KU LEUVEN

FACULTY OF SCIENCE

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

© Copyright by KU Leuven

Without written permission of the promotors and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculteit Wetenschappen, Geel Huis, Kasteelpark Arenberg 11 bus 2100, 3001 Leuven (Heverlee), Telephone +32 16 32 14 01.

A written permission of the promotor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

Preface

The following thesis work was conducted as part of the credits completion requirements of the MSc. Statistics programme at KU Leuven. The problem statement of the thesis is 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, 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 analytical calculations of the various definitions of Deviance information criteria, marginal likelihood and choice of posterior predictive checks. Very soon 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 the results. Looking backwards, I think it 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 and introduction to the Bayesian paradigm in Chapter 2. Further in Chapter 3 the definition of a Bayesian heterogeneity model and issues related to the estimation of parameters involved are presented. Chapter 4 constitutes the analytical derivations I did for various classes of Deviance information crieteria, marginal likelihood and posterior predictive checks, used in model selection. Chapter 5 includes the results of the simulation study that was performed to check the efficacy of the aforemented Bayesian model selection methods. In chapter 6 the results of modeling the Blood donor data set (Nasserinejad et al., 2015) using a Bayesian heterogeneity model are presented. The results from the simulation study are used to apply the right model selection criteria on the models formulated for the Blood donor data set.

I am grateful to my supervisor Professor Dr. Emmanuel Lesaffre for keeping faith in my capabilities and for guiding me in the right direction. I enjoyed the fact that he never spoonfed 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 Geert Molenberghs and Professor Geert Verbeke for the captivating lectures on longitudinal data analysis, which empowered me with the tools required for performing the frequentist analysis of blood donor data set. I am thankful to Kazem Nasserinejad from ErasmusMC for resolving many of my queries regarding the blood donor data set, and to Igor Milhoranca for providing the much needed inputs at crucial times. Lastly, I am grateful to my parents for the innumerable sacrifices they made to make sure 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. A mixture distribution for random effects allows to model the heterogeneity introduced by ignoring certain covariates in the mean structure of the model or to take into account the unknown non normal distriution for random effects. We then explored effectiveness of Bayesian model selection criteria (DIC, Bayes Factor, PPC) for choosing the number of component densities in the mixture distribution of random effects. Since mixture models are missing data models, we implemented various definitions of DIC as given by Celeux et al., (2006) for such models. We found that DIC 4 based on complete data likelihood was a fairly good selection criteria. However as the sample size decreased the discerning power of DIC also decreased. We then implemented Bayes Factor based on the approximation given by Chib, (1995) and found that it was not reliable for deciding on number of components required in the model. On the other hand, Posterior predictive checks were a very strong discerning method if indepdent inverse gamma priors were used for variance components, and uniform distribution for correlation, in the distribution of random effects. In regards to the choice of prior distribution for covariance parameters, we found that a Wishart prior for precision matrix(inverse of covariance matrix) overestimates the precision when within subject variance is greater than between subject variance. Thus, it could be a good idea to decrease scale of the intercept and the covariate corresponding to random slope, so that the corresponding variances increase in magnitude.

Contents

Chapter 1

Introduction

In this chapter we will first introduce a mixture distribution and then mention the challenges involved in estimation of parameters of a mixture distribution. We will also highlight the benefits of using a Bayesian approach for parameter estimation. Lastly we will 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 which firstly 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 like 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 each component density. 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 were having 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 in an example by Titterington, Smith, and Makov, (1986) it was shown that a GMM of two components could be indistinguishable from a lognormal distribution.

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

$$p(y) = \sum_{k=1}^{K} \eta_k p_k(y)$$

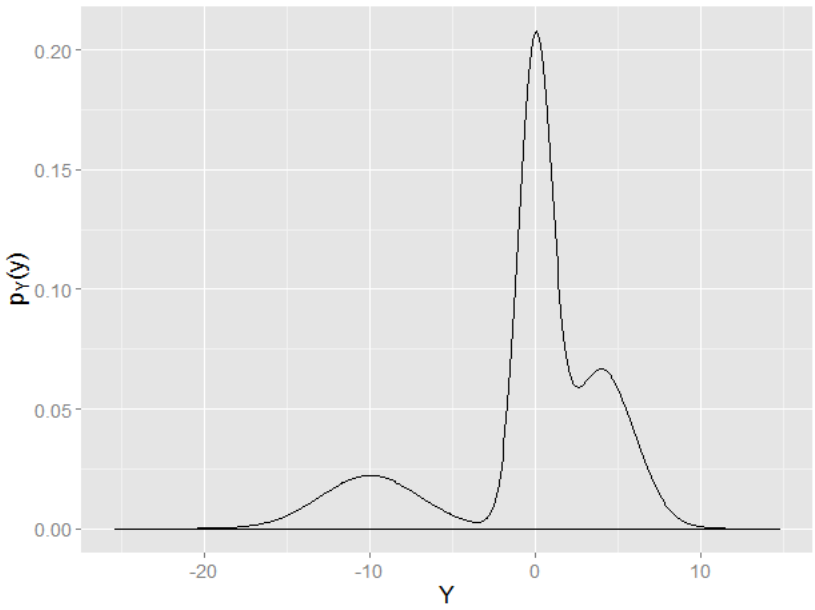


Figure 1.1: Mixture density of $Y \sim \frac{1}{6}N(-10, 3) + \frac{1}{2}N(0, 1) + \frac{1}{3}N(4, 2)$

The vector of the weights $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$ is called the weight distribution. The k^{th} weight η_k corresponds to selection probability of the k^{th} density while sampling for Y. It 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 $\boldsymbol{S} = (S_1, S_2, \ldots, S_N)$ 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 based methods such as the EM algorithm could be used to deal with the above mentioned challenges, there are certain downsides to them. Firstly it is well known that 95% confidence intervals of ML estimates are based on asymptotical 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. Secondly, in case of univariate and multivariate GMM, the maximum likelihood function,

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

is unbounded and has many spurious nodes (Day, 1969; Kiefer and Wolfowitz, 1956). A bayesian

approach however, handles this problem elegantly using priors for parameters (μ, σ^2), as shown by Frühwirth-Schnatter, (2013, pg. 176).

