converged models. DIC_6 prefers model with no mixture over models with 2 or 3 components, and hence it is not reliable. It is important to note that while we did observe some patterns for other DIC's, we only consider DIC calculations valid for up to models with 3 components as the rest did not converge.

Table 5.2: DIC and p_D for data set 2. True number of components = 3

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	9966	9959	9965	12921	10531	7855
2	9865	9849	9864	12498	10458	7860
3	9664	9665	9663	11847	10244	7870
4	9516	9654	9664	11834	10266	7888
5	7370	9729	9666	11812	10277	7870
6	9498	9661	9668	11833	10242	7857

# Comp Fitted	p_{D_1}	p_{D_2}	p_{D_3}	p_{D_4}	p_{D_5}	p_{D_6}
1	9	2	8	2670	279	269
2	15	-1	14	2344	304	272
3	21	21	20	1933	331	282
4	-125	13	23	1913	345	298
5	-2270	89	26	1889	355	280
6	-147	16	23	1912	321	269

Table 5.3 shows the values of the various DIC applied to the data set 3. Firstly we can see that DIC_6 prefers model with 2 components over a model with 3 components, which is an incorrect choice of the number of the components in mixture. DIC_5 exhibits the same problem and hence is not reliable. DIC_1 and DIC_2 both had lowest DIC for the model with 3 components among the converged models however unlike DIC_3 and DIC_4 they did not exhibit the pattern of DIC remaining almost the same for overfitted models. The presence of this pattern however, is not guaranteed for all data sets (section 6.3) but if observed it is still helpful to detect the boundary point beyond which all models were overfitted. The magnitudes of all of the the DIC for this data set was less in comparison to DIC for data set 2 because the sample size for this data set was only 36 subjects compared to 180 subjects in the former.

Table 5.4 shows the results of applying various DIC to data set 4. So far we had observed that DIC_1 to DIC_4 could be used to detect the right number of components. Since the components in this data set were fused, with the large number of subjects the following results were interesting to analyze. From the table of DIC it seems DIC_1 to DIC_4 could still be used to select the right number of components among the converged models. However only DIC_4 followed the pattern of DIC becoming stable for overfitted models. To further validate the results, we decided to decrease the number of subjects to 36.

From the table 5.5, which shows the results of DIC for data set 5, it seems that the pattern we saw so far for DIC_4 did not exist anymore. However the peculiarity of this data set was that the number of subjects were less and the components were fused. Thus we observed that some of the components remained empty even when we fitted the right number of components and

Table 5.3: DIC and p_D for data set 3. True number of components = 3

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	2013	2012	2012	2611	2118	1570
2	1989	1949	1987	2497	2013	1562
3	1942	1942	1940	2339	2039	1571
4	1943	1944	1942	2342	2034	1559
5	1936	1940	1944	2344	2049	1580
6	1695	1948	1945	2344	2053	1579

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1	8	7	7	545	52	45
2	14	-26	12	465	-20	35
3	17	17	15	370	70	46
4	16	17	15	370	62	34
5	8	11	15	370	75	56
6	-235	17	15	368	77	53

Table 5.4: DIC and p_D for data set 4. True number of components = 3

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	6568	6566	6566	8454	6899	5197
2	6531	6523	6530	8263	6946	5253
3	6497	6492	6497	8017	6898	5263
4	6347	6480	6488	7955	6898	5253
5	6321	6463	6485	7932	6743	5259
6	6382	6329	6489	7948	6611	5259

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1	9	8	8	1694	139	127
2	14	7	13	1527	210	182
3	20	14	19	1341	222	187
4	-115	17	26	1287	230	181
5	-138	4	26	1265	76	187
6	-81	-134	26	1275	-62	185

chains did not converge. As we discussed in section 3.4.4 one could use a Dirichlet prior with slightly bigger hyperparameters to avoid empty components, although at the risk of becoming too informative. We changed the prior for weight distribution from $Dir(1, 1, ..., 1)$ to $Dir(3, 3, ..., 3)$ and fitted models with 2, 3 and 4 components for the mixture. With 2 components we found $DIC_4 = 2441(p_{D4} = 461)$ and $DIC_3 = 1937(p_{D3} = 14)$. For 3 components we obtained DIC_4 = 2352 whereas for 4 components we obtained $DIC_4 = 2346$ and $DIC_3 = 1925$. In light of these results one could still justify the pattern we observed for DIC_4 so far. However it seems that the pattern was not equally clear for DIC_3. An interesting result from this exercise was that the choice of $Dir(1, 1, ..., 1)$ prior is prone to underfitting when the components are fused and observations are less. Lastly, we also modeled the data using smaller values for Dirichlet prior hyperparameters; for e.g. $Dir(0.1, 0.1, ..., 0.1)$ prior. However it performed as worse as the $Dir(1, 1, ..., 1)$ prior for the current data set.

Table 5.6 shows the results of applying DIC to the various models fitted for data set 6. The

Table 5.5: DIC and p_D for data set 5. True number of components = 3

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	1944	1943	1943	2500	1879	1364
2	1936	1941	1945	2487	1919	1408
3	1886	-3353	1944	2453	$-\infty$	1525
4	1892	1904	1944	2439	1851	1389
5	1902	1840	1942	2418	704	336
6	1883	1919	1933	2371	2023	1538

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1	9	7	7	510	-110	-119
2	2	7	11	500	-68	-73
3	-42	-5281	16	470	$-\infty$	41
4	-34	-22	18	459	-130	-93
5	-21	-83	19	442	-1272	-1146
6	-31	5	19	407	59	55

pattern of DIC_4 remaining stable for overfitted models is visible here as well. Interestingly it also exhibited by DIC_3, which can be attributed to the fact that the components are well separated. DIC_5 and DIC_6 select underfitted models whereas DIC_1 and DIC_2 still select the right model among the models which have converged chains.

Table 5.6: DIC and p_D for data set 6. True number of components = 5

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	8982	8981	8980	11847	9251	6655
2	8829	8827	8827	11327	9293	6838
3	8745	8742	8744	11036	9251	6895
4	8669	8672	8677	10737	9208	6925
5	8649	8643	8648	10601	9165	6909
6	8096	8697	8650	10594	9183	6923
7	7770	8364	8651	10593	7613	6919
8	8196	8640	8653	10597	9143	6927

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1	9	9	7	2591	-5	-14
2	14	13	12	2224	190	169
3	20	16	19	2035	250	223
4	19	23	27	1824	296	251
5	31	26	30	1725	289	232
6	-520	81	33	1711	300	246
7	-848	-254	34	1706	-1274	244
8	-424	19	33	1711	257	248

Table 5.7 shows the results of applying DIC to the various models fitted for data set 7. Based on figure 5.4b it is difficult to identify more than 3 components in this mixture. However once again DIC_4 appeared to be identifying the right number of components correctly. DIC_3 did not seem to work well in this case, which as we saw earlier happens when the components are

Table 5.7: DIC and p_D for data set 7. True number of components = 5

# Comp Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1	6708	6707	6706	8819	6977	5071
2	6606	6605	6604	8443	6946	5135
3	6539	6538	6537	8178	6944	5204
4	6506	6514	6521	8078	6915	5196
5	6505	6500	6508	7984	6896	5202
6	6465	6501	6510	7988	6895	5196
7	6200	6500	6512	7989	6883	5190
8	6448	6498	6516	7995	6901	5196

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1	9	8	7	1903	61	53
2	15	14	13	1636	139	120
3	21	20	19	1456	221	185
4	12	20	26	1381	218	176
5	26	22	29	1308	220	182
6	-14	21	30	1307	214	176
7	-282	18	30	1305	198	168
8	-36	14	32	1307	213	175

fused. Interestingly DIC_1 selects the right model only by virtue of a difference of DIC value of 1 between model with 4 components and 5 components.

