LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Friedrich Leisch

Modelling Background Noise in Finite Mixtures of Generalized Linear Regression Models

Technical Report Number 037, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Modelling Background Noise in Finite Mixtures of Generalized Linear Regression Models

Friedrich Leisch

Department of Statistics, Ludwig-Maximilians-Universität München,
Ludwigstrasse 33, 80539 München, Germany,
Friedrich.Leisch@stat.uni-muenchen.de

This is a reprint of an article that has appeared in: Paula Brito, editor, *Compstat 2008-Proceedings in Computational Statistics*. Physica Verlag, Heidelberg, Germany, 2008, pages 385–396.

Abstract. In this paper we show how only a few outliers can completely break down EM-estimation of mixtures of regression models. A simple, yet very effective way of dealing with this problem, is to use a component where all regression parameters are fixed to zero to model the background noise. This noise component can be easily defined for different types of generalized linear models, has a familiar interpretation as the empty regression model, and is not very sensitive with respect to its own parameters.

Keywords: mixture models, generalized linear models, robust statistics, R

1 Introduction

Finite mixture models have been used for more than 100 years, but have seen a real boost in popularity over the last decades due to the tremendous increase in available computing power. The areas of application of mixture models range from biology and medicine to physics, economics and marketing. On the one hand these models can be applied to data where observations originate from various groups and the group affiliations are not known, and on the other hand to provide approximations for multi-modal distributions (Everitt & Hand (1981), Titterington et al (1985); McLachlan & Peel (2000)).

In the 1990s finite mixture models have been extended by mixing standard linear regression models as well as generalized linear models (Wedel & DeSarbo (1995)). An important area of application of mixture models and also of these extensions are in market segmentation (Wedel & Kamakura (2001)), where finite mixture models replace more traditional cluster analysis and cluster-wise regression techniques as state of the art.

For mixtures without a regression part, i.e., model-based clustering, several authors have investigated the effect of outliers on parameter estimates,

and how outliers can be treated to get more robust behaviour. A comprehensive theoretical analysis for breakdown points of ML-estimators of location-scale mixtures can be found in Hennig (2004). Suggested solutions for robustification against outliers include

1. to add a noise component which is either uniform over the convex hull of the complete data set (Banfield & Raftery (1993)), or an improper constant uniform (Hennig & Coretto (2007)),
2. replace Gaussian densities with t-densities (Mclachlan & Peel (2000)), and
3. trimming observations (Cuesta-Albertos et al (1997)).

In this paper we present a new noise component to model outliers and show that our approach combines several aspects of the above. In addition, it can be easily extended to mixtures of regression models and has a natural interpretation in this context as the null model of no interaction between predictors and response.

2 Mixtures of GLMs

Consider finite mixture models with K components of form

$$h(y|x,\psi) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \tag{1}$$

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$

where y is a (possibly multivariate) dependent variable with conditional density h, x is a vector of independent variables, π_k is the prior probability of component k, θ_k is the component specific parameter vector for the density function f, and $\psi = (\pi_1, , \ldots, \pi_K, \theta_1', \ldots, \theta_K')'$ is the vector of all parameters.

If f is a univariate normal density with component-specific mean $\mu_k(x) = \alpha_k + \beta_k' x$ and variance σ_k^2, we have $\theta_k = (\alpha_k, \beta_k', \sigma_k^2)'$ and Equation (1) describes a mixture of standard linear regression models, also called *latent class regression*. If f is a member of the exponential family, we get a mixture of generalized linear models. For multivariate normal f and $x \equiv 1$ we get a mixture of Gaussians without a regression part (model-based clustering).

The posterior probability that observation (x, y) belongs to class j is given by

$$\mathbb{P}(j|x,y,\psi) = \frac{\pi_j f(y|x,\theta_j)}{\sum_k \pi_k f(y|x,\theta_k)} \tag{2}$$

The posterior probabilities can be used to segment data by assigning each observation to the class with maximum posterior probability. In the following

we will refer to $f(\cdot|\cdot, \theta_k)$ as *mixture components* or *classes*, and the groups in the data induced by these components as *clusters*.

