LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Bettina Grün & Friedrich Leisch

Finite Mixtures of Generalized Linear Regression Models

Technical Report Number 013, 2007
Department of Statistics
University of Munich

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

Finite Mixtures of Generalized Linear Regression Models

Bettina Grün[1] and Friedrich Leisch[2]

[1] Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, Wiedner Hauptstraße 8-10/1071, A-1040 Wien, Österreich
Bettina.Gruen@ci.tuwien.ac.at

[2] Institut für Statistik, Ludwig-Maximilians-Universität München, Ludwigstraße 33, D-80539 München, Deutschland
Friedrich.Leisch@stat.uni-muenchen.de

Summary. Generalized linear models have become a standard technique in the statistical modelling toolbox for investigating relationships between variables. The assumption of homogeneity of regression coefficients over all observations can be relaxed by incorporating generalized linear models into the finite mixture framework.

The model class consisting of finite mixtures of generalized linear models is presented. Model identification is discussed given that difficulties might be encountered due to trivial and generic identifiability problems. These problems have already been observed for mixtures of distributions, but the extension to mixtures of regression models introduces additional identifiability problems. Details on model estimation are given and the application is illustrated on several examples.

Key words: finite mixture models, generalized linear models, unobserved heterogeneity

1 Introduction

Finite mixture models have now been used for more than hundred years (Newcomb, 1886; Pearson, 1894). They are a very popular statistical modelling technique given that they constitute a flexible and easily extensible model class for (1) approximating general distribution functions in a semi-parametric way and (2) accounting for unobserved heterogeneity. The number of applications has tremendously increased in the last decades as model estimation in a frequentist as well as a Bayesian framework has become feasible with the nowadays easily available computing power.

The simplest finite mixture models are finite mixtures of distributions which are used for model-based clustering. In this case the model is given by a convex combination of a finite number of different distributions where each of the distributions is referred to as component. More complicated mixtures have

been developed by inserting different kinds of models for each component. An obvious extension is to estimate a generalized linear model (GLM, McCullagh and Nelder, 1989) for each component. Finite mixtures of GLMs allow to relax the assumption that the regression coefficients and dispersion parameters are the same for all observations. In contrast to mixed effects models, where it is assumed that the distribution of the parameters over the observations is known, finite mixture models do not require to specify this distribution a-priori but allow to approximate it in a data-driven way.

In a regression setting unobserved heterogeneity for example occurs if important covariates have been omitted in the data collection and hence their influence is not accounted for in the data analysis. In addition in some areas of application the modelling aim is to find groups of observations with similar regression coefficients. In market segmentation (Wedel and Kamakura, 2001) one kind of application among others of finite mixtures of GLMs aims for example at determining groups of consumers with similar price elasticities in order to develop an optimal pricing policy for a market segment.

Other areas of application are biology or medicine, see Aitkin (1999); Follmann and Lambert (1989); Wang et al (1996); Wang and Puterman (1998). An example for a biological application is illustrated by the "Aphids" data set in Boiteau et al (1998). The data contains the results of 51 independent experiments in which varying numbers of aphids were released in a flight chamber containing 12 infected and 69 healthy tobacco plants. After 24 hours, the flight chamber was fumigated to kill the aphids, and the previously healthy plants were moved to a greenhouse and monitored to detect symptoms of infection. The number of plants displaying such symptoms was recorded. The relationship between the proportion of infected plants given the number of released aphids is depicted in Figure 1.

Clearly the proportion of infected plants in dependence of the number of released aphids does not cluster around a single regression line, but around two different regression lines. For one regression line no infection takes place while for the other the proportion of infected plants increases with the number of aphids. Fitting a finite mixture of binomial logit models allows to determine the expected number of infected plants given the number of released aphids for each of the components and the proportion of times where no infection takes place.