It is important to note that when components were fused, then even with a very large number of MCMC iterations we observed a variation of 10 to 20 in the same DIC across chains. Thus, although DIC_3 preferred model 5 over model 4 by a DIC difference of 13, it is highly plausible to obtain a difference of -7; i.e. Model 4 having a lower DIC_3 than model 5. In such a case if the classical rule of thumb of DIC difference of 5 to 10 is used to select the models then an incorrect model might get selected.

5.1.4 Marginal likelihood

We implemented Chib's approximation described in section 4.2. Table 5.8 shows the results of $\log \hat{m}(\boldsymbol{y})$ for the various models we fitted to data set 1 to data set 7. One can see that even among the well converged models there is no obvious pattern visible in these results to conclude the efficacy of marginal likelihood in selection of a model. For e.g. in case of data set 6 there were 5 well separated components in the mixture distribution of random effects. When we fitted 1 to 5 components for this data set we obtained MCMC chains with good convergence. However in table 5.8 we can see that marginal likelihood prefers fitting 1 component over higher number of components. For data set 3 where we had 3 well separated components but only 36 subjects in total, we can observe that marginal likelihood incorrectly prefers model with 2 components and not model with 3 components. Thus Marginal likelihood does not seem to be an effective method to choose the right Bayesian heterogeneity model.

5.1.5 Posterior predictive check (PPC)

We implemented the PPC outlined in section 4.3. Since the PPC we used were designed to detect non identifiability due to empty components, it worked irrespective of the size of the data

Table 5.8: $\log \hat{m}(\boldsymbol{y})$ for data set 1

Fitted	1 Comp	2 Comp	3 Comp	4 Comp	5 Comp	6 Comp	7 Comp	8 Comp
Data set 1	-2120	-2128	-2139	-2142				
Data set 2	-5019	-4989	-4937	-4925	-4938	∞		
Data set 3	-1038	-1044	-1042	-645	-1003	∞		
Data set 4	-3317	-3318	-3322	-3332	-3348	∞		
Data set 5	-1001	-1016	-1032	-1041	-1058	∞		
Data set 6	-4545	-4492	-4477	-4467	-4473	∞	-4498	-3985
Data set 7	-3397	-3379	-3373	-3380	-2749	∞	∞	-3416

set or how well separated the components were. It also worked well when the chains were not fully converged. Thus we found the results of PPC to be consistent across the various data sets. Because of this reason, we will only discuss the results of fitting various number of components in data set 6. It is important to note that, like DIC we also applied PPC to models which overfitted the mixture of random effects and also did not converge. Figure 5.5a to 5.5h show the distribution of the test statistic 4.21 for the various models we fitted to data set 6. It can be seen that the distribution is positively skewed whenever overfitting is present. On the other hand the distribution of the test statistic is more or less the same in cases of underfitting.

Table 5.9: PPP values for the various models fitted to data set 6.

# Components fitted	1	2	3	4	5	6	7	8
PPP Value	0.500	0.500	0.492	0.500	0.500	0.581	0.674	0.775

We also calculated posterior predictive p-values for the various number of components we fitted to data set and they are presented in table 5.9. As one can see the PPP values are equal (0.5) for all the underfitted models. Based on the graphical PPC and these PPP values, it was not possible to choose a model among the correctly fitted and underfitted one. On the other hand for models with overfitted components, the PPP values increase as overfitting increases. It is important to note that while we look for larger PPP-values to detect overfitting, it is also possible to obtain a small PPP-value indicating that the model is badly fitting in general.

We observed a more severe impact of overfitting when we used independent inverse gamma priors for the variance components of G_k and uniform prior $U(-1, 1)$ for correlation. To show the resulting distribution of the test statistic, we had to log transform it because otherwise the values were too large to be plotted in a single graph. For data set 2 we overfitted the mixture of random effects by using 4 components in the mixture. As shown in Figure 5.6a the test statistic is inflated by a large margin, which was also discussed in section 4.3. Interestingly when we fitted the right number of components the test statistic was not inflated, but the model was not fitting well to the data as shown in Figure 5.6b. The PPP-value we observed was 1. To diagnose this problem we checked the posteriors for variance covariance matrices and found them to be underestimating the sample data's variance covariance of the random effects. This indicated that the test statistic can be used to detect badly fitting models, however differentiating among models with less number of components might be difficult.
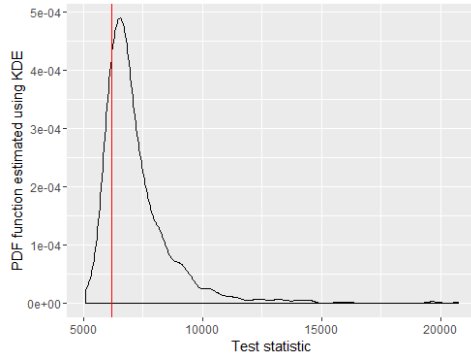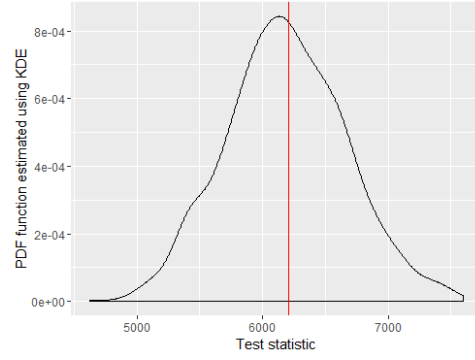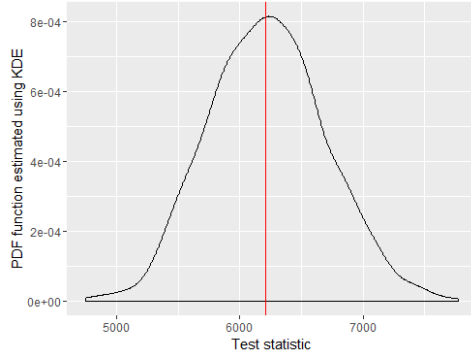
(a) # components fitted = 8
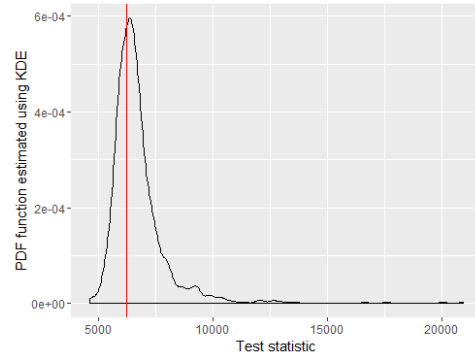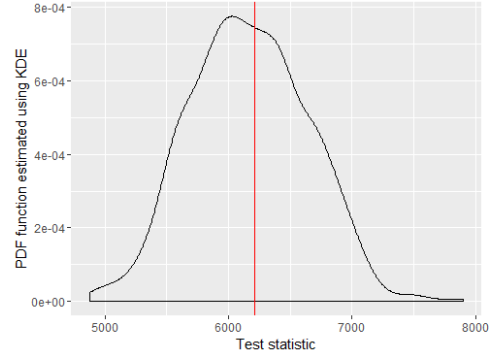
(b) # components fitted = 7
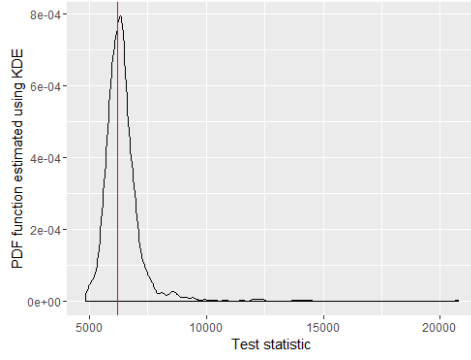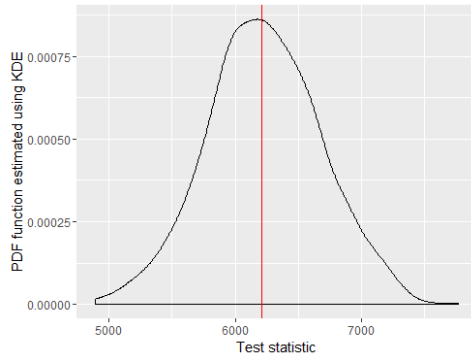
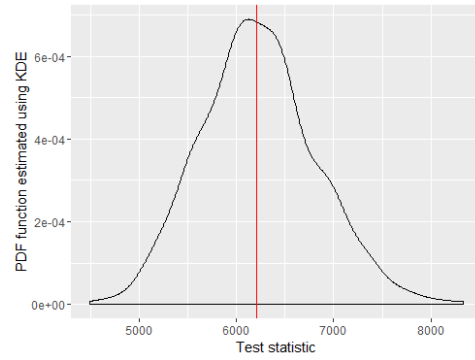(c) # components fitted = 6

