Mixture models applied to heterogeneous populations

Carolina Valani Cavalcante*
Escola Nacional de Ciências Estatísticas, Rio de Janeiro, Brazil - carolinavalanic@gmail.com

Kelly Cristina Mota Gonçalves
Universidade Federal Fluminense, Niterói, Brazil - kelly@est.uff.br

Abstract

Mixture models are useful to solve problems that involve observed phenomena in heterogeneous populations, this is, populations composed of latent subpopulations. Bayesian approaches have attracted great interest among researchers. In particular, Markov Chain Monte Carlo (MCMC) methods provide a current way to draw inference from these models. However, when the number of subpopulations is considered unknown, more sophisticated methods are required to perform the Bayesian analysis. The reversible jump MCMC (RJMCMC) is an alternative method in this case. The aim of this work is to analyze the fit of a mixture model under different settings of heterogeneity, sample sizes and estimating the number of subpopulations or not. A Normal mixture model is evaluated using simulated and real datasets.

Keywords: subpopulations; Bayesian Inference; RJMCMC.

1. Introduction

Mixture models are noted for their flexibility and are widely used in the statistical literature (see McLachlan and Peel (2004)). Such models provide a natural framework for the modeling of heterogeneity in a population. Moreover, due to the large class of functions that can be approximated by a mixture model, they are attractive for describing non-standard distributions. They have been adopted in many areas as genetics, ecology, computer science, economics, biostatistics and many others. For instance, as stated in Jordan (2004), in genetics, location of quantitative traits on a chromosome and interpretation of microarrays both relate to mixtures, while, in computer science, spam filters and web context analysis start from a mixture assumption to distinguish spams from regular emails and group pages by topic, respectively.

Statistical analysis of mixtures has not been straightforward and the Bayesian paradigm has been particularly suited to their analysis. This framework allows the complicated structure of a mixture model to be decomposed into a set of simpler structures through the use of hidden or latent variables. According to Richardson and Green (1997), when the number of components is unknown, the Bayesian paradigm is the only sensible approach to its estimation. Then, the Bayesian approach contributed to mixture models become increasingly popular in many areas.

When the number of subpopulations is assumed known, MCMC methods can be used for Bayesian estimation of the subpopulation parameters. On the other hand, when the number of components is considered unknown, there is problem with variable dimension, and more sophisticated methods are required to perform the Bayesian analysis. One method is the Reversible Jump MCMC (RJMCMC) algorithm described by Richardson and Green (1997), which they applied to univariate Normal mixture models.

Whilst MCMC provides a convenient way to draw inference from complicated statistical models, there are still many, perhaps under appreciated, problems associated with the MCMC analysis of mixtures. The problems are mainly caused by the nonidentifiability of the components under symmetric priors, which leads to so called label switching in the MCMC output, discussed in Jasra et al. (2005). It generally happens when the subpopulations are not well separated.

The aim of this work is to discuss the application of mixture models to heterogeneous populations, in particular Normal mixture models, under the Bayesian approach. The main purpose is to evaluate the

model's performance in different settings of heterogeneity and considering the number of components known and unknown.

This work is organized as follow. Section 2 presents the definition of a mixture model and discusses aspects of the inference. In Section 3 presents a simulation study for assessing the estimation of model parameters under different levels of heterogeneity. In Section 4 the performance of the methodology is assessed through application to a real dataset. Finally, Section 5 presents some conclusions and suggestions for further research.

2. Finite mixture models

The basic mixture model for independent scalar or vector observations Y_i, $i = 1, \ldots, n$ is given by:

$$Y_i \sim \sum_{j=1}^{k} w_j f(\cdot \mid \boldsymbol{\theta}_j),\ i = 1, \ldots, n, \tag{1}$$

where $f(\cdot \mid \boldsymbol{\theta})$ is a given parametric family of densities indexed by a scalar or a vector $\boldsymbol{\theta}$. In general, the objective of the analysis is to make inferences about the unknowns: the number of components, k; the parameters $\boldsymbol{\theta}_j$'s and the components' weights, w_j, $0 < w_j < 1$, $\sum_{j=1}^{k} w_j = 1$. Let $\boldsymbol{\Phi} = (\boldsymbol{w}, \boldsymbol{\theta}, k)$ be the parametric vector of the model (1).

For a random sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ observed, the likelihood function of $\boldsymbol{\Phi}$ is given by:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}) = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(y_i|\boldsymbol{\theta}_j).$$

The likelihood function leads to k^n terms, what brings a computational difficulty.

