COMPSTAT'2004 Symposium

BOOTSTRAPPING FINITE MIXTURE MODELS

Bettina Grün and Friderich Leisch

Key words: Finite mixture models, bootstrapping, identifiability, regression.
COMPSTAT 2004 section: Clustering, Resampling methods.

Abstract: Finite mixture regression models are used for modelling unobserved heterogeneity in the population. However, depending on the specifications these models need not be identifiable, which is especially of concern if the parameters are interpreted. As bootstrap methods are already used as a diagnostic tool for linear regression models, we investigate their use for finite mixture models. We show that bootstrapping helps in revealing identifiability problems and that parametric bootstrapping can be used for analyzing the reliability of coefficient estimates.

1 Introduction

During the last decades it has become popular to include covariates into mixture models leading to mixture regression models. Applications can be found, e.g., in marketing [15] or for clinical trials [13], where unobserved heterogeneity of the population is present in the data and should therefore be taken into account in the modelling process.

The mixture regression models we consider can be formulated by

$$H(y|\mathbf{x},\Theta) = \sum_{k=1}^{K} \pi_k F(y|\mathbf{x},\theta_k,\phi_k), \quad 0 < \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

where H is the mixture distribution, $\mathbf{x}$ are the regressors, y the responses, K the number of components, F the component distribution functions, θ_k the regression coefficients, ϕ_k the (possible) dispersion parameters, π_k the prior class probabilities and Θ the vector of all parameters. If the component distribution functions F are from the exponential family the generalized linear modelling framework [7] can be used leading to the so-called GLIMMIX models [14].

A very popular way of estimating mixture models is the EM method, which is a class of iterative algorithms for maximum likelihood estimation in problems with incomplete data [2]. It has been shown that during the classical EM algorithm the values of the likelihood are monotonically increased. The likelihood is in general multimodal with a unique internal global maximum (if the model is identifiable) and several local maxima. Unboundedness of the likelihood might occur at the edge of the parameter space [5]. Then, the solution found by the EM algorithm depends on its initialization.

One possibility to initialize the EM algorithm is to use a partition of the data into the number of requested segments [8]. This partition can be generated randomly or by applying some clustering algorithm, such as, e.g., k-means. In order to ensure that the global maximum is found the EM algorithm is in general run several times with different initializations.

The solution of the EM algorithm does also depend on the given data set. For eliminating random effects of a given data set the results for different samples from the same data generating process (DGP) can be compared. However, in applications there is in general only one data set available. A remedy can be then to draw samples from the empirical distribution of the given data set, i.e., bootstrapping [1]. Furthermore, the parametric bootstrap can be used to assess the stability of the estimated parameters.

2 Identifiability

Finite mixture models are trivially not identifiable with respect to the ordering of the segments and to overfitting, as this leads to empty segments or to several segments having the same parameters. By imposing constraints, e.g., on the ordering of the components, these identifiability problems can be eliminated.

It has been shown that except for these identifiability problems finite mixture distributions of several popular continuous distributions are generically identifiable, as e.g., the (multivariate) normal, gamma and exponential distribution [11], [12], [16]. A discrete identifiable distribution is the Poisson distribution [10]. In contrast the discrete and the continuous uniform distributions are not generically identifiable. The binomial and the multinomial distributions are identifiable if the number of segments is limited with respect to the repetition parameter [11], [3].

In a first analysis of mixture regression models it has been shown that the identifiability of standard linear regression models is not guaranteed even if the regressor matrix has full rank [4], which is also shown by the example given in Section 3.1.1. Furthermore, multinomial mixture regression models were analyzed in [3].

The identifiability of a mixture regression model depends on the distribution of the dependent variable, the maximum number of segments allowed, the available information per object and the regressor matrix. With respect to the regressor matrix identifiability problems might arise if there are only a limited number of different covariate points and if in addition there is only very limited information per person available. Such problems might occur in applications because the covariates are often categorical variables, as, e.g., gender, promotion in marketing, likes/dislikes, test and control group in clinical trials .... These variables are in general coded as dummy variables.

3 Simulation

Two examples are presented where bootstrap methods are applied to finite mixture models. A standard linear regression example is used for demonstrating that bootstrap samples can reveal identifiability problems, whereas a Poisson regression example shows that by parametric bootstrapping the stability of the estimated coefficients can be challenged.

Our simulation was performed using the R environment for statistical computing [9]. For EM estimation the contributed package `flexmix` [6] was taken, which implements a general framework for finite mixtures of regression models.

As the EM algorithm might be trapped in local maxima, we always made five initializations with random partitions of the data and considered only the best result with respect to the log-likelihood. Hence, the result of one "run" of the EM algorithm refers to the best result out of these five repetitions.

3.1 Bootstrapping global maxima

We investigate the convergence of the EM algorithm to different global maxima with respect to a simple standard linear mixture regression example with two global maxima.

