Identifying Mixtures of Mixtures Using Bayesian Estimation

Gertraud Malsiner-Walli[1], Sylvia Frühwirth-Schnatter [2] and Bettina Grün[1]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 22

[1] Johannes Kepler Universität Linz, Institut für Angewandte Statistik, Linz, Austria
Gertraud.Malsiner_Walli@jku.at
Bettina.Gruen@jku.at
[2] WU Wirtschaftsuniversität Wien, Institute for Statistics and Mathematics, Wien, Austria
Sylvia.Fruehwirth-Schnatter@wu.ac.at

Abstract

In a mixture of mixtures model the cluster distributions are approximated by a mixture distribution. However, idenfying the components forming one cluster is in general not straight-forword. Using a mixture of mixtures of Gaussian distributions we propose a Bayesian estimation scheme based on MCMC methods and Gibbs sampling and the specification of suitable priors to automatically fit a suitable mixture model to each cluster and determine the mixture model on the cluster level. We evaluate our proposed approach in a simulation setup with artificial data and by applying it to benchmark data sets.

Keywords: Bayesian estimation; finite mixture model; model-based clustering; multivariate normal distribution.

1 Introduction

Standard model-based clustering assigns observations to different clusters by fitting a mixture model of Gaussian distributions, where each Gaussian component corresponds to a cluster. However, often in empirical applications a cluster is poorly fitted by a single Gaussian distribution, e.g., if the cluster has a non-symmetrical shape. In this case, more than one Gaussian component is needed

to model the cluster and the number of components overestimates the number of clusters.

Previous approaches to identify the number of clusters as well as each of their distributions within this mixture of mixtures approach used methods for combining mixture components to form a cluster after first selecting the total number of components of a suitably fitting model using, e.g., the BIC [1, 4]. In general these methods group the components of the selected model into clusters according to criteria which evaluate the obtained partition.

In our approach the number of clusters and their corresponding cluster distributions are directly estimated during MCMC sampling by imposing suitable priors. In particular, we use informative hierarchical priors for the mixture parameters to encourage the components assigned to the same cluster to have overlapping distributions and to approximate a connected and dense cluster distribution.

2 Model Specification

We fit a mixture model to the data $\mathbf{y}_i$, where we assume that each mixture component is itself a mixture of multivariate Gaussian distributions:

$$\mathbf{y}_i \sim \sum_{k=1}^{K} \sum_{l=1}^{L} \eta_k w_{kl} f_{\mathcal{N}}(\mathbf{y}_i | \boldsymbol{\mu}_{kl}, \boldsymbol{\Sigma}_{kl}),$$

where K is the number of clusters, L is the number of Gaussian distributions building a cluster k, η_k are the cluster weights and w_{kl} are the component weights within each cluster k.

We estimate the number of clusters K based on the idea of *sparse finite mixtures*, introduced by Frühwirth-Schnatter at the ISBA meeting in Kyoto, 2012. Because the estimation of the exact number of "true" components within a cluster is not necessary, we use the same abundant number of normal components L for each cluster and impose a prior to ensure that each of them is filled during MCMC sampling [3, 5]. We select the priors on the component means $\boldsymbol{\mu}_{kl}$ within each cluster k to concentrate a lot of mass near the cluster center in order to pull the component means towards the cluster center.

3 Model Estimation

We perform Bayesian estimation of finite mixture models using MCMC methods based on data augmentation and Gibbs sampling. To solve the label switching problem on the cluster level, a random permutation step is added to the MCMC scheme to ensure balanced label switching during MCMC sampling [2]. In a post-processing step the MCMC output is relabeled by clustering the MCMC draws in the point process representation using k-centroid cluster analysis based

on the Mahalanobis distance. We evaluate our proposed method in a simulation setup with artificial data and by applying it to benchmark data sets.

References

[1] J.P. Baudry, A. Raftery, G. Celeux, K. Lo, R. Gottardo. **Combining Mixture Components for Clustering**. *Journal of Computational and Graphical Statistics*; 2010; 19(2); pp. 332-353.

[2] S. Frühwirth-Schnatter. **Finite Mixture and Markov Switching Models**. *Springer*; 2006.

[3] S. Frühwirth-Schnatter. **Label Switching Under Model Uncertainty**. *Mixtures: Estimation and Application*. 2011; pp. 213-239.

[4] J. Li. **Clustering Based on a Multilayer Mixture Model**. *Journal of Computational and Graphical Statistics*; 2005; 14(3); pp. 547-568.

[5] J. Rousseau, K. Mengersen. **Asymptotic Behaviour of the Posterior Distribution in Overfitted Mixture Models**. *Journal of the Royal Statistical Society B*; 2011; 73(4); pp. 689-710.