(d) # components fitted = 5

(e) # components fitted = 4
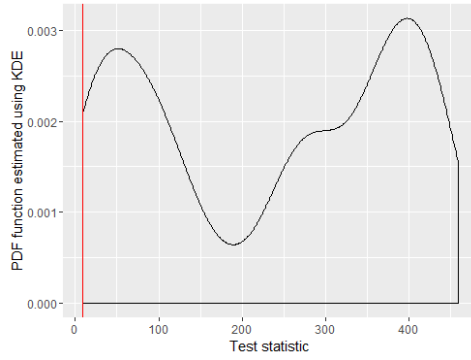
(f) # components fitted = 3

(g) # components fitted = 2

(h) # components fitted = 1

Figure 5.5: PDF function of $T(\tilde{\boldsymbol{r}})$ estimated using KDE. The red line shows the value of the test statistic $T(\boldsymbol{r})$ based on the observed data.

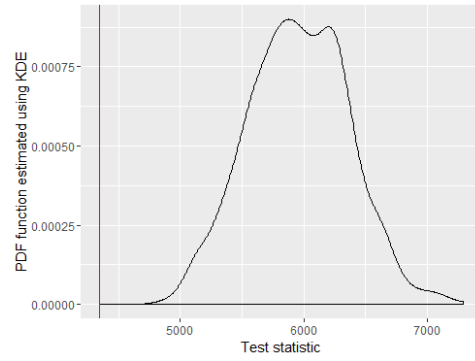(a) 4 components fitted for data set 2. log scale is used.

(b) 3 components fitted for data set 2

Figure 5.6: PDF function of $T(\tilde{\boldsymbol{r}})$ estimated using KDE. The red line shows the value of the test statistic $T(\boldsymbol{r})$ based on the observed data. Independent inverse gamma priors for variance of random effects and uniform prior for correlation is used.

Chapter 6

Analysis of blood donor data set

In this chapter we present the analysis of the blood donor data set (Nasserinejad et al., 2015) using the Bayesian heterogeneity model. The data set consists of 1595 male blood donors who donated blood multiple times over a period of many years. On each visit to the donation center, the following information was noted: Season in which blood was donated (Cold/Hot), Volume of blood donated (ml), age at the time of donation (years), a binary indicator Donate (yes/no) specifying if the donor was allowed to donate the blood, Hemoglobin (Hb) of the patient at the time of donation, and the date of visit. Other variables derived from date of visit such as time since previous donation (TSPD), time since previous visit (TSPV) were also present in the data set. Nasserinejad et al., (2013, 2015, 2016) have analyzed this data set extensively using transition models, mixed models, growth mixture models and latent class mixed effects transition models. The analysis was used to predict the Hb level of donors so that they are invited for donation at an optimal time, i.e. when their Hb levels are not too low because of the previous donations and other factors.

For the purpose of this thesis we did not use the entire data set of 1595 patients as Bayesian computation for heterogeneity model using the entire data set required large amount of computational resources. Instead we used a simple random sample of the data consisting of 250 subjects.

6.1   Motivation for analysis with Bayesian heterogeneity model

In the analysis using growth mixture models Nasserinejad et al., (2015) found 4 different underlying subpopulations in the data set. Firstly, those which had a relatively stable Hb level over donations. Secondly, those which had a higher baseline Hb level than those in category 1 and also showed a slow decline of Hb level over donations. Thirdly, those which showed a moderately sharp decline in Hb level over donations, and lastly those which showed a steep decline in Hb levels despite beginning at high initial Hb levels. Because of the presence of different subpopulations, a single methodology to decide the time of next blood donation for subjects from all subpopulations may not be effective. The aim of applying the Bayesian heterogeneity model to this dataset was to provide an alternative modeling framework for such data sets.

6.2   Frequentist analysis

We began with a frequentist analysis of the blood donor data set to select the right mean structure for our models ahead. This was necessary because omitting a required covariate in the mean structure could lead to incorrect estimates of the covariance matrix. Secondly, a

frequentist analysis provided us good starting values for the various parameters in our model. Thirdly we also wanted check if the Bayesian analysis results were consistent with the frequentist analysis results. For the random effects structure we considered a model with both random intercept and random slope. For the choice of random slope we selected the variable 'number of donations in last 2 years' because that was deemed as a suitable variable for random slope by Nasserinejad et al., (2015). For selecting the mean structure we did F-tests and likelihood ratio tests based on ML. We found the following mean structure to be suitable.

$$
\begin{aligned}
\boldsymbol{y}_{ij} = \beta_0 &+ \beta_1 * \text{Age}_{ij} + \beta_2 * \text{Season}_{ij} + \beta_3 * \text{Donate}_{ij} + \beta_4 * \text{TSPD}_{ij} \\
&+ \beta_4 * \text{\#donationLast2Years}_{ij} + \beta_5 * \text{\#donationLast2Years}_{ij} * \text{TSPD}_{ij} \\
&+ \beta_6 * \text{\#donationLast2Years}_{ij} * \text{Donate}_{ij} + \beta_7 * \text{\#donationLast2Years}_{ij}^2 \\
&+ b_0 + b_1 * \text{\#donationLast2Years}_{ij} + \varepsilon_{ij}
\end{aligned}
\tag{6.1}
$$

where, Age and TSPD have been standardized to have mean 0 and variance 1, #donationLast2Years$_{ij}$ have been downscaled by a factor of 100 so that the random slope variance scale is upscaled. Similarly the intercept we used was 0.1, thus upscaling the random intercept variance. It is important to note that this model assumes a single multivariate normal distribution for the distribution of random effects. Table 6.1 shows the parameter estimates for this model.