3.1.1 Normal mixture example Assume we have a standard linear mixture regression with one measurement per object and two different covariate points $\mathbf{x}_1 = (1, 0)'$ and $\mathbf{x}_2 = (1, 1)'$. Furthermore, let the mixture consist of two components with equal prior class probabilities.

The mixture regression can be formulated as

$$H(y|\mathbf{x}, \Theta) = \frac{1}{2}N(\mu_1, 0.1) + \frac{1}{2}N(\mu_2, 0.1)$$

where $\mu_i(\mathbf{x}) = \mathbf{x}'\theta_i$ and $N(\mu, \sigma^2)$ is the normal distribution.

As Gaussian mixture distributions are generically identifiable the means, variances and prior class probabilities are uniquely determined in each covariate point given the mixture distribution. If we assume that $\mu_1(\mathbf{x}_1) = 1$, $\mu_2(\mathbf{x}_1) = 2$, $\mu_1(\mathbf{x}_2) = -1$ and $\mu_2(\mathbf{x}_2) = 4$, the two possible solutions for θ are:

$$\begin{aligned} \theta_1^1 &= (2, \quad 2)', & \theta_2^1 &= (1, -2)' \quad \text{and} \\ \theta_1^2 &= (2, -3)', & \theta_2^2 &= (1, \quad 3)' \end{aligned}$$

A balanced sample of length 100 has been generated and can be seen together with the regression lines corresponding to the two different solutions in Figure 1.

3.1.2 Simulation results As the EM algorithm converges nearly always to the same global maximum for a given data set, we eliminate the influence

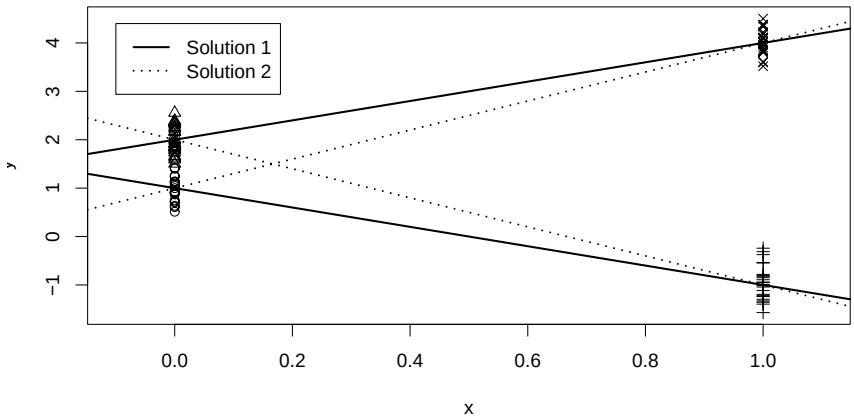

Figure 1: Data and theoretical solutions.

of a given data set by using different samples from the DGP and several bootstrap samples from a given data set. As the unidentifiability depends on the equality of the prior class probabilities we disturb them slightly in order to assess the sensitivity with respect to this parameter.

Thus, we generated 1000 balanced samples of size 100 from the mixture regression model specified in Section 3.1.1 for the DGP analysis. For the bootstrapping analysis we generated 20 samples in the same way and to each data set we added 49 bootstrap samples of the same length. For the sensitivity analysis of the prior class probabilities we applied the same setup except that we used prior class probabilities $(0.6, 0.4)$ and $(0.7, 0.3)$ respectively and the coefficients from solution 1 in Section 3.1.1 for sampling from the DGP. We made 10 runs of the EM algorithm for each sample (from the DGP or bootstrap) in order to show that if the sample is fixed the EM algorithm converges in most of the cases to the same global maximum.

We decided that a result of the EM algorithm is equal to one of the solutions if the maximum distance between the coefficients is less than 0.25 after ordering the components with respect to the intercept which is necessary due to label switching. Then we determined how often which solution was detected and how often during all ten runs only one, none and both of the solutions were found (cp. Table 1).

It can be seen that the results are similar for bootstrapping and sampling from the DGP except that the solutions are always found less often for the BS samples than the DGP samples. Even though this behavior is intuitive, it nevertheless causes that the percentage where none of the solutions are found is considerably higher for the BS samples. If the prior class probabilities are not equal but are 0.6 and 0.4, solution 2 is still found in about 20% of the runs and it is the only solution found during 10 repeated runs for nearly the same percentage. Because of the unequal priors solution 2 is only a local maximum, but obviously the attraction area is large enough that it is relatively often the best solution found in one run of the EM algorithm. The convergence

Priors	**Equal**		**0.6/0.4**		**0.7/0.3**	
	DGP	**BS**	**DGP**	**BS**	**DGP**	**BS**
Overall fraction of						
Solution 1	0.49	0.38	0.74	0.58	0.87	0.68
Solution 2	0.46	0.41	0.20	0.20	0.01	0.04
Fraction over 10 runs of						
Only solution 1	0.45	0.36	0.72	0.56	0.87	0.66
Only solution 2	0.43	0.39	0.18	0.18	0.01	0.04
None of them	0.05	0.20	0.06	0.22	0.12	0.28
Both solutions	0.06	0.04	0.03	0.03	0.00	0.01

Table 1: Simulation results.

