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

Preface

The following thesis work was conducted as part of the programme completion requirements of MSc. Statistics programme at KU Leuven. 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 Deviance information criteria definitions, marginal likelihood, choice of posteerior predictive checks, implementating them in a software, Gibbs sampler idiosyncrasies and lack of computational power required for this work. However, at every step the work became more and more enticing. Looking backwards, I think it was one of the most interesting project I have done in recent times. 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 found at
`https://github.com/anirudhtomer/MScThesis`

In chapter 1 we present an introduction to mixture distribution and their central role in the formulation of the problem statement for this thesis. In Chapter 2 we present an introduction to the Bayesian paradigm as for the work of this thesis we use Bayesian methods. Futher in Chapter 3 we present the definition of a Bayesian heterogeneity model, and the issues with estimation of parameters in it. In Chapter 4 we present the formulae for various classes of Deviance information crieteria, marginal likelihood and posterior predictive checks that we used for 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 we model the Blood donor data set (Nasserinejad et al., 2015) using a Bayesian heterogeneity model and use results from the simulation study to apply the right model selection criteria.

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 available to discuss the difficult parts of the work at hand. 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. They introduced me to mixed models and empowered me with the tools of trade required to do the frequentist analysis of blood donor data set in this report. 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 their 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 (Frühwirth-Schnatter, 2013, pg. 4). 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 however 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 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

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

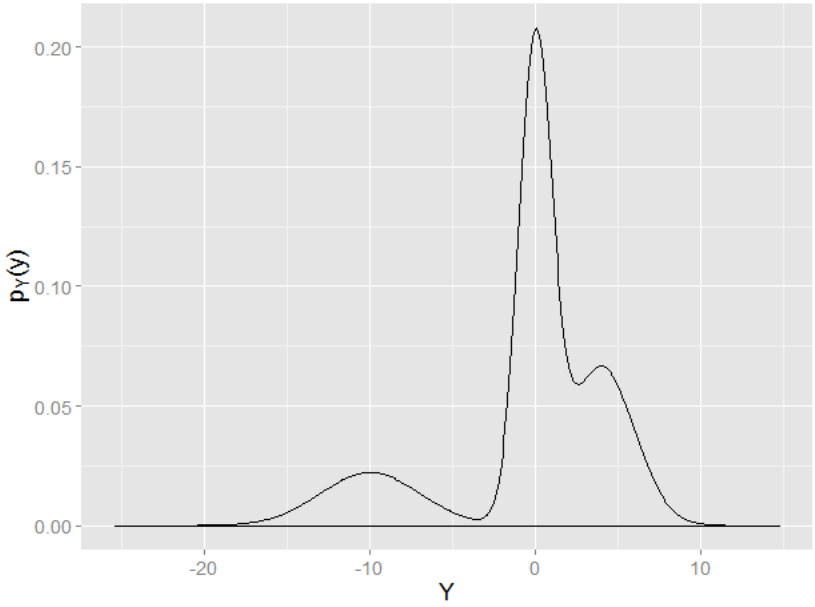


Figure 1.1: Mixture density of $Y \sim \frac{1}{6}N(-10,3) + \frac{1}{2}N(0,1) + \frac{1}{3}N(4,2)$

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

One of the biggest challenges 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 can 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 more 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 like 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 and use it as a biometric identification mechanism.
- Finance: Brigo and Mercurio, (2002) propose 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 model to adjust for a hidden attribute in the data which could not be collected. 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. This particular linear mixed effects model 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. However in both types of models one has to tackle the issues described in section 1.1.2. 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 for selecting the right number of mixture components for random effects distribution. Since we will be following the Bayesian paradigm, 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

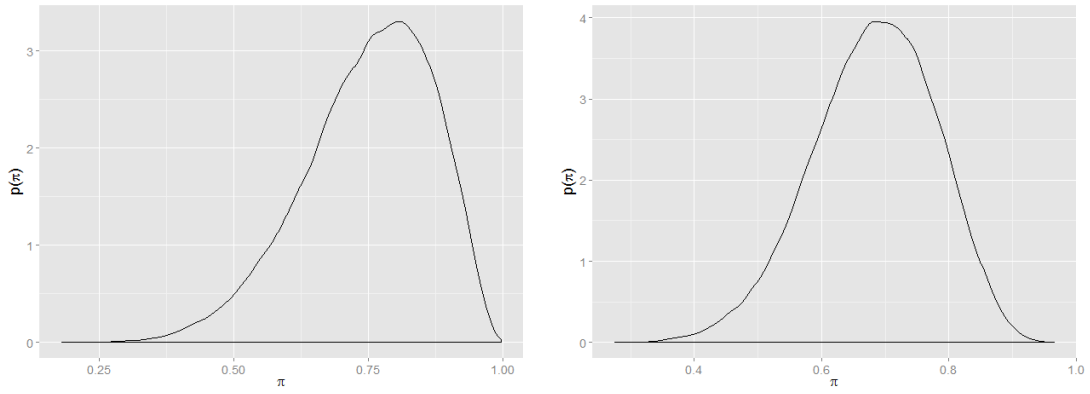
Bayesian paradigm

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 signigy the ideological difference 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 refree 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.