1.1.3 Applications of mixture distribution

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 lognormal 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 is not of known form. However mixtures have also been used as supplementary methodology in various models, a list of which can be found in Frühwirth-Schnatter, (2013, pg. 238). One such usage in linear mixed models has been proposed by Verbeke and Lesaffre, (1996) and it also forms the theme 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. For the scope of this thesis our focus will be on the Bayesian version of the linear mixed effects model(BLMM), 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 is 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 will be working in the Bayesian framework, we will use MCMC methods instead of the frequentist point estimation methods. We will simulate data sets to check efficacy of each of the aforementioned model selection criteria and then use 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

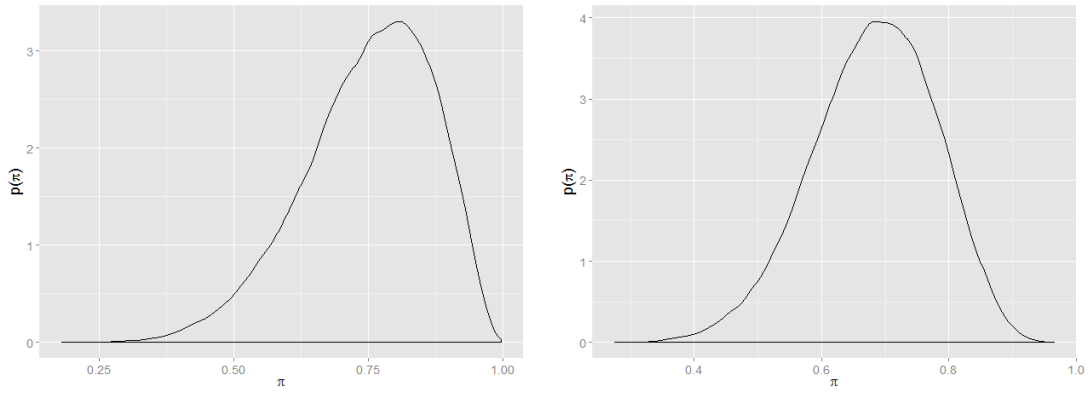
Bayesian paradigm

In this chapter we will give an introduction to the foundations of the Bayesian framework. 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 are random variables rather than being a constant. The distribution of parameters based on the data at hand is called the posterior distribution, represented by $p(\theta|y)$. Whereas the inital distribution of parameters is called the prior distribution, represented by $p(\theta)$. We will now present an example to signify the ideological differences between the Bayesian paradigm and frequentist paradigm.

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 real 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 subjective, the beliefs of captain B 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 will now present the Bayes rule which is central to the Bayesian parameter estimation process. The Bayes rule for the estimating the continuous parameter π is given by

$$p(\pi|\boldsymbol{y}) = \frac{L(\pi|\boldsymbol{y})p(\pi)}{p(\boldsymbol{y})} = \frac{L(\pi|\boldsymbol{y})p(\pi)}{\int_0^1 L(\pi|\boldsymbol{y})p(\pi)\,\mathrm{d}\pi} \tag{2.1}$$

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

We can apply Bayes rule to estimate 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; For e.g. if the prior belief $p(\pi)$ in our example is that it is trimodal then we will have to use numerical approximation for calculation of the posterior. 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 intervals 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 the interval from which parameter θ takes a value 95 out of 100 times. 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 paramters $p(\theta|y)$. 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 will stick to JAGS which is from the BUGS(Bayesian inference Using Gibbs Sampling) family. We will also 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 data which is hierarchical in structure. 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 the random effects 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 will be 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 will be correlated, and given the fact that a linear model imposes homoscedasticity, it is not suitable for use in such scenarios.

3.1.1 LMM definition

Following the notations from Lesaffre and Lawson, (2012), the LMM for the observations of the i^{th} subject among the n subjects is given by

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

where $1 \leq i \leq n$,
$\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T$ is a vector of observations for the i^{th} subject taken at m_i time points,
$\boldsymbol{X}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \ldots, \boldsymbol{x}_{im_i}^T)^T$ is the $m_i \times (d+1)$ design matrix for the i^{th} subject,
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$ 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.

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 considers the variability in parameter estimates in the first place. A similar problem occurs in the estimation of $\boldsymbol{b}_i$. The frequentist strategy is to use Empirical Bayes estimates, where the the posterior distribution of random effects uses 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, in this thesis we will model 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. This 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, the random effects are assumed to be multivariate normally distributed. It could be too strong an assumption though in certain cases. A classical example are the longitudinal studies where at any 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, however in a study where we would like to categorize patients into different groups well before diagnosis this could be difficult. Here is a toy example for it. Imagine that in longitudinal study we are measuring a response Y which is an 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 will not be knowing 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 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

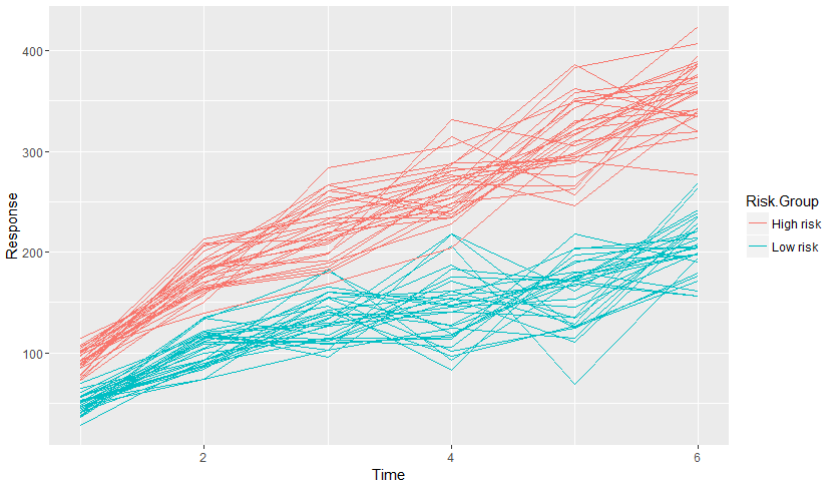


Figure 3.1: Individual profiles of 30 subjects from each group.