Table 6.1: Frequentist fixed effect estimates and random variance estimates for model 6.1. REML was used for obtaining these parameter estimates.

Fixed Effects (β)	Mean	Standard Error
intercept	94.121	0.429
Age	-0.086	0.026
#donationLast2Years	-7.265	1.555
TSPD	-0.036	0.015
Season (Hot)	-0.080	0.016
Donate (TRUE)	0.155	0.037
#donationLast2Years * TSPD	0.020	0.006
#donationLast2Years * Donate	4.689	0.995
#donationLast2Years^2	-52.796	15.237

Covariance Parameters	Estimate
$\sigma^2_{\text{Intercept}}$	19.971
$\sigma_{\text{Intercept, Slope}}$	-14.266
σ^2_{Slope}	40.183
σ^2	0.201

6.3 Bayesian analysis

We next extended the model in equation 6.1 for fitting the Bayesian version of the Heterogeneity model proposed by Verbeke and Lesaffre, (1996). Since the random component in the Bayesian heterogeneity model follows a mixture distribution we fitted mixtures of 1 to 5 components. We removed the intercept and the fixed effect of number of donations in last 2 years from the model to impose hierarchical centering on the random effects. The next step we took for model fitting was generating a plot of random part of the model using the approach mentioned in section 5.1.1. The corresponding plot in figure 6.1 indicates that there could be perhaps 1 component, or
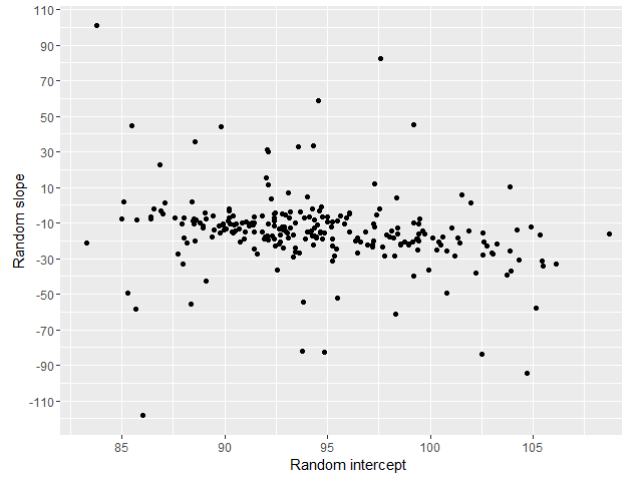
Figure 6.1: Rough estimate $\tilde{b}_i$ for random effects

if there are more components then the components are not well separated. We saw in the case of data set 5 (figure 5.3b) which also had fused components, that we required a slightly higher value for the parameters of the Dirichlet prior for weight distribution. Thus we tested $\mathrm{Dir}(1, 1, ...1)$, $\mathrm{Dir}(2, 2, ...2)$ and $\mathrm{Dir}(3, 3, ...3)$ for the blood donor data set. From the various simulations we ran we found that $\mathrm{Dir}(2, 2, ...2)$ prior gave the same results as Dirichlet priors with slightly higher values of the hyperparameters and so we stuck with it. As for the length of MCMC chains, we ran 500,000 iterations with a thinning of 150 and a burn in of 100,000 iterations.

The first model we fitted was without any mixture of random effects. The parameter estimates we obtained from this model are shown in table 6.2. It is quite clear from these results that the parameter estimates Bayesian analysis are almost the same as the parameter estimates from frequentist analysis.

Table 6.2: Bayesian fixed effect estimates and random effect estimates for model 6.1. The intercept and #donationLast2Years have been removed for hierarchical centering.

Fixed Effects (β)	Mean	Median	95% HPDI
Age	-0.086	-0.086	-0.135, -0.034
TSPD	-0.037	-0.037	-0.068, -0.009
Season (Hot)	-0.080	-0.080	-0.109, -0.048
Donate (TRUE)	0.157	0.157	0.083, 0.232
#donationLast2Years * TSPD	0.020	0.020	0.009, 0.031
#donationLast2Years * Donate	4.694	4.715	2.492, 6.492
#donationLast2Years^2	-50.247	-50.449	-77.843, -21.436

Covariance Parameters	Mean	Median	95% HPDI
$b^C_{\mathrm{Intercept}}$	94.130	94.142	93.303, 94.954
b^C_{slope}	-7.504	-7.531	-10.639, -4.531
$\sigma^2_{\mathrm{Intercept}}$	20.098	19.883	15.784, 24.559
$\sigma_{\mathrm{Intercept, Slope}}$	-13.989	-13.771	-20.879, -7.903
$\sigma^2_{\mathrm{Slope}}$	39.396	39.061	27.895, 53.443
σ^2	0.201	0.201	0.191, 0.213

Next we fitted models with 2 to 5 components in the mixture of random effects. In the case of

fitting two components to the mixture we applied identifiability constraints on the weights, i.e. $\eta_1 < \eta_2$. Without these constraints we observed label switching and thus the posteriors were bimodal. For 3 or more components however identifiability constraints were not useful. For e.g. in the case of fitting 3 components we observed convergence only in some chains when we increased the hyperparameters for the Dirichlet prior (prior for η) to 13. Since the prior was not informative we discarded the corresponding results.

Nevertheless for each of the resulting MCMC chains we calculated the various definitions of DIC given in section 4.1, the results of which are presented in Table 6.3. In the simulation study we had observed that DIC_4 was the most discerning of the DIC's, especially also when the mixture components were not well separated. In the latter case DIC_1, DIC_2 and DIC_3 were not much discerning. DIC_5 and DIC_6 did not select the right model even when components were well separated. Thus for the blood donor data set we primarily relied on DIC_4. We could not consider models with 3 or more components because the corresponding chains did not converge. Looking at the results in Table 6.3 it seems that a model with 2 components is preferable over a model with no mixture. As for the pattern of DIC_4 stabilizing for overfitted components, we did observe it for blood donor data set as well and so it was not specific to the data sets in the simulation study. However, after running multiple chains it was not observed as consistently as it was observed for the simulation study data sets.