The log-likelihood of a sample of N observations $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ is given by

$$\log L = \sum_{n=1}^{N} \log h(y_n|x_n, \psi) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f(y_n|x_n, \theta_k) \right) \tag{3}$$

and can usually not be maximized directly. The most popular method for maximum likelihood estimation of the parameter vector ψ is the iterative expectation-maximization algorithm (EM, Dempster et al (1977)):

Estimate the posterior class probabilities for each observation

$$\hat{p}_{nk} = \mathbb{P}(k|x_n, y_n, \hat{\psi})$$

using Equation (2) and derive the prior class probabilities as

$$\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^{N} \hat{p}_{nk}$$

Maximize the log-likelihood for each component separately using the posterior probabilities as weights

$$\max_{\theta_k} \sum_{n=1}^{N} \hat{p}_{nk} \log f(y_n|x_n, \theta_k) \tag{4}$$

The E- and M-steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations is reached.

Parameter estimates in standard linear models with Gaussian errors and most other GLMs are rather sensitive to outliers, because the maximum likelihood estimate is basically a mean value, which is not a robust statistic. For mixtures of regression models the problem is even more pronounced, because the variance is no longer a nuisance parameter, it needs to be estimated to compute likelihoods and posterior probabilities in each EM iteration.

One solution would be to use robust regression in the M-step, however this would violate the EM principle as the resulting estimates are no longer maximum likelihood estimates. Hence, convergence is no longer guaranteed even for clean data. In addition we run into the problem that robust estimates ususally themselves are computationally very demanding, we need estimates for every component in every EM-iteration, and convergence of EM is usually rather slow. Hence, we would need to compute expensive estimates very often.

3 Modelling background noise

Outliers or background noise can be modeled by adding a noise component f_0 to our mixture model from Equation 1:

$$h(y|x,\psi) = \pi_0 f_0(y|x,\theta_0) + \sum_{k=1}^{K} \pi_k f(y|x,\theta_k) \tag{5}$$

$$\pi_k \geq 0, \quad \sum_{k=0}^{K} \pi_k = 1$$

In the following we will call f_0 the *noise component*, and the remaining components for $k = 1, \ldots, K$ the *regular components*.

Banfield & Raftery (1993) and Hennig & Coretto (2007) use a uniform distribution for f_0, the main difference is that the former estimate the range of the uniform from the data, while the latter use either an improper uniform with pre-specified fixed value for the height of the density, or an ML estimate for the complete mixture including the noise component. Both consider only the case of model-based clustering, i.e., no regression.

3.1 Gaussian response

For mixtures of regression models there is a natural other candidate for the noise component, the null model which assumes no relationship between predictors x and response y. For notational simplicity, consider for the moment standard linear regression models with Gaussian noise, such that

$$f(y|x,\theta_k) = \phi\left(\frac{y-\mu_k(x)}{\sigma_k}\right) = \phi\left(\frac{y-\alpha_k-\beta_k' x}{\sigma_k}\right)$$

where $\phi(\cdot)$ denotes the density of the standard normal distribution. Using a noise component of form

$$f_0(y|x,\theta_0) = f_0(y|\theta_0) = \phi(\left(\frac{y-\mu_0}{\sigma_0}\right)$$

means we add a compontent corresponding to an empty regression model of form $y = \mu_0 + \epsilon$.

There are three possible ways to define the noise parameters μ_0 and σ_0:

NP1: set to fixed values in advance based on expert opinion,
NP2: estimate from data but hold fixed during EM iterations, e.g., to mean and standard deviation of y, or
NP3: treat f_0 as a regular mixture component and estimate its parameters by EM together with all other parameters of the model.

Obvously NP1 is the most robust variant, because it does not depend on the data at all. However, our simulations show that NP2 is also very robust, so we consider only the data-driven solutions NP2 and NP3 for the remainder of this paper.

Using an empty regression model as noise component has several attractive features: The noise component has the same functional form as the other components, so it is particularly easy to implement in software given the rest of the mixture model, see Section 4. There is also a natural interpretation of parameter π_0, which is the probability that an observation originated from the empty model. This is closely related to popular statistics of standalone regression models such as R^2 or analysis of variance F, which also compare a regression model with the empty model.