While subjective, these beliefs 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. For e.g. in figure 2.1a we can see one such prior distribution corresponding to the belief that the chance of getting a heads on toss is more than tails and it is more likely to be somewhere between 70 to 90%. This is in contrast to the frequentist paradigm where parameters do not



(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.

have a distribution but are rather constants.

2.2 Bayes rule

We will now present the Bayes rule which provides the framework to estimate the entire distribution of π based on the data and the prior beliefs. The Bayes rule for the continuous parameter π is given by

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

The result $p(\pi|\boldsymbol{y})$ is called the posterior distribution of the parameter based on which statistical inference about the parameter can be done. An intuitive way to get the motivation behind the Bayes rule is that the denominator can be seen as marginal probability of $\boldsymbol{y}$ based on 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 heads. Assuming that the tosses were independent, then based on the likelihood function $L(\pi|\boldsymbol{y})$ the MLE of π will be $\hat{\pi} = 0.6$. Whereas Bayes rule gives us the posterior distribution of parameter (π) as shown in figure 2.1b. The mean value of this 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 family and thus available in closed form. Such priors are termed conjugate priors. However this is not always the case; For e.g. if the prior belief for π 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 like Gibbs sampling, Metropolis hastings, 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. In absence of prior knowledge a non informative prior should be chosen. 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 are interpreted as the interval in which 95 out of 100 times one can find the population parameter θ, the Bayesian 95% credible interval can be 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 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 use 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. The specialty of the 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. Since the observations collected from a subject will be correlated a linear model will not be useful because of the restrictions it imposes on the covariance structure.

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 (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 in matrices G and R_i. Thus the uncertainty in estimation is ignored. On the other hand the bayesian approach averages out 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 for the model presented in section 3.1.1 we will have 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 of it 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 it is quite easy 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 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.

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 of Gaussian components for random effects distribution. Such a linear mixed model is termed as a Heterogeneity model.

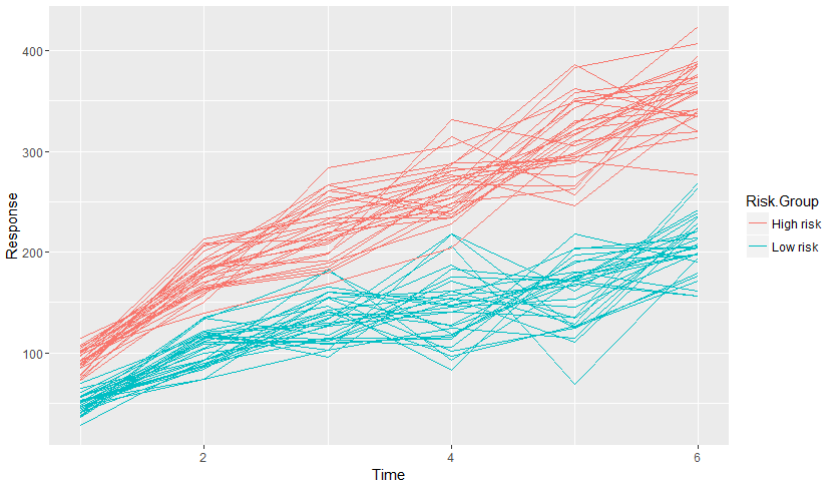


Figure 3.1: Individual profiles of 30 subjects from each group.

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 all of the subjects. Since we are following the bayesian paradigm, in addition to prior distribution for β and σ^2 we also have prior for $\boldsymbol{\nu} = (\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, \ldots, \boldsymbol{b}_K^C, \boldsymbol{\eta}, G_1, G_2, \ldots, G_K)$.

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 alloation vector S_i for each subject. Then conditional on knowing $S_i = k$ the following LMM equation has a hierarchical interpretation.

$$\boldsymbol{y}_i|\boldsymbol{b}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \boldsymbol{\varepsilon}_i), \boldsymbol{\varepsilon}_i \sim N_{m_i}(\boldsymbol{0}, R_i)$$

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

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_k^C, \boldsymbol{\varepsilon}_i^*), \boldsymbol{\varepsilon}_i \sim N_{m_i}(\boldsymbol{0}, \boldsymbol{Z}_i G_k \boldsymbol{Z}_i^T + R_i)$$

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 to for each iteration. Secondly it 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 had as much autocorrelated estimates as the marginal model and took less time for each iteration.