Table 6.3: DIC for blood donor data set

# Components Fitted	DIC_1	DIC_2	DIC_3	DIC_4	DIC_5	DIC_6
1 comp	4817	4816	4818	7077	7532	4353
2 comp	4808	4805	4811	6956	7376	4306
3 comp	4755	4825	4838	7024	7406	4360
4 comp	4608	4723	4814	7031	7360	4332
5 comp	4461	4459	4818	7031	7108	4327

# Comp Fitted	p_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
1 comp	13	12	15	13	468	302
2 comp	17	16	19	19	439	247
3 comp	-55	-15	29	5	387	294
4 comp	-183	-67	23	30	360	272
5 comp	-334	-336	23	28	104	269

6.3.1 PPC for blood donor data set

To further verify if 2 components are the correct number of components we used posterior predictive checks. The PPC which we mentioned in section 4.3.1 had an interesting property that it worked for overfitted models as well because in overfitted models the chains corresponding to fixed effects converged and only the chains corresponding to parameters of some of the component densities did not. To perform PPC we required a sample of number of donations in last 2 years. These would form the Z matrix in the Bayesian heterogeneity model definition. Instead of randomly generating the number of donations in last 2 years, we sampled them from the subjects which were not considered for analysis due to computational restrictions. We then generated random effects from the posterior predictive distribution of random effects. Figure 6.3a to 6.3c show the distribution of the test statistic 4.21 for the 5 models we considered. It can be seen that the distributions have the characteristic skewed distribution due to the posteriors of parameters being almost equal to the priors in case of 3 or more number of components in

the mixture (see section 4.3.1). Another aspect of the result is that the distribution of the test statistic with posterior predictive data overlaps the distribution of test statistic with the sample data. This shows that the model is well fitted to the data. Lastly the PPP values for the various models are shown in Table 6.4. It can be seen that the PPP values are lowest for model with 2 components. However they are not extreme for the overfitted models, a phenomenon which we also observed and discussed in simulation study results (section 5.1.5). Based on the results of PPC and DIC both, we decided to choose the model with 2 components in the mixture.

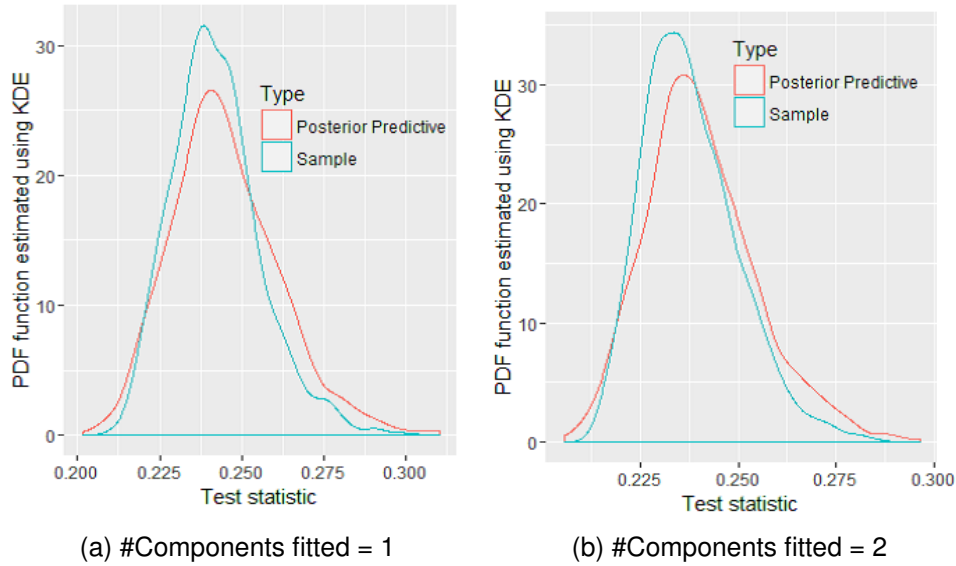(a) #Components fitted = 1   (b) #Components fitted = 2

Figure 6.2: PDF function of $T(\tilde{r})$ for underfitted/rightly fitted models applied to the blood donor data set.

Table 6.4: PPP values for the various models fitted to the blood donor data set.

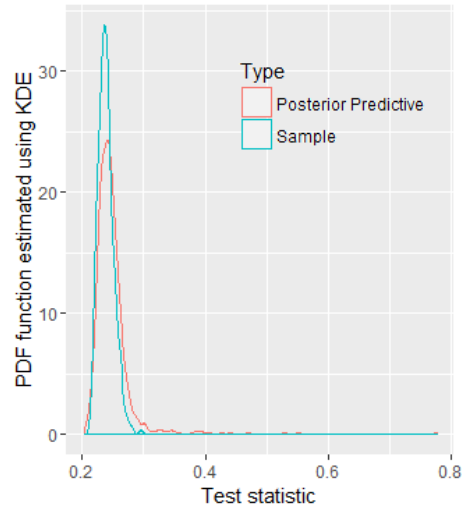| # Components fitted	1	2	3	4	5
PPP Value	0.619	0.591	0.677	0.701	0.758

6.4 Parameter estimates for the model with 2 components in mixture

Table 6.5 shows the parameter estimates for the component densities of the mixture of random effects. Further, figure 6.4 to figure 6.6 show the posterior densities of the weight distribution, mean of the component densities and covariance parameters $\sigma^2_{\text{Intercept}}$, σ^2_{Slope}, $\sigma_{\text{Intercept, Slope}}$ which form the covariance matrix of component densities. For weight distribution η if we consider the mean of the posterior distributions then the weight distribution ratio for the components is 1:3. Secondly we can see that the posterior densities for location parameters for slope have a slight overlap but that is not the case with random intercept.

The first component in the mixture corresponds to the group for which the baseline random component of Hemoglobin is low and and the second component corresponds to the group for which the baseline random component of Hemoglobin is higher. However for the second group the slope is more negative suggesting that relative to group 1 an average subject from group 2 has a more rapid decrease in Hemoglobin levels with frequent donations. Lastly, within group
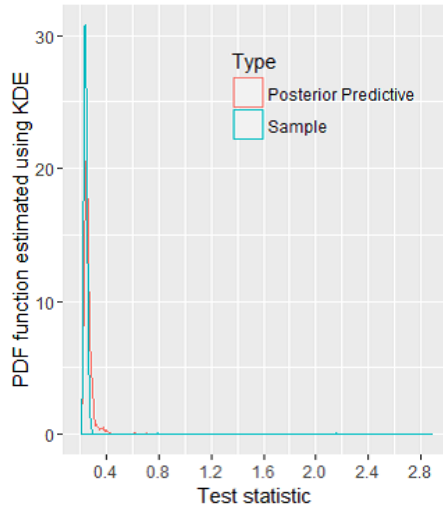
(a) #Components fitted = 3

(b) #Components fitted = 4
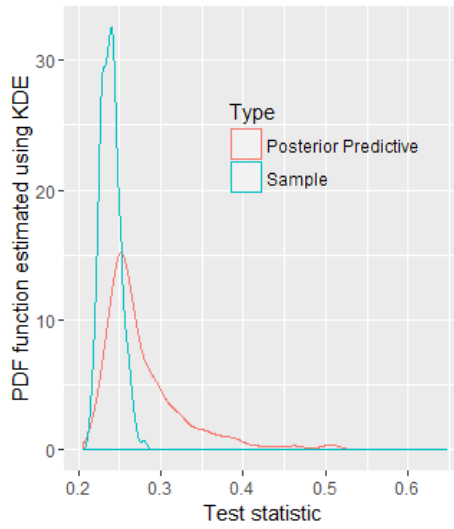
(c) #Components fitted = 5

Figure 6.3: PDF function of $T(\tilde{r})$ for overfitted models applied to the blood donor data set.