The effects of including the noise component can easily be seen by taking a look at the posterior probabilities (4). If we fix $\sigma_0^2 \equiv \mathrm{var}(y)$, then

$$\sigma_0 \geq \sigma_k, \quad k = 1, \ldots, K$$

with (approximate) equality only for components where $\beta_k \approx 0$, and usually all σ_k are smaller than σ_0. Hence, the posterior probability of the noise component equals the ratio of a normal density with large density to the sum of several normal densities, see Equation 2.

Figure 1 shows examples for $\sigma_0 = 2\sigma_1$, $\sigma_0 = 4\sigma_1$, and $\sigma_0 = 8\sigma_1$. The posterior probabilities of the noise component are larger than 0.99 outside the interval $[-4, 4]$, and larger than 0.9 outside of $[-3, 3]$. Observations which are further than 4 standard deviations away from a regular mixture component have zero weight in the M step in Equation 4 of the EM-procedure.

Choosing a Gaussian noise component rather than a uniform makes no large difference in which observations are marked as outliers. If σ_0 is large (as intended), then the Gaussian is very flat and over the main part it is very similar to the uniform. The big advantage is that the support of the Gaussian is unbounded, although it will become very small outside of, say, $\mu_0 \pm 4\sigma_0$. However, the weights used in (4) are ratios of densities (2), and due to the larger variance the density of the Gaussian noise component will always be much larger than the densities of the regular components in regions far away from the center. Thus, we knock out outliers everywhere except for the main support regions of the regular components. For uniforms, we need to solve the ill-conditioned estimation problem of the boundaries of the uniform distribution, see Hennig & Coretto (2007) for a detailed discussion. For the Gaussians exact estimation of variance is not really critical (a rather unusual situation!), Figure 1 shows that the value of σ_0 has not much influence on which observations are marked as outliers. Preliminary simulations studies (not shown here) confirm this bahaviour.

3.2 Other GLMs

The same form of noise component can easily be used in other continous members of the exponential family, as well as in some discrete distributions like the Poisson. Due to the limited space of this conference paper we cannot give full formulas or examples. The basic principle is always to have the null model with no regression part as noise component, and estimate the parameters of the noise component from the complete data set.

E.g., an exponential distribution with a large and constant mean value gives a noise component with a rather flat density on $\mathbb{R}^+$, which downweights large outliers, similar for the gamma distribution. For Poisson responses one can use overdispersed quasi-Poisson noise components. It is not so clear how the concept can be used for GLMs for categorical data (binomial, multinomial), but in this case even the definition of "outliers" or "background noise" is problematic.

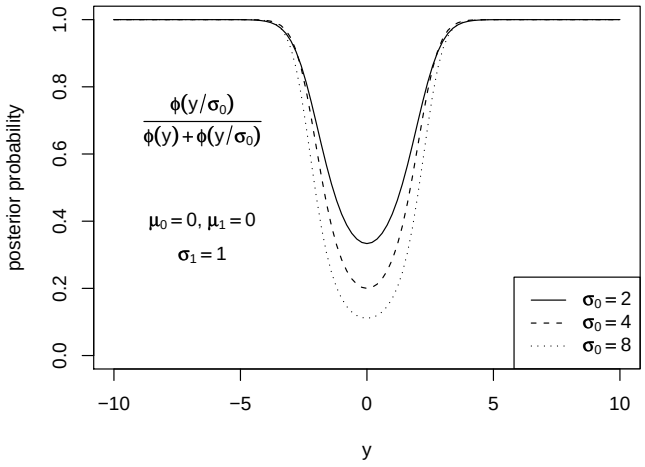

Fig. 1. Posterior probability of the noise component.