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 like time 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 time. To enforce the mean $\boldsymbol{0}$ on the random effects in a mixture distribution the following condition should be satisfied.

$$E(\boldsymbol{b}_i|\boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k N_q(\boldsymbol{b}_k^C, G_k) = 0 \tag{3.2}$$

where ϕ is the vector $(\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, \ldots, \boldsymbol{b}_K^C, \boldsymbol{\eta}, G_1, G_2, \ldots, G_K)$. This further means that $E(\boldsymbol{y}_i|\boldsymbol{\phi}) = \boldsymbol{X}_i\beta$. This parametrization, which was also used in the original paper on Heterogeneity model (Verbeke and Lesaffre, 1996) is called 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 while estimating parameters using MCMC. While doing the simulation study we figured out that imposing the constraint in equation 3.2 drastically slows the convergence as well increases the autocorrelation in parameter estimates. Thus, in this thesis we have only used hierarchically centred parametrization.

3.4.3 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 components densities in the mixture distribution of random effects. We know that both (b_k^C) and (G_k) are unknown, and hence to obtain the joint posterior in closed form one has to specify the conditionally conjugate prior $b_k^C|G_k \sim N(\boldsymbol{\mu}_0, \frac{G_K}{N_0})$ and $G_k^{-1} \sim \mathcal{W}(n_0, \Psi)$. The marginal prior of b_k^C is a multivariate T distribution. However the problem with this approach is that it is firstly difficult to implement in JAGS 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 mean and covariance matrix (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 random intercept mean and random slope mean.

Choice of prior for covariance matrix

The choice of prior for the covariance matrix is an interesting one. 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 parameter at hand. For e.g $IW(\begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix}, 2)$ could be one such prior. For precision matrix one can use the Wishart prior $W(\begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}, 2)$. i.e. the scale 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 influence the posterior more than the likelihood does.

One can also choose indepdent gamma priors for random intercept and random slope and uniform prior $U(-1, 1)$ for 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 slow as the posterior is not available as a known density. Another benefit of this approach, as we later found out during simulations is that when more components than needed are fitted to the data, 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 $\boldsymbol{b}_1^C, \boldsymbol{b}_2^C, ..., \boldsymbol{b}_K^C, G_1, G_2, ..., G_K$. The problem of choosing a conjugate prior such that the joint posterior of (β, σ^2) is available as a known density, is similar to the problem we discussed in previous two subsections. It is quite common in practice to use indepdent univariate normal priors such as $N(0, 10000)$ for each of the β_p and a Gamma$(0.0001, 0.0001)$ prior for $\tau = \frac{1}{\sigma^2}$ (Gelman and Hill, 2006, chap. 17).

3.4.4 Label Switching

3.5 Likelihood: Complete data vs Mixture

The likelihood function in a mixture distribution depends on how much we know about the data. If we know both the data and the allocation vector $\boldsymbol{S}$ then the following likelihood function of parameters is called a complete data likelihood function.

$$p(\boldsymbol{y}, \boldsymbol{S}|\boldsymbol{\nu}) = \prod_{i=1}^{N} \prod_{k=1}^{K} (p(\boldsymbol{y}_i|\boldsymbol{\theta}_k)\eta_k)^{I_{S_i=k}}$$

where $\boldsymbol{\nu} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K, \boldsymbol{\eta})$ is a vector of weight distribution and parameters of component densities. It is possible to write this complete data likelihood function in ${}^K P_K = K!$ equivalent ways by permuting the order of components. Each order of components is called a Labeling scheme. Although this idea seems trivial we will see ahead that it creates a problem during estimation called label switching. While this likelihood function is valid conditional on knowing the allocation vector $\boldsymbol{S}$, the following likelihood function called Mixture likelihood function applies when we are not aware of the allocations.

$$p(\boldsymbol{y}|\boldsymbol{\nu}) = \prod_{i=1}^{N} (\sum_{k=1}^{K} p(\boldsymbol{y}_i|\boldsymbol{\theta}_k)\eta_k)$$

It is interesting to note that the mixture likelihood function is symmetrical and has $K!$ modes. We refer readers to Frühwirth-Schnatter, (2013, pg. 45-46) for the review of geometric presentation of this likelihood function.

3.6 Mixture model identifiability: Label switching

After running a MCMC procedure to estimate the posterior distributions of the parameters involved, we will be interested in knowing the parameters for the component densities. In most cases we will also be interested in classification of observation using allocation probabilities $P(S_i = k|\boldsymbol{y})$. Now let us imagine that we fitted exactly the true number of components K^{true} from which the mixture density was formed. At this point it is possible that the posterior densities of parameters do not reflect the true posterior distribution due to label switching.