2 a person who begins with a higher baseline Hemoglobin level, is likely to have a more rapid decrease in Hemoglobin levels with donations. However within group 1 such intra subject difference cannot be generalized because 95% HPDI for the correlation between random intercept and random slope is [-0.847, 0.980].

As for the fixed effects, we can see that subjects with higher age tend to have a lower Hb. Also during hot seasons the Hb levels are lower than in cold seasons. Both of these findings are consistent with the findings of Nasserinejad et al., (2013, 2015, 2016). We did not discuss the other fixed effects as the corresponding covariates also interact with other covariates in the model.

Table 6.5: Bayesian parameter estimates for the model with 2 components in the mixture of random effects, applied to the Blood donor data set.

Fixed Effects (β)	Mean	Median	95% HPDI
Age	-0.07	-0.07	-0.118, -0.024
TSPD	-0.038	-0.038	-0.066, -0.009
Season (Hot)	-0.081	-0.081	-0.113, -0.05
Donate (TRUE)	0.160	0.159	0.091, 0.234
#donationLast2Years * TSPD	0.020	0.020	0.008, 0.030
#donationLast2Years * Donate	4.738	4.746	2.812, 6.733
#donationLast2Years^2	-48.360	-48.256	-76.647, -18.457

Component Parameters	Mean	Median	95% HPDI
η_1	0.275	0.265	0.137, 0.484
η_2	0.725	0.735	0.516, 0.863
$b^C_{\text{Intercept-1}}$	90.213	90.191	88.657, 91.67
$b^C_{\text{Slope-1}}$	-4.835	-4.896	-8.626, -0.827
$b^C_{\text{Intercept-2}}$	95.645	95.578	94.246, 97.339
$b^C_{\text{Slope-2}}$	-8.676	-8.627	-12.365, -5.386
$\sigma^2_{\text{Intercept-1}}$	2.108	1.562	0.067, 5.65
$\sigma_{\text{Intercept-1,Slope-1}}$	0.538	0.245	-2.571, 5.403
$\rho_{\text{Intercept-1,Slope-1}}$	0.202	0.334	-0.847, 0.980
$\sigma^2_{\text{Slope-1}}$	1.975	0.927	0.069, 7.579
$\sigma^2_{\text{Intercept-2}}$	18.176	18.146	11.677, 25.416
$\sigma_{\text{Intercept-2,Slope-2}}$	-14.542	-14.441	-24.496, -5.069
$\rho_{\text{Intercept-2,Slope-2}}$	-0.479	-0.492	-0.69, -0.252
$\sigma^2_{\text{Slope-2}}$	49.437	48.584	31.177, 72.200

Variance Parameters	Mean	Median	95% HPDI
σ^2	0.202	0.202	0.192, 0.212

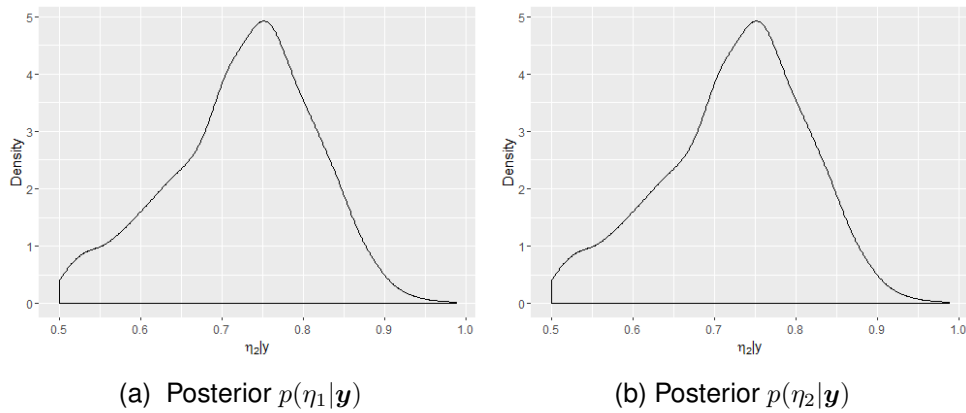(a) Posterior $p(\eta_1|\boldsymbol{y})$ (b) Posterior $p(\eta_2|\boldsymbol{y})$

Figure 6.4: Posterior densities of the weight distribution for the components of the mixture of random effects for the blood donor data set.

(a) Posterior $p(b_{\text{Intercept-1}}^C|\boldsymbol{y})$

(b) Posterior $p(b_{\text{Slope-1}}^C|\boldsymbol{y})$

(c) Posterior $p(b_{\text{Intercept-2}}^C|\boldsymbol{y})$

(d) Posterior $p(b_{\text{Slope-2}}^C|\boldsymbol{y})$
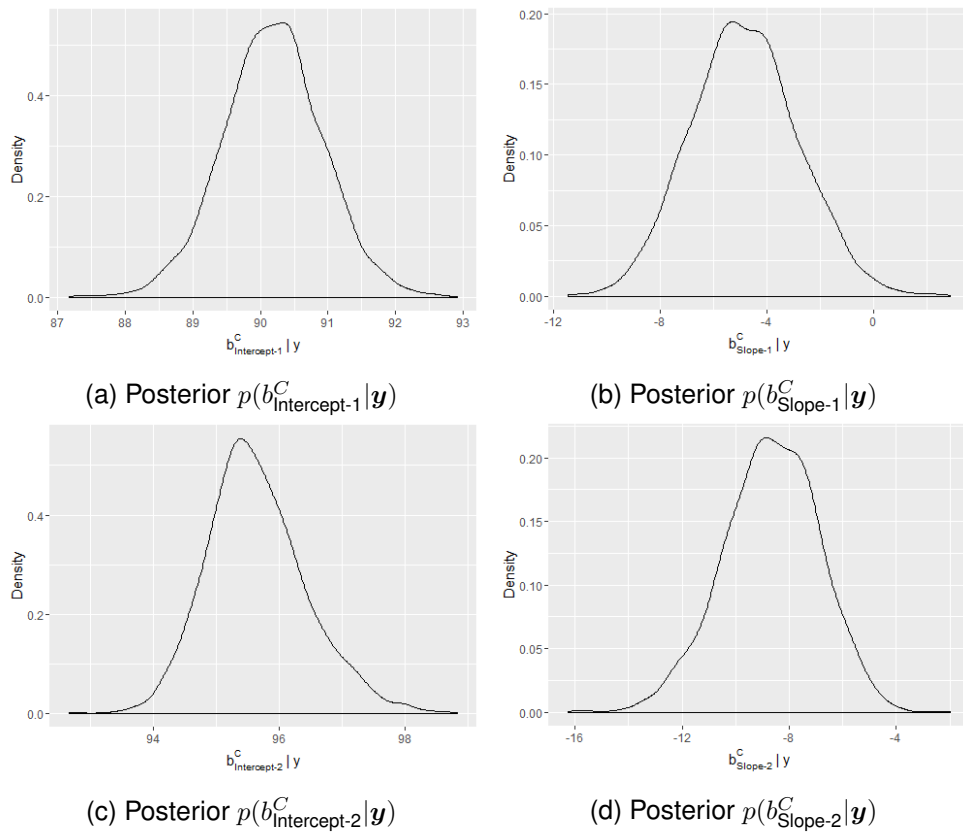
Figure 6.5: Posterior densities of the location parameters for the components of the mixture of random effects for the blood donor data set.