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 k^{th} component in the mixture distribution respectively. The vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$ is the weight distribution for the component densities. The vector $\boldsymbol{S} = (S_1, S_2, ..., S_n)$ 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 will discuss some of the challenges in Bayesian estimation of parameters in the Bayesian heterogeneity model. We will also discuss the approaches we used to deal with them in this thesis.

3.4.1 Marginal vs. Hierarchical model

Suppose that in our heterogeneity model we know the alloations S_i for each subject. Then conditional on knowing $S_i = k$ the following LMM equation has a hierarchical interpretation.

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

One can however integrate out the random effects $\boldsymbol{b}_i$ and obtain the corresponding marginal Bayesian heterogeneity model,

$$\begin{aligned} \boldsymbol{y}_i|S_i &\sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{\varepsilon}_i^*) \\ \boldsymbol{\varepsilon}_i^* &\sim N_{m_i}(\boldsymbol{0}, \boldsymbol{Z}_i G_k \boldsymbol{Z}_i^T + R_i) \end{aligned} \tag{3.3}$$

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, the marginal model took quite a long time for each iteration. It also did not give posterior estimates of the random effects $\boldsymbol{b}_i$ 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 $\boldsymbol{b}_i$ in a mixed model could be seen as random deviations from the fixed effects (β) with a mean $\boldsymbol{0}$. 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ both share columns corresponding to the variable intercept. To enforce the mean $\boldsymbol{0}$ on the random effects in a mixture distribution of random effects, the following condition should be satisfied.

$$E(\boldsymbol{b}_i|\boldsymbol{\nu}) = \sum_{k=1}^{K} \eta_k N_q(\boldsymbol{b}_k^C, G_k) = 0 \tag{3.4}$$

where $\boldsymbol{\nu}$ is defined in section 3.3.1. This further means that $E(\boldsymbol{y}_i|\boldsymbol{\nu}) = \boldsymbol{X}_i\boldsymbol{\beta}$. This parametrization, which was also used in the original paper on Heterogeneity model (Verbeke and Lesaffre, 1996) is called the noncentralized parametrization. The centralized parametrization assumes that the random effects are not deviations from the fixed effects and are centred 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 hierarchically centred parametrization.

3.4.3 Starting values

3.4.4 Choice of priors

Since we are following a Bayesian paradigm, parameters in the Bayesian heterogeneity model are random variables and thus need to have a prior distribution. There are certain difficulties in specifying the prior though, especially that it can be difficult to implement theoretically preferred prior distributions. As an example we will begin with the choice of prior for the mean (b_k^C) and covariance matrix (G_k) of component densities in the mixture distribution of random effects. Both b_k^C and G_k are unknown, and hence to obtain the joint posterior as a known density one is forced to specify the conditionally conjugate prior $b_k^C|G_k \sim N(\boldsymbol{\mu}_0, \frac{G_K}{N_0})$ and $G_k^{-1} \sim \mathcal{W}(n_0, \Psi)$. Here $\boldsymbol{\mu}_0, N_0, n_0, \Psi$ are the hyperparameters for the corresponding prior distributions. Since JAGS only allows specifying marginal priors, one will have to specify a multivariate T distribution for b_k^C. However the problem with this approach is that the choice of the right hyperparameters is debatable (Frühwirth-Schnatter, 2013, pg. 192) and even if one does, the extra computationally intensive procedure does not provide much advantage in practice. It is thus a widespread practice to use 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 to specifying indepedent diffuse univariate normal priors for the mean of random intercept and for the mean of random slope.

Choice of prior for covariance matrix

The choice of prior for the covariance matrix (G_k) is an interesting problem. 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. For precision matrix G_k^{-1} one can use 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 diagnoal elements of scale matrix of wishart distribution influenced the posterior more than the likelihood did.

Lesaffre and Lawson, (2012, pg. 260) also suggest using indepdent gamma priors for random intercept and random slope and uniform prior $U(-1, 1)$ for the correlation between the two. The upside of this approach is that it gives 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. This property can be used to make decisive posterior predictive checks.

Choice of priors for β and σ^2

We assume that the parameters β and σ^2 are indepdent from $\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, ..., \boldsymbol{b}_K^C, G_1, G_2, ..., G_K, \boldsymbol{\eta}$. The problem of choosing a conjugate prior for β and σ^2 is similar to what we discussed in the section above. The solution thus is alike, i.e. using indepdent 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 $\boldsymbol{\eta}$

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 cases where one of the components is nearly empty. If one chooses the hyperparameters to be equal to 1 then label switching is observed whenever one of the components is nearly empty. We found out during the simulation study that choosing larger values for the hyperparameter indeed mitigated the issue of label switching, however it did also gave parameter estimates far from the real ones.

3.4.5 Label Switching

We use a mixture distribution for random effects in the Bayesian heterogeneity model. However we do not know the allocation vector $\boldsymbol{S}$ in advance. In this case the mixture likelihood for the response $\boldsymbol{y}$ is given by the 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 doing 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. ie. we begin with some random allocation vector $\boldsymbol{S}_{\text{initial}}$ 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 $\boldsymbol{S}_{\text{initial}}$ 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.

Now let us imagine a scenario where because of a certain $\boldsymbol{S}_{\text{initial}}$, the sampled parameter estimates are $(\mu_1, \mu_2) = (5.5, 6.5)$. So far it seems μ_1 represents μ_{C1} and μ_2 represents μ_{C2}. Now in the next step the Gibbs sampler will estimate the allocation vector $\boldsymbol{S}$ conditional on these estimates. Suppose that in this step an observation with value 6.5 which was originally from component C_2, gets allocated to component C_1 and similarly an observation with value 5.5 from component C_1 gets allocated to C_2. Unless we impose some constraint like $\mu_1 < \mu_2$, under the current scenario even $(\mu_1, \mu_2) = (6.5, 5.5)$ is likely to be be sampled by MCMC. If the sampler keeps on arbitrarily switching between the two equivalent posterior regions, then because of this label switching, one may obtain a bimodal posterior for both μ_1 and μ_2. Thus both posteriors may well remain partially explored and thus any inference based on these posteriors will not be useful.