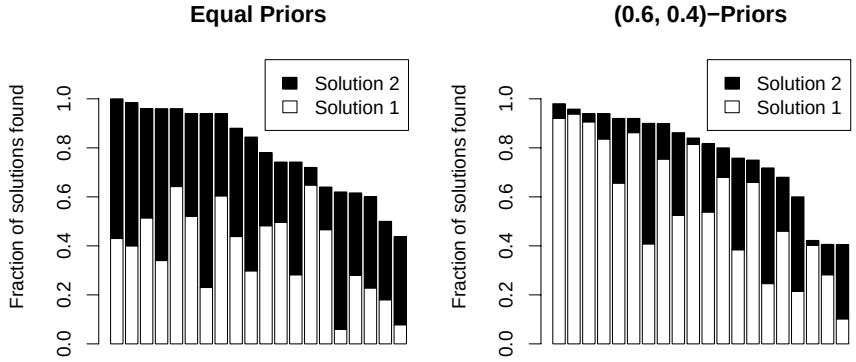

Figure 2: Fraction of only solution 1 or 2 found separately for each of the 20 data sets generated from the DGP when BS.

to solution 2 gets less the more the prior class probabilities deviate from 0.5 such that for $(0.7, 0.3)$ solution 2 is hardly ever found.

In Figure 2 the bootstrap results are analyzed separately for the 20 different data sets generated from the DGP. It can be seen how often which solution was the only one found during 10 runs. Even though the results are varying a lot for the different data sets, it can nevertheless be seen that both valid solutions are found every time by bootstrapping the data for equal priors and that for unequal priors there are data sets, where both solutions are nearly equally often found.

3.2 Parametric bootstrapping

Given a solution bootstrapping from the estimated distribution can be used for assessing the stability of the estimates. Note that we do not intend to replace the standard tools for estimating standard deviations but we propose that by additionally applying the parametric bootstrap it can be assessed if

the standard asymptotic theory is appropriate. We use a Poisson mixture regression model to analyze the application of this method to finite mixture models.

3.2.1 Poisson mixture In [13] a Poisson mixture regression is fitted to data from a clinical trial where the effect of intravenous gammaglobulin on suppression of epileptic seizures is investigated. The data used were 140 observations from one treated patient, where treatment has started on the 28^{th} day. In the regression there were three independent variables included: treatment, trend and interaction treatment-trend. Treatment is a dummy variable indicating if the treatment period has already started. Furthermore, the number of parental observation hours per day were available and it is assumed that the number of epileptic seizures per observation hour follows a Poisson mixture distribution. The fitted mixture distribution consisted of two components which can be interpreted as representing 'good' and 'bad' days of the patients.

The mixture model can be formulated by

$$H(y|\mathbf{x}, \Theta) = \pi_1 P(\lambda_1) + \pi_2 P(\lambda_2)$$

where $\lambda_i = e^{\mathbf{x}'\theta_i}$ for $i = 1, 2$ and $P(\lambda)$ is the Poisson distribution.

By reestimating this model we became nearly equal results than in [13]. Our solution for θ_i $i = 1, 2$ with the corresponding standard deviations is:

$$\theta_1 = (2.84, 1.30, -0.41, -0.43)' \quad \text{with SD}(\theta_1) = (0.23, 0.47, 0.09, 0.13)'$$
$$\theta_2 = (2.07, 7.43, -0.27, -2.28)' \quad \text{with SD}(\theta_2) = (0.09, 0.52, 0.04, 0.14)'$$

The size of the first component representing 'bad' days is 0.28.

3.2.2 Simulation results We generated 100 samples from the estimated mixture distribution with the same structure as the sample used in [13] and applied the EM algorithm to them.

In Figure 3 the theoretical means are given together with the 95% confidence intervals derived with standard asymptotic theory in both plots. The means estimated for the bootstrap samples classified with respect to their prior class probability have been separated. While the confidence intervals for baseline and treatment period are of similar width, it can clearly be seen from the bootstrap application that estimation in the baseline period is much less stable than in the treatment period.

Furthermore, there can be component label switching observed. As the dummy variable treatment is included with its interaction terms and the component sizes are not separated enough, there are solutions which join days with low numbers of seizure episodes during baseline period with those with high numbers during treatment period.