(a) Posterior $p(\sigma^2_{\text{Intercept-1}}|\boldsymbol{y})$

(b) Posterior $p(\sigma_{\text{Intercept-1,Slope-1}}|\boldsymbol{y})$

(c) Posterior $p(\sigma^2_{\text{Slope-1}}|\boldsymbol{y})$

(d) Posterior $p(\sigma^2_{\text{Intercept-2}}|\boldsymbol{y})$

(e) Posterior $p(\sigma_{\text{Intercept-2,Slope-2}}|\boldsymbol{y})$

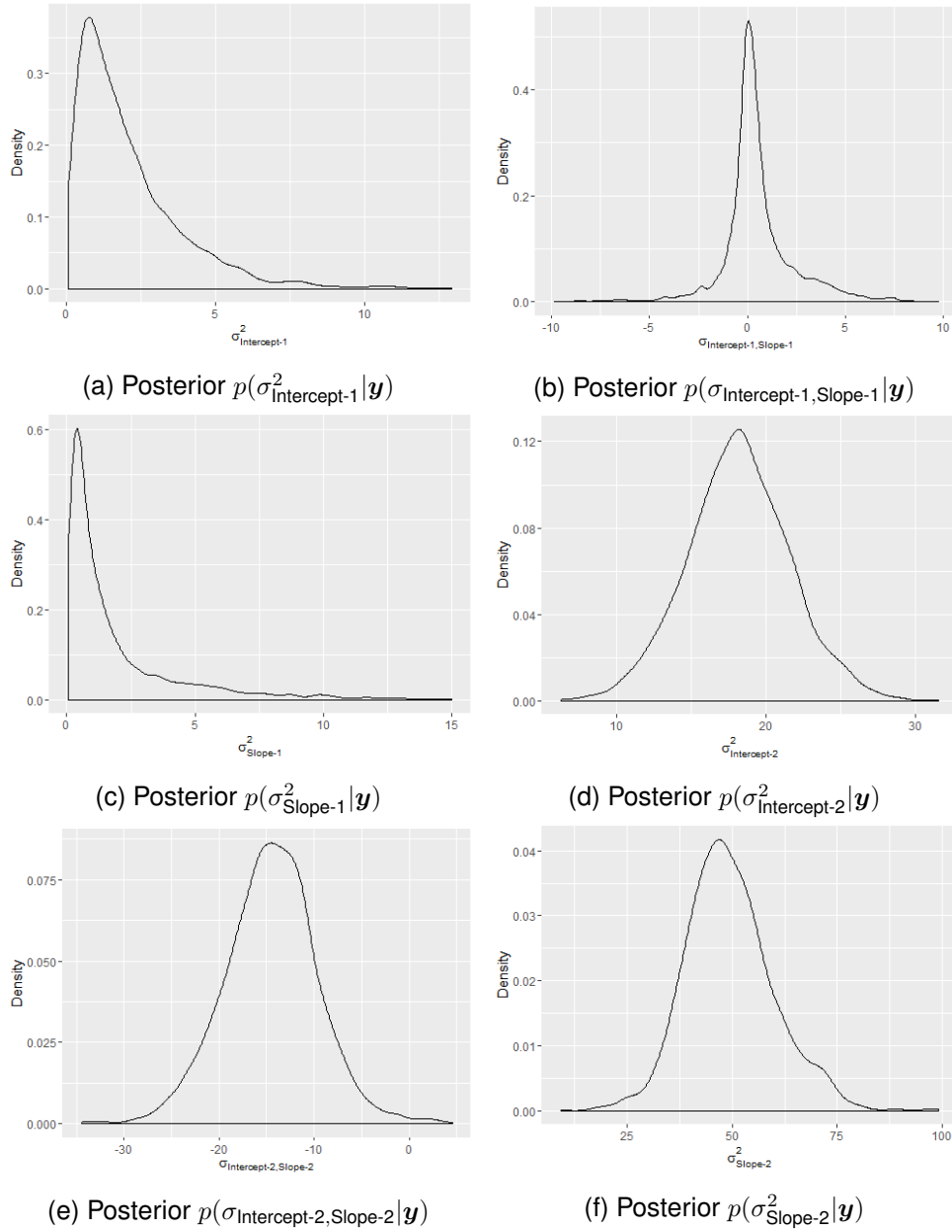(f) Posterior $p(\sigma^2_{\text{Slope-2}}|\boldsymbol{y})$

Figure 6.6: Posterior densities of the variance covariance parameters for the components of the mixture of random effects for the blood donor data set.

Chapter 7

Conclusion

As part of this thesis we evaluated the efficacy of DIC, Marginal likelihood and Posterior predictive checks for selecting the right number of components in the mixture distribution of random effects for a Bayesian heterogeneity model. We first gave the various definitions of DIC based on observed data likelihood, complete data likelihood and conditional data likelihood. We found that DIC_4 which was a complete DIC was the most reliable among the various definitions of the DIC's. This result also matched the findings of Celeux et al., (2006). We also found that when the components are not well separated or if the number of subjects are less then DIC_1, DIC_2 and DIC_3 may not discern among the various models much and may also lead to the choice of underfitted models. They can also be used alongside DIC_4 when the components are well separated. We also confirmed the suggestion of Frühwirth-Schnatter, (2013) that in cases where components are not well separated, to avoid sampling of empty components one may have to use a Dirichlet prior with slightly large values for the hyperparameter.

Secondly, we found that Bayes Factor via Chib's approximation was not a reliable method to detect the number of components even when the chains are converged. To use Chib's approximation one has to put constraints on the ordering of the components in the MCMC simulations, otherwise while doing chib's approximation one may get further MCMC chains with a different ordering of components. This can result into very large Bayes factor estimates.

For posterior predictive checks we employed the mixed predictive check technique suggested by Marshall and Spiegelhalter, (2003) and found that overfitting models can be detected easily by exploiting non identifiability due to empty components. Empty components have posterior estimates sampled from the priors, thus giving estimates which do not support the data. More specifically we obtained right skewed distribution for the test statistic in case of the overfitted models. However because the nearly components have very small weights, the posterior predictive check requires a very large sample size. We also found that it was difficult to detect models with less number of components in the mixture because such models also gave a very good fit to the data. Having said that, in case the number of components are fitted correctly, but the posterior estimates are poor then the poor model fit can still be detected via PPC. This was especially observed in the case of using uniform prior for correlation and inverse gamma priors for random component variances, because the random component variances were underestimated.

