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

Write something in todo

Summary

Summary yet to be done. Following is copied from Professor Lesaffre's proposal document for sake of not leaving this section empty.

In this master thesis we wish to explore Bayesian methods to model finite mixture random effects distributions in a Bayesian linear mixed effects model. By assuming that the random effects are a finite mixture of normal distributions, we can account for random effects that are not normally distributed. We wish to address two problems: finding the correct number of mixture components and checking the fit of the mixture distribution. The master thesis involves fitting finite mixture linear mixed models to real longitudinal data, such as blood donor data. The proposed approaches for choosing the number of components in the random effects distribution (e.g. marginal likelihood, posterior predictive checks, DIC, etc) will be evaluating using simulation studies. The analyses will programmed in WinBUGS or JAGS, but also in combination with R.

Contents

Chapter 1

Introduction

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. In our example all the mixture components were having the same parametric family i.e. Normal distribution, but 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).

1.1.1 Formal definition for finite mixture distribution

Mention that we follow notation from (Frühwirth-Schnatter, 2013) ???

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

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

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,

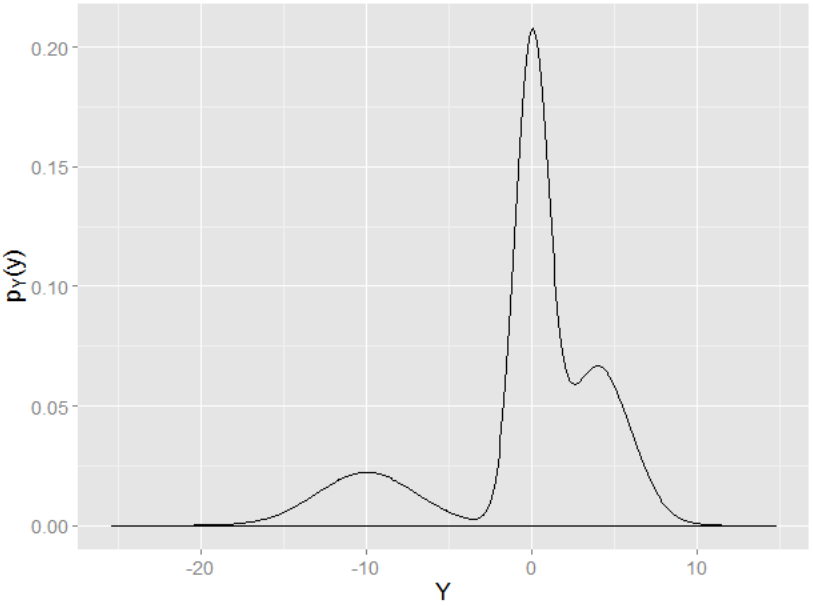


Figure 1.1: Mixture density of $Y \sim \frac{1}{6}N(-10,3) + \frac{1}{2}N(0,1) + \frac{1}{3}N(4,2)$

$$\sum_{i=1}^{K} \eta_i = 1$$

1.1.2 Challenges

One of the biggest challenges while modeling 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 in fact solving the clustering problem, albeit using parametric methods in our case. Rephrase the last line

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)

I have personally seen NERF researchers using GMM for spike sorting in their lab at IMEC during a visit

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 (see section 3.1 for 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 (see chapter 3). 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 approaches name some approaches like marginal likelihood etc. for solving these issues. Since we will be following the bayesian paradigm, we will use MCMC methods instead of the frequentist point estimation methods. While we will use the longitudinal data set [name of the data set here] to fit bayesian linear mixed effect model, we will also simulate data sets to compare various approaches for choosing the number of components.