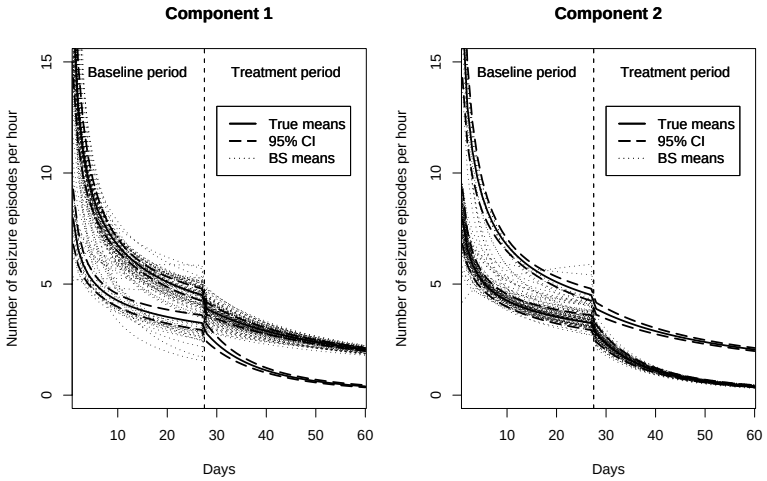

Figure 3: Theoretical means and estimated means of bootstrap samples.

4 Conclusion and future research

We showed that bootstrapping can be a valuable diagnostic tool when estimating finite mixture models as it can reveal identifiability problems and give further insight into the stability of parameter estimates.

Obviously, there exist data sets where the attraction area to one global maximum is that large that only this solution is found by using different initializations to the EM algorithm. Bootstrapping can be used in such a situation for revealing other global maxima. Furthermore, the stability of a solution can be analyzed by parametric bootstrap. In an example it was shown that by including categorical variables with their interaction terms label switching can also occur within components, because there are solutions found where components are joined differently for the different values of the factor due to the flexibility of the estimated model.

As computers are nowadays fast enough to repeat the EM algorithm with different initializations and different input samples within a reasonable amount of time, we recommend to use bootstrapping in addition to the already commonly accepted strategy to use different initializations, in order to ensure a higher stability of the method with respect to the dependency on a certain initialization and data set. Furthermore, by applying the parametric bootstrap additional insights can be gained on the stability of the estimates complementing the results derived with standard asymptotic theory.

References

[1] Davison A.C., Hinkley D.V. (1997). *Bootstrap methods and their application*. Cambridge series on statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK.

[2] Dempster A., Laird N., and Rubin D. (1977). *Maximum likelihood from incomplete data via the EM-alogrithm.* Journal of the Royal Statistical Society B **39**, 1 – 38.

[3] Grün B. (2002). *Identifizierbarkeit von multinomialen Mischmodellen.* Master's thesis, Technische Universität Wien, Kurt Hornik and Friedrich Leisch, advisors.

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

[5] Kiefer N.M. (1978). *Discrete parameter variation: Efficient estimation of a switching regression model.* Econometrica **46** (2), 427 – 434.

[6] Leisch F. (2003). *FlexMix: A general framework for finite mixture models and latent class regression in R.* Report 86, SFB "Adaptive Information Systems and Modeling in Economics and Management Science".

[7] McCullagh P., Nelder J. (1989). *Generalized linear models.* Chapman and Hall.

[8] McLachlan G., Peel D. (2000). *Finite mixture models.* John Wiley and Sons Inc.

[9] R Development Core Team. (2003). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[10] Teicher H. (1960). *On the mixture of distributions.* Annals of Mathematical Statistics **31**, 55 – 73.

[11] Teicher H. (1963). *Identifiability of finite mixtures.* Annals of Mathematical Statistics **34**, 1265 – 1269.

[12] Titterington D., Smith A., Makov U. (1985). *Statistical analysis of finite mixture distributions.* Chichester: Wiley.

[13] Wang P., Puterman M., Cockburn I., Le N. (1996). *Mixed poisson regression models with covariate dependent rates.* Biometrics **52**, 381 – 400.

[14] Wedel M., DeSarbo W. (1995). *A mixture likelihood approach for generalized linear models.* Journal of Classification **12**, 21 – 55.

[15] Wedel M., Kamakura W. (2001). *Market segmentation — conceptual and methodological foundations.* Kluwer Academic Publishers.

[16] Yakowitz S., Spragins J. (1968). *On the identifiability of finite mixtures.* The Annals of Mathematical Statistics **39** (1), 209 – 214.

Acknowledgement: This piece of research was supported by the Austrian Science Foundation (FWF) under grant SFB#010 ('Adaptive Information Systems and Modeling in Economics and Management Science').

Address: B. Grün, F. Leisch, Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Wien, Austria

E-mail: `Bettina.Gruen@ci.tuwien.ac.at`, `Friedrich.Leisch@ci.tuwien.ac.at`