A context in which the model (1) can arise and we are interested in this paper is when we postulate a heterogeneous population consisting of heterogeneous groups $j = 1, 2, \ldots, k$ of sizes proportional to w_j, from which a random sample is drawn. The label of the group from which each observation is drawn is unknown. As stated in Richardson and Green (1997), it is natural to regard the group label z_i, for the i-th observation as a latent variable and rewrite (1) as the following hierarchical model:

$$Y_i \mid \boldsymbol{\theta}_j, z_i = j \sim f(\cdot \mid \boldsymbol{\theta}_j),\ \text{with } P(z_i = j) = w_j,\ i = 1, \ldots, n,\ j = 1, \ldots, k. \tag{2}$$

Integrating $\mathbf{z} = (z_1, \ldots, z_k)$ out from (2) we return to model (1). The formulation given by (2) is convenient for interpretation and calculation, decreasing the computational cost.

The mixture model in (2) is invariant to permutation of the labels $j = 1, \ldots, k$. Therefore, it is important to adopt unique labeling to ensure identifiability. For example, if θ is a scalar, we can impose an ordering constraint on θ_j's, such as $\theta_1 < \theta_2 < \ldots < \theta_k$.

Finally, as we are in a Bayesian framework, the prior elicitation is an important question. Being fully non-informative and obtaining proper posterior distributions are not possible in a mixture content. An alternative on this case is to define weakly informative priors, which may or may not be data dependent. Moreover, the constraint mentioned above, in order to ensure identifiability, is usually imposed on the prior distribution assumed of the parametric vector $\boldsymbol{\Phi}$.

2.1 Normal mixture model

In this work we are particularly interested in the univariate Normal case presented in Richardson and Green (1997), where $\boldsymbol{\theta}_j$ becomes a vector of expectation and variance (μ_j, σ_j^2). Assuming that the parameters in $\boldsymbol{\Phi}$ are prior independent, the model is described below: for $i = 1, \ldots, n$ and $j = 1, \ldots, k$,

$$
\begin{aligned}
Y_i \mid \mu_j, \sigma_j^2, z_i = j &\sim \text{Normal } (\mu_j, \sigma_j^2), \\
P(z_i = j) &= w_j, \\
\boldsymbol{w} &\sim \text{Dirichlet } (\boldsymbol{\gamma}), \\
\mu_j &\sim \text{Normal } (\mu_a, \sigma_a^2), j = 1, \ldots, k, \\
\sigma_j^{-2} &\sim \text{Gamma } (\alpha, \beta), j = 1, \ldots, k, \\
\beta &\sim \text{Gamma } (g, h), \\
k &\sim \text{Uniform } \{1, k_{max}\}.
\end{aligned} \tag{3}
$$

The default hyperparameter choices are described in Richardson and Green (1997) and are elicited making minimal assumptions on the data. For identifiability, we can use for example that the μ_j are in increasing numerical order, thus the joint prior distribution of $\boldsymbol{\Phi}$ is $k!$ times the product of their marginal prior distributions.

2.2 Inference

As we are in a Bayesian framework, the inference consists in obtain the posterior distribution of the parametric vector $\boldsymbol{\Phi}$ of model (2). In general, this distribution cannot be obtained in closed form. Therefore, it is necessary to use some numerical approximation methods. One alternative, which is often used and is feasible to implement, is to generate samples from the marginal distributions of the parameters based on the MCMC algorithm. Nevertheless, this method, as originally formulated, requires the posterior distribution to have a density with respect to some fixed measure. Thus, in the mixture context, the method can only be applied when the number of components k in the model (2) is considered known.

However, rarely the number k is known, and fix it on a incorrect value can bring important consequences to the posterior distribution. Other times, the target of the study is exactly the estimation of k. The approach based on RJMCMC is an alternative in this case. It basically consists of jumps between the parameter subspaces corresponding to different numbers of components in the mixture.

To do all the inference in this work we particularly used the R package `mixAK` presented in Komárek (2009).

3. Simulation study

To examine the performance of the model for heterogeneous populations, we generated simulated samples and obtained samples from the posterior distribution of the parametric vector, supposing k known and estimating it. The estimates were then compared with the true values to evaluate the model's performance.