Lastly, we applied the Bayesian heterogeneity model to the blood donor data set and found a mixture of 2 components to be ideal for the distribution of random effects. This was in contrast with the number of components proposed by Nasserinejad et al., (2015) using growth mixture models. It is however interesting to note that the components we found were a subset of the components proposed by Nasserinejad et al., (2015). More specifically the two components that we observed differed in a way such that the subgroup of subjects with higher baseline

Hemoglobin level also had a rapid decrease in Hemoglobin level with number of donations.

Bibliography

Brigo, Damiano and Fabio Mercurio (2002). "Lognormal-mixture dynamics and calibration to market volatility smiles." In: *International Journal of Theoretical and Applied Finance* 05.04, pp. 427–446. DOI: 10.1142/S0219024902001511.

Celeux, G. et al. (2006). "Deviance information criteria for missing data models." EN. In: *Bayesian Analysis* 1.4, pp. 651–673. DOI: 10.1214/06-BA122.

Chib, Siddhartha (1995). "Marginal Likelihood from the Gibbs Output." In: *Journal of the American Statistical Association* 90.432, pp. 1313–1321. DOI: 10.2307/2291521.

Congdon, Peter D. (2010). *Applied Bayesian Hierarchical Methods*. English. Boca Raton: Chapman and Hall/CRC.

Day, N. E. (1969). "Estimating the Components of a Mixture of Normal Distributions." In: *Biometrika* 56.3, pp. 463–474. DOI: 10.2307/2334652.

Frühwirth-Schnatter, Sylvia (2013). *Finite Mixture and Markov Switching Models*. English. 2006 edition. Springer.

Frühwirth-Schnatter, Sylvia, Regina Tüchler, and Thomas Otter (2004). "Bayesian Analysis of the Heterogeneity Model." In: *Journal of Business & Economic Statistics* 22.1, pp. 2–15. DOI: 10.1198/073500103288619331.

Fu, Zhaoxia and Liming Wang (2012). "Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm." en. In: *Multimedia and Signal Processing*. Ed. by Fu Lee Wang et al. Communications in Computer and Information Science 346. Springer Berlin Heidelberg, pp. 61–66.

Gelman, Andrew and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. English. 1 edition. Cambridge ; New York: Cambridge University Press.

Gianola, Daniel et al. (2007). "Mixture models in quantitative genetics and applications to animal breeding." In: *Revista Brasileira de Zootecnia* 36, pp. 172–183. DOI: 10.1590/S1516-35982007001000017.

Gruen, Bettina and Martyn Plummer (2015). *bayesmix: Bayesian Mixture Models with JAGS*.

Kiefer, J. and J. Wolfowitz (1956). "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." EN. In: *The Annals of Mathematical Statistics* 27.4, pp. 887–906. DOI: 10.1214/aoms/1177728066.

Komárek, Arnošt (2015). *mixAK: Multivariate Normal Mixture Models and Mixtures of Generalized Linear Mixed Models Including Model Based Clustering*.

Lesaffre, Emmanuel and Andrew B. Lawson (2012). *Bayesian Biostatistics*. English. 1 edition. Chichester, West Sussex: Wiley.

Lesosky, Maia et al. (2012). "A live weight–heart girth relationship for accurate dosing of east African shorthorn zebu cattle." en. In: *Tropical Animal Health and Production* 45.1, pp. 311–316. DOI: 10.1007/s11250-012-0220-3.

Lewicki, Michael S. (1994). "Bayesian Modeling and Classification of Neural Signals." In: *Neural Computation* 6.5, pp. 1005–1030. DOI: 10.1162/neco.1994.6.5.1005.

Marshall, E. C. and D. J. Spiegelhalter (2003). "Approximate cross-validatory predictive checks in disease mapping models." en. In: *Statistics in Medicine* 22.10, pp. 1649–1660. DOI: 10.1002/sim.1403.

Marín, Xavier Fernández i (2016). *ggmcmc: Tools for Analyzing MCMC Simulations from Bayesian Inference*.

Nasserinejad, Kazem et al. (2013). "Predicting hemoglobin levels in whole blood donors using transition models and mixed effects models." In: *BMC Medical Research Methodology* 13, p. 62. DOI: 10.1186/1471-2288-13-62.

Nasserinejad, Kazem et al. (2015). "Prevalence and determinants of declining versus stable hemoglobin levels in whole blood donors." eng. In: *Transfusion* 55.8, pp. 1955–1963. DOI: 10.1111/trf.13066.

Nasserinejad, Kazem et al. (2016). "Prediction of hemoglobin in blood donors using a latent class mixed-effects transition model." eng. In: *Statistics in Medicine* 35.4, pp. 581–594. DOI: 10.1002/sim.6759.

Povey, Daniel et al. (2011). "The subspace Gaussian mixture model—A structured model for speech recognition." In: *Computer Speech & Language*. Language and speech issues in the engineering of companionable dialogue systems 25.2, pp. 404–439. DOI: 10.1016/j.csl.2010.06.003.

Richardson, Sylvia. and Peter J. Green (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)." en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.4, pp. 731–792. DOI: 10.1111/1467-9868.00095.

Shoham, Shy, Matthew R. Fellows, and Richard A. Normann (2003). "Robust, automatic spike sorting using mixtures of multivariate t-distributions." In: *Journal of Neuroscience Methods* 127.2, pp. 111–122. DOI: 10.1016/S0165-0270(03)00120-1.

Sim, Adelene Y. L. et al. (2012). "EVALUATING MIXTURE MODELS FOR BUILDING RNA KNOWLEDGE-BASED POTENTIALS." In: *Journal of bioinformatics and computational biology* 10.2, p. 1241010. DOI: 10.1142/S0219720012410107.

Simancas-Acevedo, Eric et al. (2001). "Speaker Recognition Using Gaussian Mixtures Models." en. In: *Bio-Inspired Applications of Connectionism*. Ed. by José Mira and Alberto Prieto. Lecture Notes in Computer Science 2085. Springer Berlin Heidelberg, pp. 287–294.

Spiegelhalter, David J. et al. (2002). "Bayesian measures of model complexity and fit." en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639. DOI: 10.1111/1467-9868.00353.

Stephens, Matthew (2000). "Dealing with label switching in mixture models." en. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, pp. 795–809. DOI: 10.1111/1467-9868.00265.

Su, Yu-Sung and Masanao Yajima (2015). *R2jags: Using R to Run 'JAGS'*.

Titterington, D. M., Adrian F. M. Smith, and U. E. Makov (1986). *Statistical Analysis of Finite Mixture Distributions*. English. 1 edition. Chichester ; New York: Wiley.

Verbeke, Geert and Emmanuel Lesaffre (1996). "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population." In: *Journal of the American Statistical Association* 91.433, pp. 217–221.

Verbeke, Geert and Geert Molenberghs (2009). *Linear Mixed Models for Longitudinal Data*. en. Springer Science & Business Media.

Xiang, Bing and T. Berger (2003). "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network." In: *IEEE Transactions on Speech and Audio Processing* 11.5, pp. 447–456. DOI: 10.1109/TSA.2003.815822.

Yang, Narendra Ahuja Ming-hsuan (1998). "Gaussian Mixture Model for Human Skin Color and Its Applications in Image and Video Databases." In: *Proc SPIE* 3656. DOI: 10.1117/12.333865.