Dealing with label switching

One of the techniques we used for dealing with label switching was imposing an identifiability constraint such as $\mu_1 < \mu_2$. The difficulty with this approach is that it is easier to do in univariate mixtures but not with multivariate mixtures. For e.g. a multivariate mixture that we have is mixture distribution for the joint distribution of random intercept and random slope. In this thesis the multivariate case was handled by putting an identifiability constraint only on either the random intercept or the random slope depending upon the variance of each random effect. It is interesting to note that if more components than needed were chosen, then label swtiching is unavoidable, and should also be seen as an indicator for overfitting (Frühwirth-Schnatter, 2013, pg. 104).

One of the other interesting techniques to deal with label switching is postprocessing of MCMC chains by relabeling the output(Richardson and Green, 1997; Stephens, 2000). We too employed this technique in the approximation of marginal likelihood (section 4.2), as that procedure also involves running further MCMC chains (expression 4.16). As we later figured out in our simulations, without careful relabeling of output one may obtain a Bayes factor $\to 0$ and thus reject the model outright.

Chapter 4

Model selection criteria

In most cases we do not know the right number of mixture components in advance unless we have some expert knowledge available or we know them from a previous/similar study. As part of this thesis we will compare many of the existing methods for finding the right number of mixture components.

4.1 Deviance information criteria

The Deviance information criteria or DIC was first proposed by Spiegelhalter et al., (2002) for Bayesian model selection. The idea 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

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

where $\bar{\boldsymbol{\theta}} = \mathsf{E}(\boldsymbol{\theta}|\boldsymbol{y})$,
$\mathsf{p_D} = -2\mathsf{E}_{\boldsymbol{\theta}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\theta}|\boldsymbol{y}])) + \log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}))$ is the penalty for model complexity, and can also be written as

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

where $D(\boldsymbol{\theta}) = -2\log p(\boldsymbol{y}|[\boldsymbol{\theta}|\boldsymbol{y}]) + 2\log f(\boldsymbol{y})$ is called the Bayesian deviance. The term $f(\boldsymbol{y})$ however cancels out in the expression for $\mathsf{p_D}$ and hence is not discussed.

4.1.1 DIC for missing data models

Mixture models and mixed models both are both a member of the clas of models called missing data models. The reason is that the allocation vector $\boldsymbol{S}$ in a mixture model and matrix of random effects $\boldsymbol{b} = (\boldsymbol{b}_1, \boldsymbol{b}_1, ..., \boldsymbol{b}_n)$ in a LMM, both 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 different impact on model selection. In this thesis we will take some of those definitions and apply them in context of the Bayesian heterogeneity model.

Observed DIC

The first category of DIC's is associated with observed data likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$ or in our case $f(\boldsymbol{y}|\beta, \sigma^2, \boldsymbol{\nu})$, where $\boldsymbol{\nu}$ is as defined in section 3.3.1. The observed likelihood can be obtained

by marginalizing over the allocation vector of subjects $\boldsymbol{S}$ and random effects $\boldsymbol{b}$. This give us the following formula for observed data likelihood.

$$f(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}) = \prod_{i=1}^{n}\sum_{k=1}^{K} f_N(\boldsymbol{y}_i;\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{Z}_i G_k \boldsymbol{Z}_i^T + R_i)\eta_k \tag{4.1}$$

Based on equation 4.1 we will now extend the definition of DIC_1, DIC_2 and DIC_3 proposed by Celeux et al., (2006) to give

$$\text{DIC}_1 = -4\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}])) + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\beta}},\bar{\sigma^2},\bar{\boldsymbol{\nu}})) \tag{4.2}$$

where $\bar{\boldsymbol{\beta}} = \text{E}(\boldsymbol{\beta}|\boldsymbol{y})$, $\bar{\sigma^2} = \text{E}(\sigma^2|\boldsymbol{y})$ and $\bar{\boldsymbol{\nu}} = \text{E}(\boldsymbol{\nu}|\boldsymbol{y})$,

DIC_2's definition is similar to DIC_1 but intead of posterior mean, posterior mode is used in calculation of $D(\bar{\boldsymbol{\theta}})$. It is given by

$$\text{DIC}_2 = -4\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}])) + 2\log p(\boldsymbol{y}|\hat{\boldsymbol{\beta}},\hat{\sigma^2},\hat{\boldsymbol{\nu}})) \tag{4.3}$$

where $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} p(\boldsymbol{\beta}|\boldsymbol{y})$, $\hat{\sigma^2} = \arg\max_{\sigma^2} p(\sigma^2|\boldsymbol{y})$ and $\hat{\boldsymbol{\nu}} = \arg\max_{\boldsymbol{\nu}} p(\boldsymbol{\nu}|\boldsymbol{y})$,

Celeux et al., (2006) suggest that for models where non identifiability of parameters is endemic, as is the case for mixtures usually, one should use an estimator $\hat{f}(\boldsymbol{y})$ for the approximation of the density $p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu})$. They further propose the following estimator for $\hat{f}(\boldsymbol{y})$ which uses posterior samples $\boldsymbol{\theta}^{(l)}$ from the l^{th} MCMC iteration of a chain of length m.

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

This gives us the following definition of DIC.

$$\text{DIC}_3 = -4\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}])) + 2\log \hat{f}(\boldsymbol{y}) \tag{4.4}$$

The other interesting aspect of this exercise was that for DIC_3 hasd a positive pd

In each of the equations 4.2, 4.3, 4.4, the calculation of $\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}]))$ can be done by approximating it using the results from the MCMC iterations in the following way.

$$\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y}])) = \frac{1}{m}\sum_{l=1}^{m}\log p(\boldsymbol{y}|\boldsymbol{\beta}^{(l)},\sigma^{2(l)},\boldsymbol{\nu}^{(l)})) \tag{4.5}$$

Complete DIC

The second class of the DIC is based on the complete data likelihood. Complete data for the i^{th} subject in a Bayesian heterogeneity model will be $(\boldsymbol{y}_i, S_i, \boldsymbol{b}_i)$. The following equation shows the complete data likelihood of the data at hand.