To idea of label switching could be explained with this simple 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 sampled a few observations from it. The MCMC procedure we will estimate parameters using data augmentation. i.e. we begin with some random allocation vector $\boldsymbol{S}^0$ and estimate parameters using complete data likelihood. For MCMC labels μ_1 and μ_2 exist rather than μ_{C1} and μ_{C2} and it does not associate labels with actual components. We begin with a vague joint prior for these parameters $p(\mu_1, \mu_2) = p(\mu_1) * p(\mu_2) = p(\mu_1) * p(\mu_1) = p(\mu_2) * p(\mu_2)$.

Assume that the allocation vector we began with assigns all observations from component C_1 to label 1 and all observations from component C_2 to label 2. Under such a scheme $(\mu_1, \mu_2) = (5, 7)$ is likely. However if we take a conjugate of this allocation vector $(\mu_1, \mu_2) = (7, 5)$ will also be accepted. This because we have a mixture likelihood function which is bimodal. Now let us imagine a scenario where because of our initial allocation vector, 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 we estimate allocation vector conditional on these estimates in MCMC. Supposing that an observation with value 6.5 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 situations even $(\mu_1, \mu_2) = (6.5, 5.5)$ could be sampled by MCMC. This because under the mixture likelihood it is also likely. However this scenario could've been unlikely if the true means were very far apart. In our scenario the issue is that posterior for μ_1 will have a multiple modes. Not only that but if the sampler kept on arbitrarily switching between the two equivalent posterior regions then both regions will be partially explored. Thus any inference based on this posterior will be useless. Frühwirth-Schnatter, (2013, pg. 82) suggest to use a

balanced label switching, which gives multimodal posterior albeit with a full exploration.

It is interesting that if our prior for the parameters was not vague, but exactly equal to the true distribution of parameters then label switching might not have happened. However the problem with a strong prior is that an incorrect strong prior could also inadvertently cause label switching or it might not allow a complete exploration of the posterior. Other techniques to stop label switching are imposing an identifiability constraint, which in our case was $\mu_1 < \mu_2$. However in higher dimensions it could become difficult to find a constraint which imposes a unique labeling scheme. For more details we refer the reader to Stephens, (2000).

3.7 Mixture model identifiability: Equal or empty components... DIrichlet prior

A mixture model will also be unidentified if we have an empty component or two components with the same parameters. Suppose the true number of components is K^{true} and we fit $K = K^{true}+1$ components. In the MCMC sampler suppose one of the components is assigned any observation. Thus the posterior for the parameters will remain the same as the prior. Assuming that the prior for weight distribution $\boldsymbol{\eta}$was a Dirichlet prior $\mathcal{D}(0.5, 0.5, \ldots, 0.5)$, then we will have a posterior for $\boldsymbol{\eta}$ such that the K^{th} component will always have an almost 0 weight in the mixture distribution. This means that at the end of MCMC sampling the component density's posterior will be same as its prior and no observations will be allocated to the component. In such cases the true number of components could be estimated by the count of components with non zero number of allocations. However this situation could be avoided with a stronger prior on the weight distribution which forces the posterior distribution of $\boldsymbol{\eta}$ to be such that no components are empty at the end of MCMC sampling. Identification of number of components in such case is explained in the next section.

Gosh! I should explain it in terms of pulling away the posterior eta from the boundary where components of eta are linearly dependent..

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 $f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\nu})$, 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.

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

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

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.

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

where $\bar{\boldsymbol{\beta}}$, $\bar{\sigma^2}$ 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. $\text{E}_{\boldsymbol{b},\boldsymbol{S}}(\log p(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{S}|\bar{\boldsymbol{\beta}}, \bar{\sigma^2}, \bar{\boldsymbol{\nu}}))$, 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 $\bar{\boldsymbol{\beta}}$, $\bar{\sigma^2}$ 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 $\bar{\boldsymbol{\beta}}^{(l)}$, $\bar{\sigma^2}^{(l)}$ and $\bar{\boldsymbol{\nu}}^{(l)}$.

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

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

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

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

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

The next definition of DIC under the class of complete data DIC's 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 using MAP(maximum a posteriori) as the estimate instead, the following definition of DIC is proposed.

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

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 $\boldsymbol{\theta}_{\text{cond}} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\nu}, \boldsymbol{S}, \boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_n)$. This leads to the conditional data likelihood,

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

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

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

where $\hat{\boldsymbol{\theta}}_{\mathsf{cond}} = \arg\max_{\boldsymbol{\theta}_{\mathsf{cond}}} p(\boldsymbol{\theta}_{\mathsf{cond}}|\boldsymbol{y})$, and $\mathsf{E}_{\boldsymbol{\theta}_{\mathsf{cond}}|\boldsymbol{y}}(\log p(\boldsymbol{y}|[\boldsymbol{\theta}_{\mathsf{cond}})|\boldsymbol{y}])$ 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 3.4.3. 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 3.4.3 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

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.

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

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.

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.

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.

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