4 Software implementation

All simulation results shown below were computed using R (R Development Core Team (2007)) extension package `flexmix` (Leisch (2004), Gruen & Leisch (2007)). The standard driver for mixtures of GLMs in `flexmix`

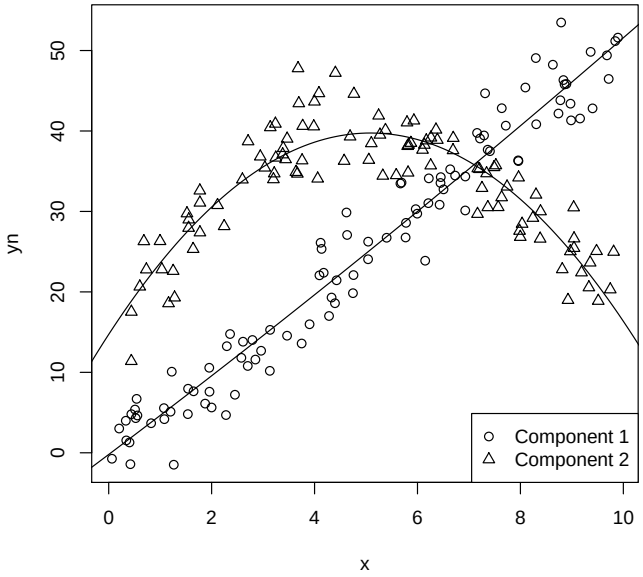

Fig. 2. A two component mixture regression example. The lines correspond to the fitted values of a model estimated with the EM algorithm.

is `FLXMRglm`. The new extension fixes the first component to be the noise component, and dispatches to the standard driver for the rest. The current development version of the software can be obtained from the author upon request and will be released on CRAN (`http://cran.r-project.org`) as part of `flexmix` later this year.

It allows to estimate the parameters of the noise component either fixed from the complete data set, in which case only π_0 is estimated by maximum likelihood, or by weighted maximum likelood with weights proportional to the probability of being a member of the noise component. The latter approach has the advantage that the null model can be interpreted at par with the regular components, but is not robust against outliers which are located close to each other.

5 Artificial example

First we consider a simple example introduced by Leisch (2004) with two latent classes of size 100 each:

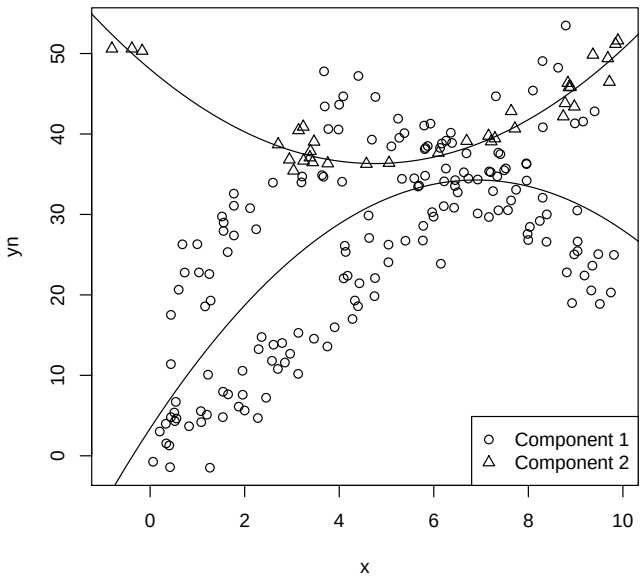

Fig. 3. The same data set as in Figure 2 with three outliers. The lines correspond to the best model found by EM, which is completely broken.

$$\text{Class 1: } y = 5x + \epsilon$$
$$\text{Class 2: } y = 15 + 10x - x^2 + \epsilon$$

with $\epsilon \sim N(0, 9)$ and prior class probabilities $\pi_1 = \pi_2 = 0.5$. The data set can be loaded into R with the command `data("NPreg", package="flexmix")`. The result of fitting a mixture model of with two quadratic polynomial components to the data can be seen in Figure 2.

If we add three outliers on the top left corner to the data set, EM estimation breaks down and gives completely wrong results, see Figure 3. Note that this is the result with the best likelihood of 20 replications of the EM algorithm, and not simply a problem of convergence in a local minimum. Estimating the model with an additional noise component correctly identifies the three outliers with posterior probabilities numerically equal to 1. As a result, estimation of the two regular components is now correct again, see Figure 4.

