# Master Thesis Progress Report
# Literature Review

Thesis title: The use of mixture distributions in a Bayesian linear mixed effects model.

Name of student: Anirudh Tomer

Name of promoter: Professor Emmanuel Lesaffre

## Introduction

In this literature review I will present a brief overview of the concepts I studied in the past few months to proceed in my thesis work. I first began with the material in Lesaffre and Lawson (2012) to understand the concepts for modelling Bayesian linear mixed effects model (BLMM). I then proceeded to the material in Frühwirth-Schnatter (2013) to learn about Mixture models and estimation of parameters in mixture models using Bayesian techniques. The other material I studied were publications related to Bayesian heterogeneity model. All of the references that I have used in my literature survey are included in the References section of this document. I also started writing a thesis document, and it contains a detailed overview of the material that I have studied so far. It can be found at this link:

https://github.com/anirudhtomer/MScThesis

## Mixture distributions

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 twostep process in which firstly a particular random variable (also called component) 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. Mixtures 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.

Formally, 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:

$$p(y) = \Sigma_{i=1}^{K} \eta_i p_i(y)$$

The vector of weights $\boldsymbol{\eta} = \eta_1, \eta_2, \ldots, \eta_K$,is called the weight distribution. The $k^{th}$ weight $\eta_k$ 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 $R^{+K}$ with an additional constraint $\Sigma_{i=1}^{K} \eta_i = 1$.

It is possible to have mixture components from different parametric families as well (Frühwirth-Schnatter, 2013, pg. 4).

### Challenges

One of the biggest challenges while modelling a mixture density for an observed random variable is that the number of mixture components ($K$), weight distribution $\boldsymbol{\eta}$ and the corresponding parameters for component densities might not be known in advance. Another issue is that 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 analogous to the clustering problem. There are other issues that I will have to deal with, one particular being the Label switching problem, and the other being the choice of priors. However for sake of brevity I am not including them in this document.

### Applications of mixture distributions

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

- Spike sorting of neural data: Both GMM and mixture of multivariate t-distributions have been used. (Lewicki, 1994; Shoham, Fellows, and Normann, 2003). I also visited researchers at NERF-IMEC in Heverlee and personally saw them using GMM for spike sorting.
- 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 I 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.

## Goals of this 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 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 challenges described above. The aim of this master thesis is to evaluate existing Bayesian approaches such as deviance information criterion (DIC), marginal likelihood, posterior predictive checks etc. for selecting the right number of mixture components for random effects distribution. Since I will be following the Bayesian paradigm, I will use MCMC methods instead of the frequentist point estimation methods. While I will use a longitudinal data set (mostly from Erasmus Medical Center in Rotterdam) to fit Bayesian linear mixed effect model, I will also simulate data sets to compare various approaches for choosing the number of components.

# Bayesian heterogeneity model

The Bayesian heterogeneity model is an extension of the Bayesian linear mixed model (BLMM), in the sense that BLMM assumes that the random effects have a joint multivariate normal distribution. The heterogeneity model however allows having a mixture of random effects. This is useful in many medical studies where patients cannot be categorized into certain categories in advance as that information becomes available with time. This heterogeneity which should've been modelled with fixed effects, can be accounted for, using a mixture for random effects. Formally a Bayesian heterogeneity model, considering the $i^{th}$subject among $n$ subjects is given by:

$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b}_i + \boldsymbol{\varepsilon_i}$ , where $1 \leq i \leq n$,

$\boldsymbol{y_i} = \left(y_{i1}, y_{i2}, \dots, y_{im_i}\right)^T$ is a vector of observations for the $i^{th}$ subject taken at $m_i$ time points,

$Xi = \left(x_{i1}^T, x_{i2}^T, \dots, x_{im_i}^T\right)^T$ is the $m_i \times (d + 1)$ design matrix for the $i^{th}$ subject,

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T$ is a $(d + 1) \times 1$ vector of fixed effects with $\beta_0$ being the intercept,

$\boldsymbol{Z_i} = \left(z_{i1}^T, z_{i2}^T, \dots, z_{im_i}^T\right)^T$ is the $m_i \times q$ design matrix of covariates varying for a subject at each observation,

$\boldsymbol{b_i} = \left(b_{0i}, b_{1i}, \dots, b_{(q-1)i}\right)^T$ is a $q \times 1$ vector of random effects with $b_{0i}$ being the random intercept. The random effects $\boldsymbol{b_i} \sim \Sigma_{i=1}^{K} \eta_i N_q(\boldsymbol{b}_k^C, G_k)$, with $b_k^C$ and $G_k$ being the mean vector and covariance matrices for the $k^{th}$ component in the mixture distribution respectively,

$\varepsilon_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{im_i}\right)^T$ is a $m_i \times 1$ vector of measurement errors. The errors $\varepsilon_i \sim N_{m_i}(0, R_i)$ with the covariance matrix of errors $R_i$ usually being $\sigma^2 I_{m_i}$.
The parameters $\boldsymbol{\beta}, \boldsymbol{\sigma^2}, \boldsymbol{b_1^C}, \boldsymbol{b_2^C}, \dots, \boldsymbol{b_k^C}, G_1, G_2, \dots G_k$ all have a prior distribution, the choice of which is not discussed in this document for brevity.

## Bayesian software

The software tools I will use are from the BUGS family like JAGS or WinBUGS. While WinBUGS provides its own integrated development environment it definitely lacks the usability and visualization capabilities offered in R. JAGS on the other hand relies on third party tools completely for visualization and analysis of MCMC chains. There are R packages namely R2jags, R2WinBUGS which allow users to execute JAGS/WinBUGS code via R. The R package coda provides a rich array of functions to do analysis and diagnosis of MCMC chains. For Bayesian linear mixed models the R package 'blme' will be used. However it seems I will have to modify it a bit to allow for heterogeneity model. I have contacted the author of the package Vincent Dorie in this regard. For Bayesian mixture models I will evaluate the R package 'bayesmix' and see if I can merge is with 'blme'.

## Plan ahead

I have already contacted the author of 'blme' package, Vincent Dorie, so that I can modify his code and merge 'bayesmix' package in it. Once that is done the next obvious step will be to model the longitudinal dataset from Erasmus MC using Bayesian heterogeneity model. I am thinking of completing both of these tasks before the midterm presentation in March 2016.

## References

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.

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.

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

Johannes Berkhof, Iven Van Mechelen (2003). "A Bayesian approach to the selection and testing of mixture models." In: Statistica Sinica 13.2, pp. 423–442.

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.

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.

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.

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.

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.