$$f(\boldsymbol{y},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}) = \prod_{i=1}^{n} f_N(\boldsymbol{y}_i;\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\boldsymbol{b}_i, R_i) f_N(\boldsymbol{b}_i;\boldsymbol{b}_{S_i}^C, G_{S_i})\eta_{S_i} \tag{4.6}$$

The formulation of complete data DIC is straightforward as we assume $(\boldsymbol{b},\boldsymbol{S})$ to be observed. It can be written down as,

$$\text{DIC} = -4\text{E}_{\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y},\boldsymbol{b},\boldsymbol{S}}(\log p(\boldsymbol{y},\boldsymbol{b},\boldsymbol{S}|[\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu}|\boldsymbol{y},\boldsymbol{b},\boldsymbol{S}])) + 2\log p(\boldsymbol{y},\boldsymbol{b},\boldsymbol{S}|\bar{\boldsymbol{\beta}},\bar{\sigma^2},\bar{\boldsymbol{\nu}})) \tag{4.7}$$

where $\bar{\beta} = \mathsf{E}(\beta|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S})$, $\bar{\sigma^2} = \mathsf{E}(\sigma^2|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S})$ and $\bar{\boldsymbol{\nu}} = \mathsf{E}(\boldsymbol{\nu}|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S})$,

Unfortunately $(\boldsymbol{b}, \boldsymbol{S})$ are latent and thus Celeux et al., (2006) propose integrating the expression in 4.7 with respect to $(\boldsymbol{b}, \boldsymbol{S})$ to obtain the following definition of DIC.

$$\mathsf{DIC}_4 = -4\mathsf{E}_{\beta,\sigma^2,\boldsymbol{\nu},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}|[\beta, \sigma^2, \boldsymbol{\nu}|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}])) + 2\mathsf{E}_{\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}|\bar{\beta}, \bar{\sigma^2}, \bar{\boldsymbol{\nu}})) \quad (4.8)$$

where $\bar{\beta}$, $\bar{\sigma^2}$ and $\bar{\boldsymbol{\nu}}$ remain same as for 4.7. The first part of the formula for DIC_4 is not available in closed form for Bayesian heterogeneity model, however it can still be approximated using the output of Gibbs sampler in the same way as in 4.5. However although one has to also simulate $(\boldsymbol{b}, \boldsymbol{S})$ in the MCMC iterations. The reason it works is that during each iteration of the Gibbs sampler, it simulates parameter values from the conditional distribution of the parameters. i.e. conditional on every other parameter being simulated in the chain, including the unobserved data. We further verified this approach by comparing the results of DIC_4 approximation for mixture distribution given by Celeux et al., (2006) with our approach and found the results to be differing only by a few decimal places.

The second part of DIC_4, i.e. $\mathsf{E}_{\boldsymbol{b},\boldsymbol{S}}(\log p(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}|\bar{\beta}, \bar{\sigma^2}, \bar{\boldsymbol{\nu}}))$, is also not straightforward to compute. While the expectation over $\boldsymbol{b}, \boldsymbol{S}$ can be approximated in the same way as in 4.5 but for calculating $\bar{\beta}$, $\bar{\sigma^2}$ and $\bar{\boldsymbol{\nu}}$, Celeux et al., (2006) suggest using the posterior estimates $(\boldsymbol{b}, \boldsymbol{S}|\boldsymbol{y})$ of the unobserved data. We will now give the formulae for the expected values of parameters of interest during the l^{th} iteration $\bar{\beta}^{(l)}$, $\bar{\sigma^2}^{(l)}$ and $\bar{\boldsymbol{\nu}}^{(l)}$.

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

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

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

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

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

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

$$\mathsf{DIC}_5 = -4\mathsf{E}_{\beta,\sigma^2,\boldsymbol{\nu},\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y}}(\log p(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}|[\beta, \sigma^2, \boldsymbol{\nu}|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}])) + 2\log p(\boldsymbol{y}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{S}}|\hat{\beta}, \hat{\sigma^2}, \hat{\boldsymbol{\nu}}) \quad (4.9)$$

Conditional DIC

The third class of the DIC is based on the assumption that missing data i.e. allocation vector $\boldsymbol{S}$ and random effects $\boldsymbol{b}_i$ can be seen as additional parameter rather than as missing data. We will represent the new posterior parameter space as $\boldsymbol{\theta}_{\mathsf{cond}} = (\beta, \sigma^2, \boldsymbol{\nu}, \boldsymbol{S}, \boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_n)$. This leads to the conditional data likelihood,

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

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

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

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

4.2 Marginal Likelihood

The marginal likelihood of data represents the probablilty of data given the model. This can be calculated by margilizing 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 proposed models for the data, M_1 and M_2, one can further use the quantity in equation 4.12 to calculate model evidence. The idea is to calculate the odds of model M_1 against the model M_2 given the data. i.e. Posterior odds. This ofcourse means that it is a comparative measure as $\sim M_1 = M_2$. One can write the posterior odds 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 belief in each of these models, prior odds is equal to 1 apriori. 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 $\arg\max_\theta p(\boldsymbol{\theta}|\boldsymbol{y})$ of parameters or the maximum likelihood estimate $\arg\max_\theta L(\boldsymbol{\theta}|\boldsymbol{y})$. We decided to choose the latter of the two. Further, in context of the Bayesian heterogeneity model, we will denote the selected parameter values as $\boldsymbol{\beta}^*$, σ^{2^*} and $\boldsymbol{\nu}^*$. Thus Chib's approximation for marginal likelihood on log scale is given by,

$$\log \hat{m}(\boldsymbol{y}) = \log L(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*|\boldsymbol{y}) + \log p(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*) - \log p(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*|\boldsymbol{y}) \tag{4.14}$$

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

Firstly $\log L(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\beta}^*, \sigma^2, \boldsymbol{\nu}^*)$ can be easily determined using the formula given in equation 4.1. As for the calculation of $\log p(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*)$, it is also straightforward as we take independent priors for these parameters and they are well known in advance. The details of the priors we chose are given in section 3.4.4. Assuming that the parameters of component densities of the mixture distribution of random effects are indepdent, one can use the following to calculate $\log p(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^*|\boldsymbol{y})$.