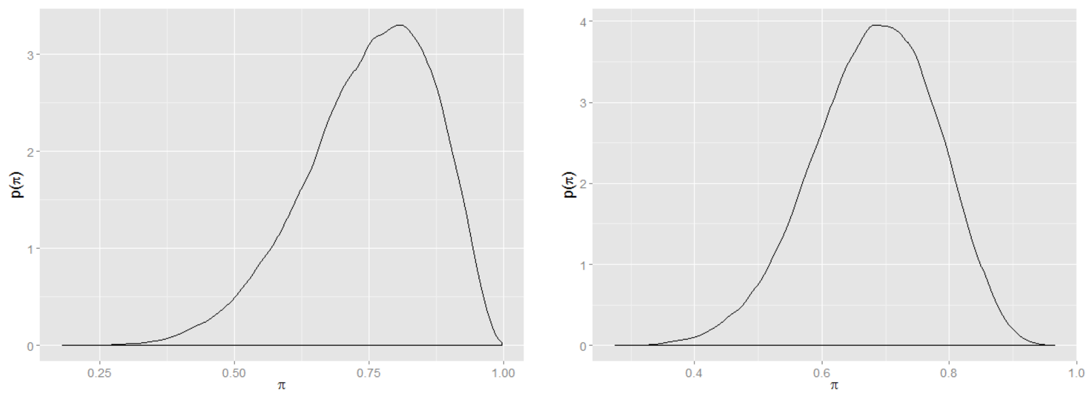
Chapter 2

Bayesian paradigm

2.1 The bayesian motivation: A toy example

To explain the motivation behind the bayesian paradigm, we will use an informal approach via the following toy example. Suppose there are three people A,B and C of whom A and B each are captains of a cricket team and C is the refree who tosses the coin. Given the importance of the toss in this sport each side would like to win the toss. 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 toy problem 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 have a distribution but are rather constants. Also the point and interval estimation in frequentist paradigm do not take prior beliefs into account and rely completely on the data at hand. While a detailed discussion of Bayesian vs. Frequentist approaches can be found in Lesaffre and Lawson (2012), we will present a brief overview of Bayes theorem and its usage in bayesian parameter estimation.



(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.

2.2 Bayes theorem

To understand the use of Bayes theorem in parameter estimation let us first look at one of the frequentist approach Maximum likelihood estimation to estimate the probability of getting heads (π). Suppose after 10 matches captain B observed that 6 times out 10 the toss resulted in heads. Assuming that conditions in each toss were such that the tosses were independent, then based on the likelihood function $L(\pi|\boldsymbol{y})$ the MLE of π will be $\hat{\pi} = 0.6$. In contrast, the Bayes rule 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 theorem 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. In figure 2.1b we can see the posterior PDF for the probability of getting heads that we obtained after applying Bayes rule. The mean value of this distribution is 0.7 which if you compare with the MLE of 0.6 you can see that bayesian posterior mean is influenced by the prior as well.

The intuition part is needed to be expanded upon, and shall I talk of CI at this moment???

2.3 Bayesian software

We can see in equation 2.1 that the computation of posterior involves solving the integral in the denominator. While the calculations are simple in certain cases, for e.g. in exponential family the choice of conjugate prior leads to a posterior belonging to the same family. However it also depends on the prior beliefs. For e.g. if the prior belief for π in our toy example is that it is trimodal then we may 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, Slice sampling, Metropolis hastings and their variants etc.

I am not sure which ones I will end up using. Pretty confused what to write in this section. I've only used JAGS and WinBugs via R (hated WinBugs user interface). Tried SAS proc mcmc recently, can't comment much on it.

The software tools we 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. We will also compare the results from the Bayesian procedures present in SAS.

For a detailed overview of the Bayesian software and their procedures Lesaffre and Lawson (2012, Chap. 8) can be referred. SAS also has an online manual with examples available on their website *Bayesian Analysis Using SAS/STAT Software*.

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 with the help of 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

$\boldsymbol{y_i} = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$, where

$1 \le i \le n$,
$\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T$ is a vector of observations for the 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 varying for a subject at each observation,
$\boldsymbol{b_i} = (b_{0i}, b_{1i}, \ldots, b_{(q-1)i})^T$ 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. For this and an in depth coverage of LMM we refer the reader to Verbeke and Molenberghs (2009).

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:

- $\sigma^2 \sim p(\sigma^2)$
- $\boldsymbol{\beta} \sim p(\boldsymbol{\beta})$
- $G \sim p(G)$

talk about prior heterogeneity and mixture talk about checks for model fit

Chapter 4

Data set

Write something here

Chapter 5

Analysis of data

Write something here

Chapter 6

Conclusion

Write something here

Bibliography

Bayesian Analysis Using SAS/STAT Software. SAS support.

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.

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.

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.

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. 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.

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