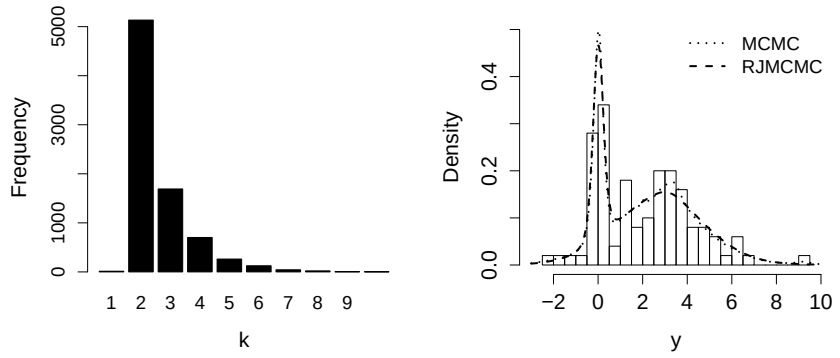
We generated samples of size $n = 100$, fixing $k = 5$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2) = (0.22, 1.95, 0.92, 0.74, 1.13)$ and $(w_1, w_2, w_3, w_4, w_5) = (0.21, 0.51, 0.23, 0.03, 0.02)$. We considered two different levels of heterogeneity fixing values for the means more similar and very different. We considered $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (-3, 0, 4, 11, 16)$ for the sample more heterogeneous and $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (0, 2, 4, 6, 8)$ for the homogeneous one.

We used the same prior distribution for $\boldsymbol{\Phi}$ described in Richardson and Green (1997) and for identifiability we considered that the μ_j are in increasing numerical order. It should be noted that on the case more heterogeneous, this is, when the means are well separated, labelling of the realizations from the posterior by ordering their means generally coincides with the sample labelling. However, on the more homogeneous case the so called label switching eventually occurred.

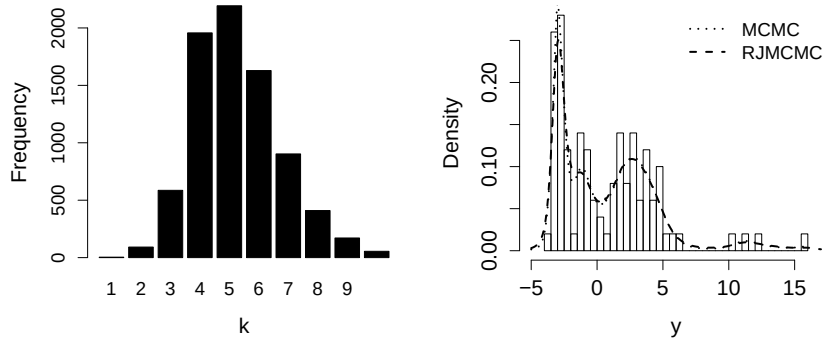
On the inference we considered the model fit estimating k and fixing it on its true value. For the RJMCMC and MCMC simulations, we generated 45,000 samples from the posterior distribution, discarded the first 5,000, then thinned the chain by taking every 5th sample value, and the convergence was achieved.

Figure 1 displays (a) the histogram with the posterior densities of k and (b) the predictive density estimate of the distribution of Y, for the two datasets generated, estimating k (RJMCMC) and fixing $k = 5$ (MCMC), represented by the dashed and dotted lines, respectively. It should be noted that k is better estimated for the

data more heterogeneity and the predictive for the heterogeneous one identifies easier the subpopulations. Moreover, the predictive density obtained considering k unknown and known are very similar. Even when fixing $k = 5$ for the homogeneous sample, the predictive follows the same behavior obtained by the RJMCMC, thus fixing k does not improve the estimation.



(a) Homogeneous



(b) Hetereogeneous

Figure 1: *Posterior densities of k and predictive densities for the (a) homogeneous and (b) heterogeneous artificial data sets.*

Finally, it should be considered that the fourth and fifth components have a few observations, which complicates the inference, principally with respect to k.

4. Application to a real dataset

We applied the methodology on a real dataset that concerns antibody levels of Cytomegalovirus (CMV) in individuals, both males and females, from 6 years to 49 years old. This data set was extracted from the 2003 - 2004 National Health and Nutrition Examination Survey (NHANES) and can be obtained from NHANES website. The CMV is a member of the *Herpesviridae* family of viruses and, according to Kusne et al. (1999), is the most common viral pathogen in organ transplant recipients. The CMV IgG test allows to know if someone is infected or not. The range of values for the antibody levels CMV IgG are from 0.048 to 3.001. To the values reported as "out of range" (i.e. over the detectable range, > 3.00) the survey specialists usually assign the value of 3.001. Thus, there are a lot of individuals with this particular value. Figure 2 shows the

distribution antibody levels of CMV IgG for 5126 individuals infected and not infected. The interest here is in identifying subgroups of IgG as a marker of the presence of the disease.