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

An interesting problem one faces in such an expansion is that the posteriors may not be available as well known density. For e.g. we began with choosing indepdent gamma priors for precision parameters of random effects and uniform prior for correlation. However the posterior density was not well known. One could try to fit it with a wrapper density however as we will show ahead this is practically improbable. An obvious alternative is to choose conjugate priors in such situation. However as we mentioned in section 3.4.4 the joint conjugate prior in the case of unknown mean and precision matrix is a Normal-Wishart-Prior and the joint posterior is a

Normal-Wishart-Posterior. Although one does not use them in practice while using BUGS family of software, the problem of posterior being from a unknown family remains the same. Chib, (1995) suggested using the Rao-Blackwellization method to solve this problem. For e.g. the Rao-Blackwellized estimate of $p(G_k^*|\boldsymbol{y})$ is given by

$$\begin{aligned}\prod_{k=1}^{K} p(G_k^*|\boldsymbol{y}) &= \int \prod_{k=1}^{K} p(G_k^*|\boldsymbol{y},\boldsymbol{b},\boldsymbol{S},\boldsymbol{b}_k^C)p(\boldsymbol{b}_1^C,\boldsymbol{b}_2^C,...,\boldsymbol{b}_K^C,\boldsymbol{b},\boldsymbol{S}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{b}_1^C\,\mathrm{d}\boldsymbol{b}_2^C...\,\mathrm{d}\boldsymbol{b}_K^C\,\mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{S} \\ &\approx \frac{1}{m}\sum_{l=1}^{m}\prod_{k=1}^{K} p(G_k^*|\boldsymbol{y},\boldsymbol{b}^{(l)},\boldsymbol{S}^{(l)},\boldsymbol{b}_k^{C(l)}) \\ &\approx \frac{1}{m}\sum_{l=1}^{m}\prod_{k=1}^{K} f_{\mathcal{W}^{-1}}(G_k^*;n_k^{(l)}+n_0,\Psi+\sum_{i=1}^{n_k^{(l)}}(\boldsymbol{b_i}^{(l)}-\boldsymbol{b}_k^{C(l)})(\boldsymbol{b_i}^{(l)}-\boldsymbol{b}_k^{C(l)})^T)\end{aligned} \tag{4.16}$$

where, $n_k^{(l)}$ are number of subjects classified under component k in iteration l and (n_0,Ψ) are the parameters for the inverse wishart distribution specified as prior for the variance covariance matrix of the component densities. The approximation in 4.16 is done by approximating 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. However in cases when this posterior is not well known, then given that the large number of MCMC iterations one does, it is not possible to manually check and fit wrapper densities to posterior densities. The use kernel density estimation procedures can also be dismissed as they require significant computational power. It is because of these reasons we avoided calculation of Bayes factor in the case where we took indepdent gamma priors for precision of random effects and uniform prior for correlation.

Proceeding further with the Rao-Blackwellization procedure one can obtain the following approximations for the other parameters.

$$\prod_{k=1}^{K} p(\boldsymbol{b}_k^{C*}|G_k^*,\boldsymbol{y}) \approx \frac{1}{m}\sum_{l=1}^{m}\prod_{k=1}^{K} f_N(\boldsymbol{b}_k^{C*};(G_0^{-1}+n_k^{(l)}G_k^{*-1})^{-1}(G_0^{-1}\boldsymbol{\mu}_0+n_k^{(l)}G_k^{*-1}\bar{\boldsymbol{b}_{ik}^{(l)}}),(G_0^{-1}+n_kG_k^{*-1})^{-1}) \tag{4.17}$$

$$p(\sigma^{2^*}|G_k^*,\boldsymbol{b}_k^{C*},\boldsymbol{y}) \approx \frac{1}{m}\sum_{l=1}^{m} f_{Inv-Gamma}(\sigma^{2^*};\alpha_0+\frac{\sum_{i=1}^{n} m_i}{2},\beta_0+\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(y_{ij}-\boldsymbol{x_{ij}}\boldsymbol{\beta}^{(l)}-\boldsymbol{z_{ij}}\boldsymbol{b_i}^{(\boldsymbol{l})})}{2}) \tag{4.18}$$

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

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

where, $(\boldsymbol{\mu}_0,G_0)$ are the parameters for the multivariate normal prior for the mean $\boldsymbol{b}_k^C$ of the k^{th} component density,
$\bar{\boldsymbol{b}_{ik}^{(l)}} = \frac{\sum_{i=1}^{n} I(S_i^{(l)}=k)\boldsymbol{b_i}^{(\boldsymbol{l})}}{n_k^{(l)}}$ is the mean of the estimated random effects corresponding to the n_k subjects classified under the k^{th} component in the l^{th} 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 $\boldsymbol{\eta}$.

Using these values an estimate of $\log p(\boldsymbol{\beta}^*, \sigma^{2^*}, \boldsymbol{\nu}^* | \boldsymbol{y})$ is available which can be further substituted in equation 4.14 to obtain $\log \hat{m}(\boldsymbol{y})$. In ideal cases, i.e. 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 chosen.

4.3 Posterior predictive checks

The idea of the posterior predictive checks is to evaluate the model fit using simulations from the posterior predictive distribution(PPD) $p(\tilde{\boldsymbol{y}}|\boldsymbol{y}$. As an informal check one could sample 1000 values from the PPD 20 times and make 20 histograms to show the density. If the histograms do not match with the histogram of the original sample one could say that the model did not fit the data well.
A formal way to do this is using Posterior predictive p-values(PPP) or Bayesian p-values. In the frequentist paradigm after fitting a model based on parameter $\hat{\theta}$ one could test the model using test statistic. Let us represent that test statistic value for original sample to be $T(\boldsymbol{y})$. Now based on the sampling distribution of $T(\tilde{\boldsymbol{y}})$ we could check the probability $P(T(\tilde{\boldsymbol{y}}) > T(\boldsymbol{y})$. In the bayesian paradigm the parameter θ has a posterior distribution and so we find the same probability like before albeit averaged over the entire posterior $p(\theta|\boldsymbol{y})$. A small PPP value indicates bad fit of model to the data. For a complete interpretation of this p-value we refer the readers to Gelman, (2012).