In Section 2 the finite mixture model of GLMs is specified starting with the standard GLM formulation. The general mixture model class is presented and several special cases which are included in this model class are discussed. In Section 3 the identifiability of finite mixtures of GLMs is analyzed and sufficient conditions to guarantee "generic" identifiability are given. As additional problems to the case of finite mixtures of distributions can occur in the regression setting, model identification has to be again investigated and results obtained for mixtures of distributions can not be directly transferred without further consideration. After an outline of model estimation using the EM algorithm and a brief overview on Bayesian methods in Section 4 the

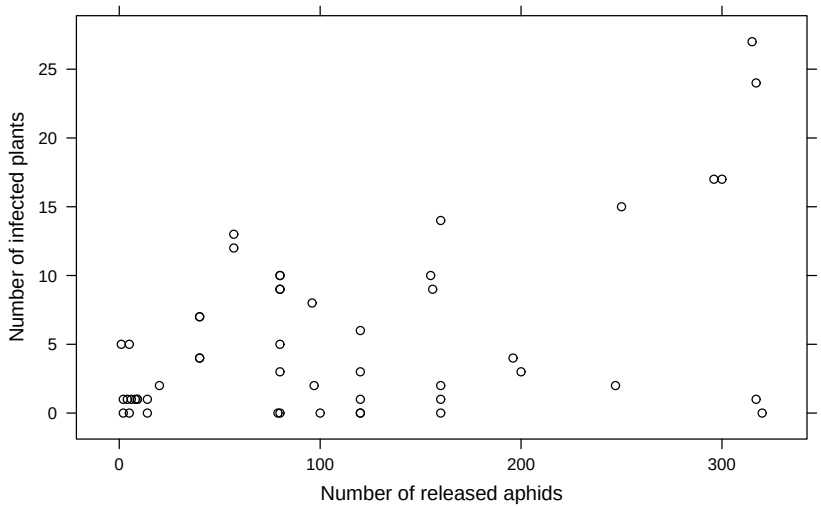

Fig. 1. "Aphids" data set.

application of the model class is illustrated in a cluster-wise regression setting as well as in a situation where overdispersion in a Poisson standard GLM is observed and a random intercept model is fitted to account for this overdispersion. An outlook on several possible extensions is given in the last section. All computations and graphics in this paper have been done using package **flexmix** (Leisch, 2004b; Grün and Leisch, 2006, 2007) in R, an environment for statistical computing and graphics (R Development Core Team, 2007).

2 Model specification

In the standard linear model the dependent variable y is assumed to follow a Gaussian distribution where the mean value is determined through a linear relationship given the covariates x:

$$\mathbb{E}[y|x] = x'\beta,$$

where β are the regression coefficients. This signifies that $y|x \sim N(x'\beta, \sigma^2)$.

The assumption that the dependent variable follows a Gaussian distribution is relaxed in the generalized linear model framework. The distribution of the dependent variable is assumed to be from the exponential family of distributions (e.g. Gaussian, binomial, Poisson or gamma). This allows to take certain data characteristics into account such as that the dependent variable y is for example a counting variable with values in $\mathbb{N}$ which is then in general assumed to follow a Poisson distribution.

The density of a distribution from the exponential family is given by

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi) + c(y,\theta)}\right\}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. For the Gaussian distribution $N(\mu,\sigma^2)$ with mean μ and variance σ^2 and the assumption that $\theta = \mu$ and $\phi = \sigma^2$ these functions are for example given by

$$a(\phi) = \phi, \qquad b(\theta) = \frac{\theta^2}{2}, \qquad c(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\phi} + \log(2\pi\phi)\right\}.$$

The relationship between the linear predictor η and the expected value μ of the dependent variable y is modelled via a link function

$$\eta = g\left(\mathbb{E}[y|x]\right) = x'\beta,$$

where η is the linear predictor and $g(\cdot)$ the link function. Different link functions are possible. A special link function is the canonical link which is given by

$$\eta = x'\beta = \theta.$$

For the Gaussian distribution the identity function is the canonical link, for the Poisson the log function, for the binomial the logit function and for the gamma distribution the reciprocal function.

The GLM framework is embedded in the finite mixture framework by inserting GLMs into the components. The resulting models are also referred to as GLIMMIX models (Wedel and Kamakura, 2001). A finite mixture density of GLMs with K components is given by

$$h(y|x,\Theta) = \sum_{k=1}^{K} \pi_k f_k(y|x,\theta_k)$$

where Θ denotes the vector of all parameters for the mixture density $h()$. The dependent variable is y and the independent variables are x. f_k is the component specific density function which is assumed to be univariate and from the exponential family of distributions. The component specific parameters are given by $\theta_k = (\beta_k', \phi_k)$ where β_k are the regression coefficients and ϕ_k is the dispersion parameter. The mean of each component is given by

$$\mu_k(x) = g_k^{-1}(x'\beta_k),$$

where $g_k()$ is the component specific link function.

