Finite Mixture Model Diagnostics Using the Parametric Bootstrap

Bettina Grün
Department of Statistics & Prob. Theory
Vienna University of Technology
Vienna, Austria
Email: Bettina.Gruen@ci.tuwien.ac.at

Friedrich Leisch
Department of Statistics
University of Munich
Munich, Germany
Email: Friedrich.Leisch@stat.uni-muenchen.de

Abstract — *Finite mixture models are a popular tool for modelling unobserved heterogeneity. As these models are in general very complex, it is essential to have suitable methods for model diagnostics which allow e.g. to check for model identifiability, model fit and possible model restrictions. In this paper we propose to use the parametric bootstrap for model diagnostics and to visualize the bootstrap results using parallel coordinate plots. The application of the proposed methods is illustrated using an artificial example.*

I. Introduction

This paper outlines the use of the parametric bootstrap for finite mixture model diagnostics as a special case of the general framework presented in [1], which encompasses different resampling methods.

II. Finite mixture models

The finite mixture models considered in this paper are given by

$$H(y|\mathbf{x}, \Theta) \quad = \quad \sum_{k=1}^{K} \pi_k F(y|\mathbf{x}, \boldsymbol{\vartheta}_k)$$

where H is the mixture distribution, $\mathbf{x}$ is an optional vector of regressors, y the vector of responses, K the number of components, F the component distribution function, $\boldsymbol{\vartheta}_k$ the component specific parameters and π_k the subcomponent probabilities. Θ is the vector of all parameters with $\Theta \in \Omega$, where Ω denotes the space of admissible parameters for K-component mixtures. The restrictions on the parameters are

- $0 < \pi_k \leq 1$, $\forall k = 1, \ldots, K$,
- $\sum_{k=1}^{K} \pi_k = 1$, and
- $\boldsymbol{\vartheta}_k \neq \boldsymbol{\vartheta}_l$, $\forall l \neq k$ with $l, k \in \{1, \ldots, K\}$.

Given the number of components K and the component distribution function F, $\mathscr{A}_K = \mathscr{A}_K(F, \Omega)$ denotes the set of all finite mixture models with K components and mixture distributions of form $H(\cdot|\cdot, \Theta)$

In a frequentist framework the Expectation-Maximization (EM) algorithm is the most popular method to determine the model $a(\mathcal{X}_N) \in \mathscr{A}_K$ with the maximum likelihood. As the EM algorithm might be trapped in a local optimum, it is in general recommended to choose the best solution of several runs with different starting values in order to detect the global maximum.

III. Model diagnostics using resampling methods

Resampling methods are already a popular technique for model diagnostics of linear and generalized linear models. The use of the parametric bootstrap for finite mixture models has been proposed for

- determining the number of components [2, 3]
- estimating standard deviations [4] and
- checking for identifiability problems [5].

The parametric bootstrap procedure can be outlined by:

1. Estimate $\hat{a}(\mathcal{X}_N) \in \mathscr{A}_K$ and determine a corresponding parameterization $\hat{\Theta} \in \Omega$.
2. Sample B bootstrap samples $\mathcal{X}_N^b$ $(b = 1, \ldots, B)$ independently with the parametric bootstrap: $\mathcal{X}_N^b \sim \hat{a}(\mathcal{X}_N)$.
3. Fit models to the bootstrap samples using the EM algorithm with either
   (a) random initialization: $\hat{a}^b(\mathcal{X}_N^b) \in \mathscr{A}_{K_0}$ with possibly $K_0 \neq K$, or
   (b) initialization in $\hat{\Theta}$: $\hat{a}^b(\mathcal{X}_N^b, \hat{\Theta}) \in \mathscr{A}_{K_0}$ with $K_0 = K$.

Depending on whether global or local characteristics of the fitted model are analyzed either random initialization or initialization in the solution is used. Random initialization introduces the problem of *label switching* which has already received some attention in Bayesian analysis and which makes it necessary to suitably relabel the components before making component-specific analyses.