Chapter 5

Simulation study

In this chapter we will share 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 for extreme values of the wishart random variable. We eventually used mixAK (Komárek, 2015) as the MCMCpack package predicted 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. Simulatenously we made sure that these cow specific random effects were mixture distributed. We will refer to the cows as subjects hereforth.

5.1.1 Description of each data set

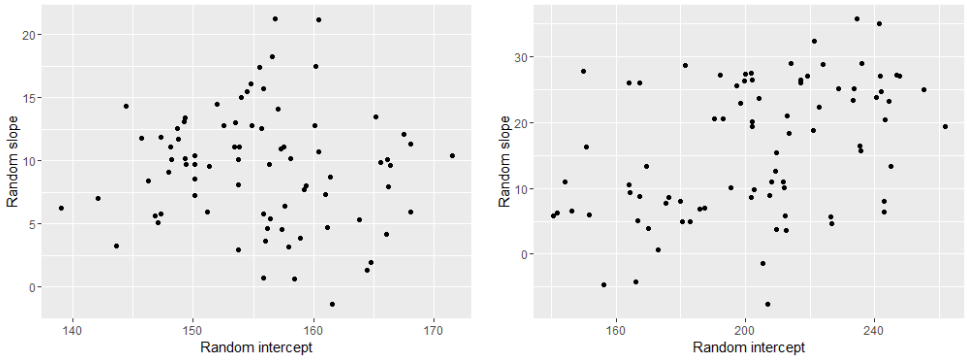
Our aim was to create data sets differing in number of mixture components for random effects, number of subjects, statistical power to detect the fixed effects, separation of mixture components, number of subjects per component. To analyze the efficacy of model selection criteria under these different scenarios we created the multiple data sets. To get a rough idea about the random effects in each of these data sets, we did graphical analysis. For this purpose we first regressed the response $\boldsymbol{y}$ on the 3 predictors age, gender and birth year of cattle using OLS. We then subjectwise regressed the residuals from OLS on the intercept and time of measurement to obtain a rough estimate of the random effect $\tilde{\boldsymbol{b}_i}$ of the i^{th} subject. There are two important aspects of this method. Firstly this estimator overestimates the actual size of the random effects. Secondly, if in the mean structure, one misses out on a covariate other than the one which causes the mixture then it could be virtually impossible to decide on the number of components 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. $\boldsymbol{b}_i \sim N(0, G)$. 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

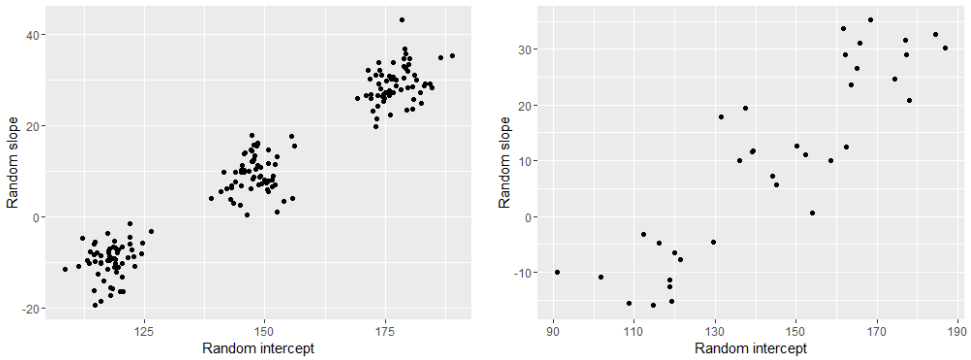
Figure 5.1: Rough estimate $\tilde{\boldsymbol{b}_i}$ for random effects

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 5.2b) then it is likely that they select 1 or 2 components.



(a) Data set 2 (b) Data set 3

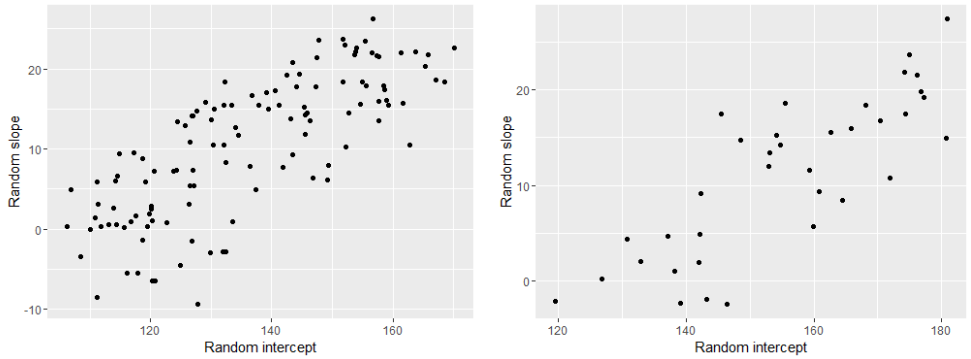
Figure 5.2: Rough estimate $\tilde{\boldsymbol{b}_i}$ for random effects

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. It is important to note that while we generated equal number of components per group earlier, depending upon how many components we model we might still deal with the problem of nearly empty components. The plot of the rough estimates of random effect values for this data set is shown in figure 5.4a.