For the component weights π_k it holds

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad \pi_k > 0, \forall k. \tag{1}$$

Several special cases and extensions of this model class exist. Often it is assumed that the component specific densities are from the same parametric family for each component, i.e. $f_k \equiv f$ for notational simplicity, and that the link function is also the same for all components ($g_k \equiv g$). In a cluster-wise regression setting this will be an obvious model choice as no a-priori knowledge about differences in distributional families of the components is available. Another popular extension is to have a so-called concomitant variable model for the prior class probaibilities, such that the π_i also depend on a set of explanatory variables (e.g., using a multinomial logit model).

A special case where different component specific distributions are used is a model where only a single component is specified to follow a different distribution in order to allow this component to capture outlying observations (Dasgupta and Raftery, 1998). This approach is similar to the specification of zero-inflated models (Böhning et al, 1999). Even though the component specific densities are in general assumed to be from the same parametric family (e.g. Poisson or binomial), the parameters are fixed a-priori for one component such that this component absorbs all excess zeros in the zero-inflated model.

In order to decrease the number of parameters equality constraints can be imposed over the components for a subset of the component specific parameters θ_k. A special case are random intercept models where only the intercept follows a finite mixture distribution while all other regression coefficients are constant over the components, see Follmann and Lambert (1989). These models are often used if overdispersion is encountered in Poisson or binomial GLMs in order to determine a model which describes the data in an appropriate way.

3 Identification

Statistical models are in general represented by parameter vectors. For finite mixture models the parameter vector Θ which consists of the component weights and the component specific parameters determines a mixture distribution, i.e. there is a mapping from the parameter space to the model space. For identifiability this mapping has to be injective, i.e. for each model in the model space there is a unique parameter vector in the parameter space which is mapped to the model. Lack of identifiability can be a problem for model estimation or if parameters are interpreted.

In the following let Ω denote the space of admissible parameters for K-component mixtures where the following conditions are fulfilled

- $\pi_k > 0 \; \forall k = 1, \ldots, K$, and
- $\forall k, l \in \{1, \ldots, K\}$: $k \neq l \Rightarrow \theta_k \neq \theta_l$.

These two conditions prevent overfitting and identifiability problems which occur due to empty components where θ_k cannot be uniquely determined and due to components with equal component parameter vectors where different values for π_k are possible.

Let $\mathscr{A}_K = \mathscr{A}_K(f, \Omega)$ be the set of all finite mixture models with K components, component specific density function f and mixture densities of form $h(\cdot|\cdot, \Theta)$, $\Theta \in \Omega$. Each parameter vector $\Theta \in \Omega$ corresponds to one model $a \in \mathscr{A}_K$, but each model a has at least $K!$ parameterizations Θ due to all possible permutations of the components, also known as *label switching* (Redner and Walker, 1984).

$\mathscr{A}_K$ induces a system of equivalence classes Ξ on Ω where two elements of Ω are in the same equivalence class if they correspond to the same model a:

$$\Theta, \tilde{\Theta} \in \Xi \Leftrightarrow h(\cdot|\cdot, \Theta) \equiv h(\cdot|\cdot, \tilde{\Theta}).$$

The usual definition of model identifiability is that either all equivalence classes contain only one element (which is trivially not true for mixture models), or that at least a unique representative for each equivalence class can be selected.

Let $\text{ident}(\Omega) \subset \Omega$ be the subset of parameterizations which contain only one permutation of each possible set of component parameters. $\text{ident}(\Omega)$ can be obtained from Ω by imposing an ordering constraint on the components with respect to a certain parameter (or a combination of several parameters). We refer to any identifiability problems which are present for $\text{ident}(\Omega)$ as *generic* (Frühwirth-Schnatter, 2006).

