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. 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.

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 $p_1(y), p_2(y), ..., p_K(y)$ and weights $\eta_1, \eta_2, ..., \eta_K$, 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 $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_K)$ 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,

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

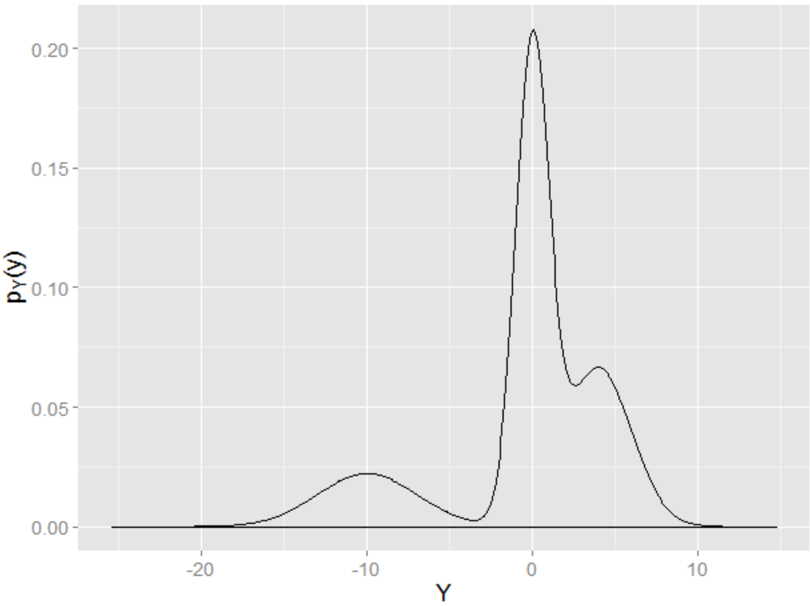

Figure 1.1: Mixture density of $Y \sim \frac{1}{6}N(-10,3) + \frac{1}{2}N(0,1) + \frac{1}{3}N(4,2)$

1.1.2 Issues

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 $y_1, y_2, ..., y_N$ sampled from the mixture density $p(y)$ one may not know which observation belongs to which component density. Formally, an allocation vector $\boldsymbol{S} = (S_1, S_2, ...S_N)$ 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

Put some applications here. Section 6.1.4 and section 7.4.5

1.2 Goal of master thesis

citations for LMM and BLMM to verbeke et. al and lesaffre et al.

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). more on this in section..refer. This 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 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 2.1 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 the prior probability of getting a heads is a constant. However one can also model the frequentist scenario in bayesian paradigm by having all the probability mass at one single point (e.g. Direc delta distribution).

talk about posterio and then introduce to the 2. bayes theorem, 3. bayesian software and MCMC methods

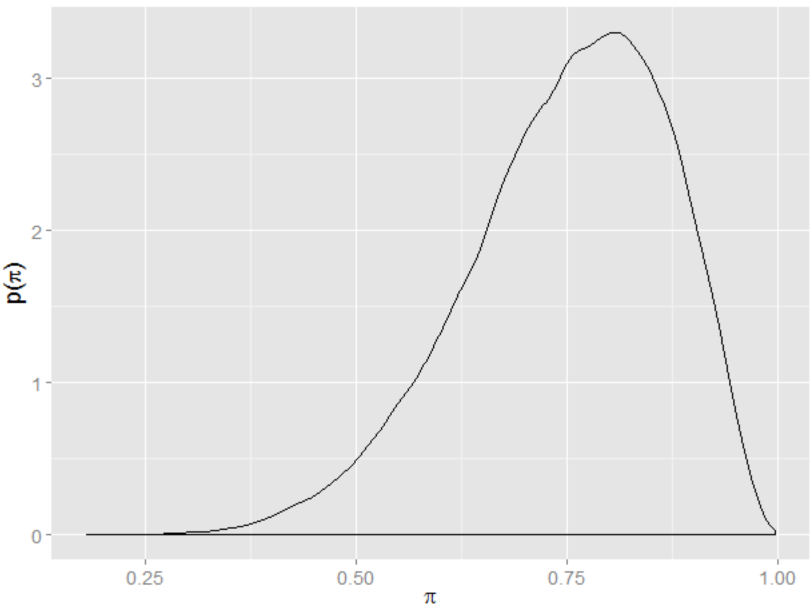

Figure 2.1: PDF of the probability of getting heads on coin toss (π)

Chapter 3

Mixture models

Write something here

Chapter 4

Data set

Write something here

Chapter 5

Analysis of data

Write something here

Chapter 6

Conclusion

Write something here

Bibliography

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

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.

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