Mean and variance of the noise component were fixed to the corresponding empirical estimates from the response variable. If we have only a few outliers in the same spot, we cannot reliably estimate the parameters μ_0 and σ_0 by EM. Another situation is shown in Figure 5, where 20 uniform noise

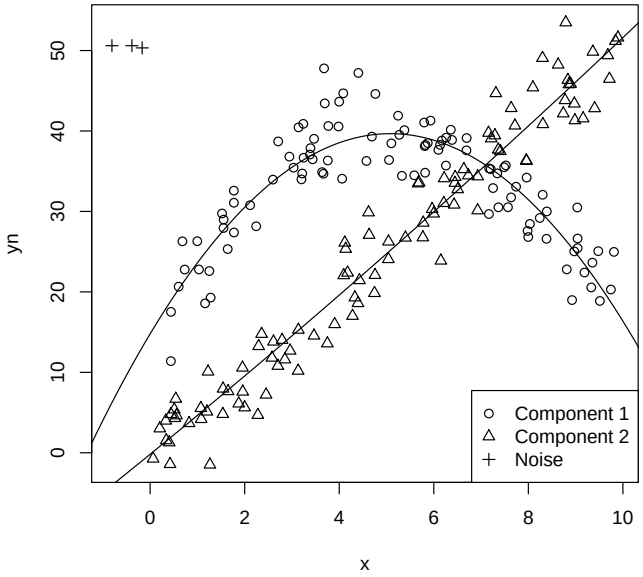

Fig. 4. The same data set as in Figure 3 using a model with a noise component. The three outliers are correctly identified.

observations have been added on a rectangle that is larger than the original data range. Again, outliers not located in the main part of the original data set are correctly identified, and both the linear and the parabolic components were almost exactly identified. There is now a little bit more curvature in the fitted model for the linear class, but note that both components have a linear model with parameter estimates for intercept, x and x^2. It is impossible to distinguish original data points from background noise that is located close to the original data, so some effect is to be expected.

6 Simulation study

We also conducted several simulation studies to see whether it makes a huge difference if we estimate the parameters of the noise component by NP2 or NP3 for the case of uniform background noise. We fixed the data set described above and added $0, 5, 10, \ldots, 50$ noise observations from a uniform distribution on $[-5, 15] \times [-10, 60]$ in the same way as we did in Figure 5. For each number of noise points we drew 100 data sets, ran the EM algorithm 5 times on each and kept only the best model to avoid local minima. The

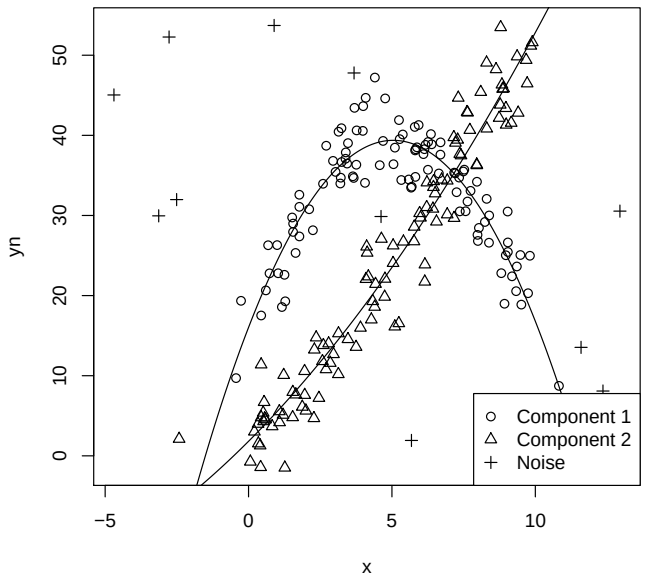

Fig. 5. The same data set as in Figure 2 with 20 outliers distributed uniformly on $[-5, 15] \times [-10, 60]$.

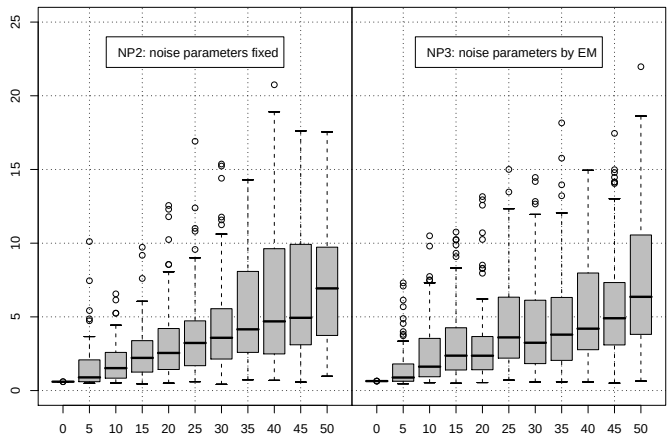

Fig. 6. Distance between estimated and true parameter values for data sets with 0–50 uniform background noise values.

estimated regression coefficients of the mixture models where then compared to the true parameter values.