3.1 Generic identifiability

Generic identifiability problems have already been analyzed for finite mixtures of distributions by Titterington et al (1985). In nearly all cases only mixtures where the component distribution is from the same distributional family have been considered. General results for certain kinds of distribution as well as specific results for a given component specific distribution function have been derived. Generic identifiability is guaranteed for important continuous distributions such as the Gaussian, gamma and Poisson distribution. A special case are finite mixtures of binomial distributions which are only identifiable if the number of components is limited. For the model class of finite mixtures of binomial distributions $\text{Bi}(\pi, T)$ with success probability π and repetition parameter T a necessary and sufficient condition for identifiability is $T \geq 2K-1$.

The analysis of identifiability of mixtures of Gaussian regression models revealed that requiring a covariate matrix of full rank – as postulated previously for example by Wang and Puterman (1998) – is not sufficient (Hennig, 2000). Contrarily, it is necessary to check a coverage condition in order to ensure identifiability. With respect to generic identifiability of finite mixtures of regression models three influencing factors can therefore be distinguished:

- component distribution f,
- covariate matrix and
- repeated observations/labelled observations.

Repeated observations where the class membership is fixed are necessary for mixtures of binomial distributions to be identifiable. In a regression setting repetitions over different covariate points can help in making a mixture identifiable as it changes the set of feasible hyperplanes for the coverage condition. Labels for some observations indicating that they belong to the same component have the same influence.

In order to present a theorem on sufficient conditions for identifiability of finite mixtures of GLMs a data representation is necessary which takes repeated observations of the same individual where the component membership is fixed into account. The observations for an individual t are combined and given by:

$$(y_t, x_t) = (y_i, x_i)_{i \in I_t},$$

where I_t contains the set of indices corresponding to the observations of individual t. In the following X and Y denote the matrix of all x and y observations of all N individuals.

Theorem 1. *The model defined by*

$$h(Y \mid X, \Theta) = \prod_{t=1}^{N} \left[\sum_{k=1}^{K} \pi_k \prod_{i \in I_t} f(y_i | \mu_i^k, \phi_k) \right]$$

and

$$g^{-1}\left(\mu_i^k\right) = x_i' \beta_k$$

is identifiable if the following conditions are fulfilled:

1. *(a)* $\exists \tilde{I} \neq \emptyset$*:* $\tilde{I} \subseteq \bigcup_{t=1}^{N} I_t$*: The mixture of distributions given by*

$$\sum_{k=1}^{K} \pi_k f(y_i | \mu_i^k, \phi_k)$$

   is identifiable $\forall i \in \tilde{I}$*.*

   (b) $q^* > K$ *with*

$$q^* := \min \Big\{ q : \forall i^* \in \tilde{I} : \exists H_j \in \{H_1, \ldots, H_q\} : \{x_i : i \in I_{t(i^*)} \cap \tilde{I}\} \subseteq H_j \wedge H_j \in \mathcal{H}_U \Big\}$$

   where $\mathcal{H}_U$ *is the set of* $H(\alpha) := \{x \in \mathbb{R}^U : \alpha' x = \mathbf{0}\}$ *with* $\alpha \neq \mathbf{0}$*.*

2. *The matrix* X *has full column rank.*

The proof is straight-forward given the previous results for finite mixtures of standard linear regression models by Hennig (2000) and finite mixtures of GLMs and multinomial logit models with varying and fixed effects in the regression coefficients by Grün (2006); Grün and Leisch (2007).

For Condition (1a) the generic identifiability of finite mixtures with the given component specific distribution is essential. If the component specific distribution is either Gaussian, Poisson or gamma this condition is no restriction as mixtures of these distributions are generically identifiable, i.e. $\tilde{I} = \bigcup_{t=1}^{N} I_t$. In the case of the binomial distribution the repetition parameter has to be checked for each observation in order to determine if it can be included in $\tilde{I}$. Condition (1b) indicates that for each individual t there has to be one of the q hyperplanes through the origin H_j which covers all identifiable observations of this individual. The rank condition (2) ensures that the regression coefficients can be uniquely determined given the linear predictor.

These conditions indicate that identifiability problems can especially occur if the covariate matrix contains categorical variables. We refer to identifiability problems due to the violation of the coverage condition as

Intra-component label switching: If the labels are fixed in one covariate point according to some ordering constraint, then labels may switch in other covariate points for different parameterizations of the model.