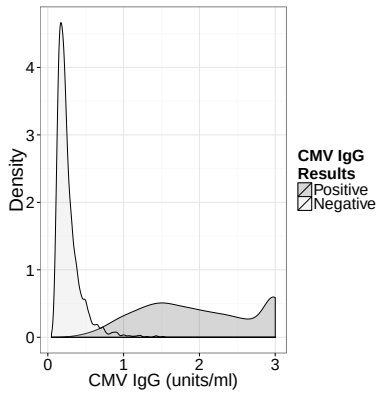


Figure 2: *Distribution of antibody levels of CMV IgG (units/ml) for 5126 individuals infected by the virus or not.*

As shown in Figure 2, we clearly identify two or three heterogeneous subpopulations, so we will analyze this data using a Normal mixture model. On the inference we considered also the one estimating k and fixing it on its true value. For the RJMCMC simulations, we generated 45,000 samples from the posterior distribution, discarded the first 5,000, then thinned the chain by taking every 5th sample value. For the MCMC simulations, we considered 50,000 sweeps, then discarded the first 10,000 and thinned the chain by taking every 10th sample value.

Figure 3 displays the posterior distribution of k and predictive densities of antibody levels estimating k (RJMCMC) and fixing $k = 3$ (MCMC), represented by the dashed and dotted lines, respectively. The posterior obtained from RJMCMC for k favours 3 components, however, the third one is due to the group assigned as 3.001. It should be noted that predictive plots for RJMCMC and MCMC are very similar, showing a great performance even when k is estimated.

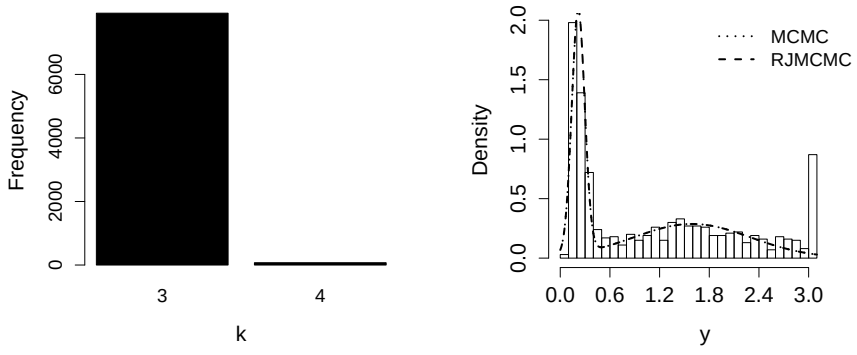


Figure 3: *Posterior distribution of k and predictive densities for the real dataset.*

Another criterion used as comparison for both methods was the Deviance Information Criteria (DIC), which evaluates the goodness of fit of the model, and smaller values of DIC indicates a better fit. The model fit estimating k and fixing it in $k = 3$ results, respectively, a DIC of -3692.59 and -3693.99. Thus, as DIC increases with the number of parameters, it is expected that RJMCMC presents a higher DIC. However,

since both DICs were very similar, it is possible to conclude that both methods are efficient in this case.

5. Conclusions

We have considered the problem of evaluating the fit of a Normal mixture problem to populations with different levels of heterogeneity. The estimation was done considering the number of components k known and unknown, this is using MCMC and RJMCMC methods, respectively. The simulation study shows that as heterogeneity between subpopulations decreases, the number of subpopulations tends to be underestimated. Even when the number of components is fixed in the real value, the subpopulations are not all easily identified.

The mixture models are a flexible alternative to heterogeneous populations, but as heterogeneity between subpopulations decreases it might be worth using the mixture model. Moreover, there are still many, perhaps under appreciated, problems associated with the MCMC analysis of mixtures. Therefore, we suggest the use of this modeling to problems where the heterogeneity is really expected. On the other hand, when there is no idea of the number of components we also recommend the inference considering k unknown. The results are very similar for both approaches.

The main findings of this work encourage to do a bigger simulation study with several samples to evaluate better the mixture model's performance not only in different cases of heterogeneity, but also under different sets of w and sample sizes.

References

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2003 - 2004][http://www.cdc.gov/nchs/nhanes].

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 50-67.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, 140-155.

Komárek, A. (2009). A new R package for Bayesian estimation of multivariate Normal mixtures allowing for selection of the number of components and interval-censored data. *Computational Statistics & Data Analysis*, 53(12), 3932-3947.

Kusne, S., Shapiro, R., & Fung, J. (1999). Prevention and treatment of cytomegalovirus infection in organ transplant recipients. *Transplant infectious disease*, 1(3), 187-203.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4), 731-792.