IV. Visualization

Parallel coordinate plots are a visualization technique for hyperdimensional data [6]. For visualizing the bootstrap results the data used are the parameter estimates of each component of the models fitted to each bootstrap sample. This visualization technique can be enhanced by

- adding the confidence intervals for the parameter estimates derived using standard asymptotic theory and

- using different colors or line types for each of the components after appropriate relabelling, e.g. by imposing an ordering constraint on one of the parameters.

V. Example

All computations are done in R [7] using package **flexmix** [8]. For illustration we use an artificial example of a finite mixture of Gaussian regression models which is not identifiable due to intra-component label switching [9]. It is assumed that the following mixture distribution consisting of three components has been fitted to a sample with 50 observations for each value of x where x is a binary variable with values $\{0, 1\}$:

$$\begin{array}{ll} \text{Class 1:} & \pi_1 = 0.45,\ y = x + \epsilon \\ \text{Class 2:} & \pi_2 = 0.45,\ y = 2 + x + \epsilon \\ \text{Class 3:} & \pi_3 = 0.10,\ y = -2 + \epsilon \end{array}$$

with $\epsilon \sim N(0, 0.1)$.

The parametric bootstrap procedure is applied to this model with $B = 200$. As identifiability problems are investigated, the EM algorithm is randomly initialized and the best solution of 5 repetitions is reported. The fitted parameters are visualized in Figure 1. It can clearly be seen that the estimated parameters cluster around three distinct values for the coefficient of the intercept, while they cluster around a single point for σ. The identifiability problem is indicated by the different bundles which connect the estimates of the coefficients of the intercept to those of x.

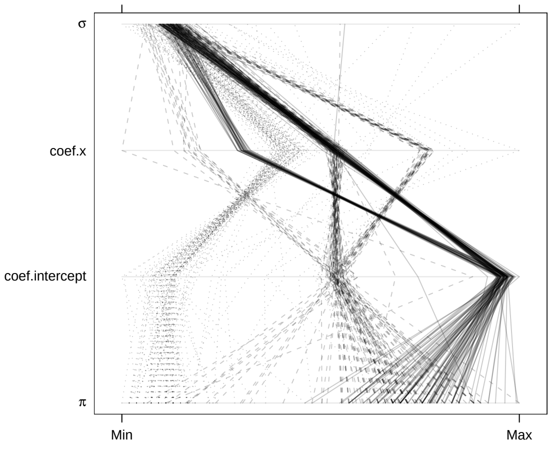

Figure 1: Parallel coordinate plot of the parameters fitted to 200 parametric bootstrap samples. The line types are according to an ordering constraint on the intercept.

VI. Conclusions & future work

The presented procedure is based on resampling methods and can be seen as complementary to methods using standard asymptotic theory. In order to facilitate the interpretation of the results a visualization method is proposed which uses parallel coordinate plots.

In the future these methods shall be implemented by extending the R package **flexmix**. As **flexmix** allows the user to easily extend available functionality and develop new mixture models, the diagnostic tools can be used for all these models as this approach is general enough to be applied to different kinds of mixture models.

Acknowledgments

This research was supported by the Austrian Academy of Sciences (ÖAW) through a DOC-FFORTE scholarship and the Austrian Science Foundation (FWF) under grant P17382.

References

[1] B. Grün and F. Leisch. Finite mixture model diagnostics using resampling methods. Unpublished manuscript, 2005.

[2] G. J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society C*, 36(3):318–324, 1987.

[3] Z. D. Feng and C. E. McCulloch. Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society B*, 58(3):609–617, 1996.

[4] K. E. Basford, D. R. Greenway, G. J. McLachlan, and D. Peel. Standard errors of fitted means under normal mixture model. *Computational Statistics*, 12:1–17, 1997.

[5] B. Grün and F. Leisch. Bootstrapping finite mixture models. In Jaromir Antoch, editor, *Compstat 2004 — Proceedings in Computational Statistics*, pages 1115–1122. Physica Verlag, Heidelberg, 2004.

[6] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675, 1990.

[7] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[8] F. Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 2004.

[9] C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296, 2000.