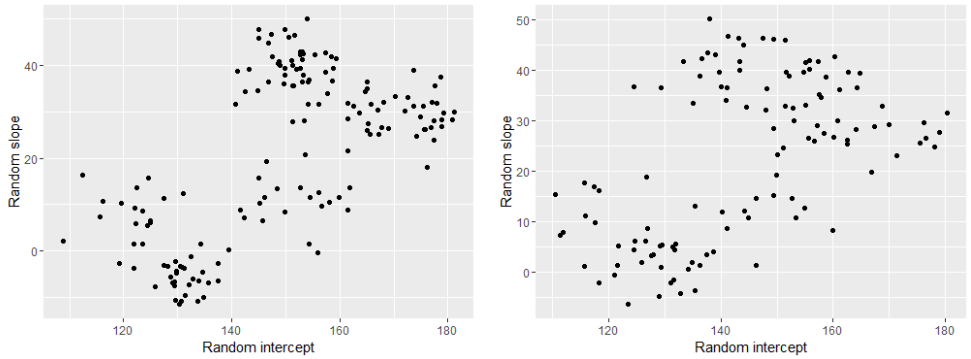


(a) Data set 4 (b) Data set 5

Figure 5.3: Rough estimate $\tilde{b_i}$ for random effects

Data set 7: 5 fused components

his data is similar to Data set 6 in all regards except that the number of subjects per component are lesser, 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

Some of the issues we faced in running MCMC simulations were, label switching across chains, 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 mechanism given in section 3.4.5, it decreased the speed of simulations drastically. To make matters worse we had to use thinning of around 1 per 100 iterations to make sure the resulting chains were not autocorrelated. In light of these reasons and given the time constraints we had to be content with chains of 1300 iterations. Although the following results are based on a single chain and are rounded to the nearest integer, we ran multiple chains and found results to be differing only in the ones's place or by a very small margin.

5.1.3 Deviance information criteria

Table 5.1 shows the values of the various Deviance information criteria (section 4.1) applied to data set 1. We can see DIC_1 gives misleading results whereas the rest of the DIC remain more or less the same for overfitted models. The value of p_{D1} is negative (-602) when we fitted 3 components. Lunn et al., (2012, pg. 161) noted that this can happen if the posterior is multimodal, which as we know can happen due to label switching when models are overfitted. For the same reason Celeux et al., (2006) suggest using DIC_3 instead of the other two observed data DIC measures.

Table 5.1: DIC and p_D for data set 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_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
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 applied to the data set 2. One of the patterns we inferred out of these results was that for DIC_4 the DIC values decrease continually by a very large margin till the right number of components are fitted, whereas for the overfitted models that the DIC either decreases by a relatively smaller margin or remains more or less the same. The behavior of DIC_5 and DIC_3 is similar to DIC_4, however the magnitude difference in DIC that we observe for DIC_4 is not as much for them. These patterns, however do not seem applicable to DIC_1 and DIC_6. For DIC_1 the pattern is that the values of p_{D1} become negative for 4 or more components, i.e. sign of overfitting as we discussed above.

Table 5.3 shows the values of the various DIC applied to the data set 3. Firstly we can see that the pattern we inferred for DIC_1 above is not applicable here. Similar to the previous data set DIC_6 does not reveal any meaningful pattern. We can also see that the pattern of DIC_5 that we noted above is also not applicable here. However for DIC_3 and DIC_4 one can still see that the pattern of DIC decreasing by big magnitude till the right number of components are fitted is still

Table 5.2: DIC and p_D for data set 2.

# 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_{D1}	p_{D2}	p_{D3}	p_{D4}	p_{D5}	p_{D6}
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

valid. The magnitudes of the DIC have decreased as the sample size for this data set was only 36 subjects compared to the 180 subjects in data set 2.

Table 5.3: DIC and p_D for data set 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 shows the results of applying various DIC to data set 4. So far we have observed that DIC_3 and DIC_4 can be used to detect the number of components. Since the components in this data set are fused, and the subjects count is moderately high the results following are interesting to analyze. Firstly, we can see that DIC_4 still follows the pattern we have discussed so far, but with DIC_5 it is a bit difficult to justify. i.e. even if the components are fused, DIC_4 may work well.

To see the patterns in further light, we decided to decrease the number of subjects to 36. Table 5.5 shows the corresponding results of DIC. At first glance one can see that the pattern we saw so far for DIC_4 doesn't exist anymore. However, the catch here is that these results are with a dirichlet prior $Dir(1, 1, ..., 1)$ for the weight distribution. When we fitted 2 or more components we found that the MCMC chains had not converged with this prior. Given the fused data set, even

Table 5.4: DIC and p_D for data set 4

# 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

with fitting 2 components we risk unidentifiability due to empty components. Thus we changed the prior to Dir$(3, 3, ..., 3)$ and fitted for 2 and 3 components respectively. With 2 components we found DIC_4 = 2441 and p_D = 461. Whereas DIC_3 was 1937 with p_D = 14. Further fitting with 3 components we obtained DIC_4 = 2352 and DIC_4 = 2346 respective. In light of these results one can still justify the pattern we have observed for DIC_4 so far. An interesting result from this exercise was that using only $Dir(1, 1, ..., 1)$ can lead to severe underfitting.

Table 5.5: DIC and p_D for data set 5.

# 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

Table

5.1.4 Marginal likelihood

We implemented Chib's approximation mentioned in section 4.2. Table 5.8 shows the results of $\log \hat{m}(\boldsymbol{y})$ for the various data sets and the models fitted for them. One can see that there is no obvious pattern visible in these results to conclude the efficacy of Bayes factor in selection of

Table 5.6: DIC and p_D for data set 6

# 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_{D_1}	p_{D_2}	p_{D_3}	p_{D_4}	p_{D_5}	p_{D_6}
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

a model. We had observed so far that given our data sets, the results of overfitting was label switching. However fitting 1 and 2 components did not have label switching and the chains had good convergence as well. However even if we take the cases where we had large number of observations and the components were well separated such as data set 6, we can see marginal likelihood prefers fitting 1 component over 2. Also the results of marginal likelihood for 3, 4 and 5 components for data set 6, are more or less the same. It is important to note that there is a higher margin of error in these results as the MCMC iterations were done multiple times for each data set and each component. However as we have already shown even for data sets where these chances were very less, marginal likelihood doesn't help choosing the right model.

Table 5.7: My caption

# 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

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

Chapter 6

Analysis of blood donor data set

Write something here

Chapter 7

Conclusion

Write something here

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.

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 (2012). *Understanding posterior p-values*.

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.

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.

Lunn, David et al. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. English. 1 edition. Boca Raton, FL: Chapman and Hall/CRC.

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

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.

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.

Leuven Statistics Research Centre (LStat)
Celestijnenlaan 200 B bus 5307
3001 HEVERLEE, BELGIË
tel. + 32 16 32 88 75
fax + 32 16 32 28 31
www.kuleuven.be