Figure 6 shows boxplots of the Euclidean distance between estimated and true parameters. Without noise ("zero points added") EM converged to the same solution all the time, these values can be used as reference baseline. As expected, estimation error increases when more and more noise points are added, but there is no large difference between schemes NP2 and NP3. NP2 seems to be slightly better for fewer outliers, while NP3 is slightly better for more outliers. There are 2 components with 3 regression coefficents each, i.e., a total of 6 coefficients. Estimation errors range between almost zero to a median of about 7 for 50 noise points. If we divide this by the number of coefficients, we get an average error of $7/6 \approx 1.1$ per coefficient. This is not too bad, considering that 20% of the complete data set are noise and the sample size is not that large.

If we fit a mixture model without noise component, we get a median error of about 7 if we add only 5 noise points, and a median error of 15 for 10 noise points. In both cases variation is very large and EM often gets stuck in bad solutions like Figure 3. For more than 10 noise points EM estimation breaks completely down and yields only random results with median errors of 45 and larger. Thus, by using a noise component, we can add 10 times as many noise points for comparable increase in estimation error. Simulations with other data sets of different size, dimension and number of mixture components showed similar results.

7 Outlook

We have successfully applied the proposed methodology in a consulting project modelling customer satisfaction. The data are surveys of tourists rating Austrian alpine skiing resorts. Each respondent rated dozens of detailed aspects of the resort (quality of slopes, lifts, restaurants, entertainment, ...), the task was to identify which items had a strong impact on the overall satisfaction. A global model for all tourists makes no sense, as different subgroups of the tourist population will have different preferences. For most tourists it can be assumed, that only few items have a strong impact on overall satisfaction, the remainder being more or less noise.

We are currently working on a systematic benchmark study to confirm the findings of our preliminary simulations studies like the one presented above. This also includes GLMs with other response distributions, which were only discussed shortly in this paper due to space limitations. Another line of research is to see how other approaches presented in the literature for model based-clustering can be adapted to the case of mixtures of regression models. E.g., it should be rather straightforward to replace the normal distribution with a t-distribution if the degrees of freedom are fixed in advance.

Acknowledgements

Flexmix is joint work with Bettina Grün. This research was supported by the Austrian Science Foundation (FWF) under grant P17382.

References

BANFIELD, J. D. and RAFTERY, A. E. (1993): Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49(3), 803–821.*

CUESTA-ALBERTOS, J. A., GORDALIZA, A. and MATRAN, C. (1997): Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, *25(2), 553–576.*

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.*

EVERITT, B. S. and HAND, D. J. (1981): *Finite Mixture Distributions.* London: Chapman and Hall.

GRÜN, B. and LEISCH, F. (2007): Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, *51(11), 5247–5252.*

HENNIG, C. (2004): Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, *32(4), 1313–1340.*

HENNIG, C. and CORETTO, P. (2007): The noise component in model-based cluster analysis. In: *Proceedings of GfKl-2007.* Springer Verlag, Studies in Classification, Data Analysis, and Knowledge Organization.

LEISCH, F. (2004): FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, *11(8), 1–18.*

MCLACHLAN, G. and PEEL, D. (2000): *Finite Mixture Models.* John Wiley and Sons Inc.

R Development Core Team (2007): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

TITTERINGTON, D., SMITH, A. and MAKOV, U. (1985): *Statistical Analysis of Finite Mixture Distributions.* Chichester: Wiley.

WEDEL, M. and DESARBO, W. S. (1995): A mixture likelihood approach for generalized linear models. *Journal of Classification*, *12, 21–55.*

WEDEL, M. and KAMAKURA, W. A. (2001): *Market Segmentation - Conceptual and Methodological Foundations.* Kluwer Academic Publishers, Boston, MA, USA, 2nd edition.