For mixtures where the component distributions are identifiable this means that the component weights and possible dispersion parameters are unique, but the regression coefficients vary because they depend on the combination of the components between the covariate points. This identifiability problem is also of concern for prediction, because given the class membership the predicted value for new data depends on the chosen solution.

Unidentified mixture models with several isolated non-trivial (global) modes in the likelihood are to some extent more of a theoretical problem, because, e.g., minimal changes of the component weights π_k often make the model identified by breaking symmetry. However, models "close" to an unidentified model will have multiple local modes.

The following example presents a simple mixture of regression models with intra-component label switching. The model is unidentified (with two non-trivial modes) only if both components have exactly the same probability.

Example 1. Assume we have a mixture of standard linear regression models with one measurement per object and a single categorical regressor with two levels. The usual design matrix for a model with intercept uses the two covariate points $x_1 = (1, 0)'$ and $x_2 = (1, 1)'$. Furthermore, let the mixture consist of two components with equal component weights. The mixture regression is given by

$$h(y|x, \Theta) = \frac{1}{2} f_N(\mu_1(x), 0.1) + \frac{1}{2} f_N(y|\mu_2(x), 0.1)$$

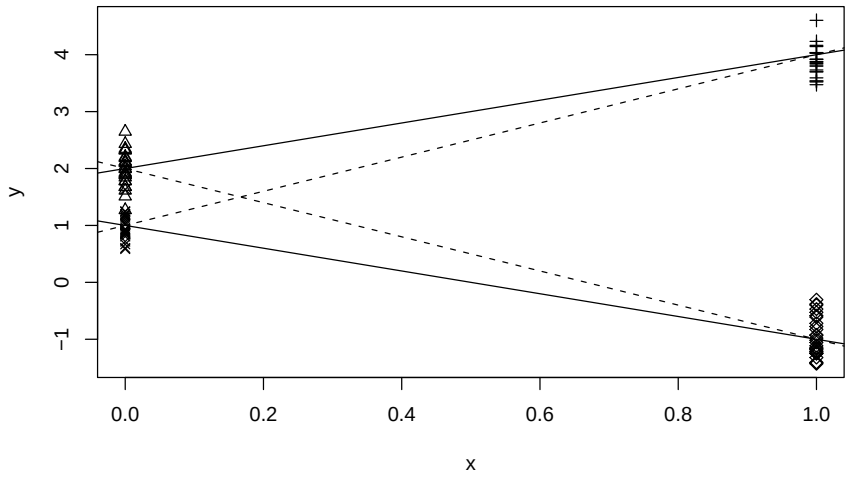

Fig. 2. Balanced sample from the artificial example with the two theoretical solutions. The solid lines correspond to solution 1 and the dashed lines to solution 2.

where $\mu_k(x) = x'\beta_k$ and $f_N(y|\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

Now let $\mu_1(x_1) = 1$, $\mu_2(x_1) = 2$, $\mu_1(x_2) = -1$ and $\mu_2(x_2) = 4$. As Gaussian mixture distributions are generically identifiable the means, variances and component weights are uniquely determined in each covariate point given the mixture distribution. However, as the coverage condition is not fulfilled, the two possible solutions for β are:

Solution 1: $\beta_1^{(1)} = (2, \quad 2)'$, $\beta_2^{(1)} = (1, -2)'$
Solution 2: $\beta_1^{(2)} = (2, -3)'$, $\beta_2^{(2)} = (1, \quad 3)'$

In Figure 2 a balanced sample with 50 observations in each covariate point is plotted together with the two solutions for combining x_1 and x_2.

This mixture model would be identifiable if either

1. three different covariate points were available, or
2. observations for both covariate points for the same object were available, or
3. the component weights were unequal, e.g. $\pi_1 = 0.6$.

Condition 1 is not an option, for instance, for a single 2-level categorical regressor. Condition 2 is not possible if the categorical regressor cannot change for repeated observations of the same subject like, for instance, the gender of a person. However, when developing a suitable measurement design, the

possibility of these problems to occur should be considered in order to develop a suitable design matrix.

The identifiability conditions given in Theorem 1 have the drawback that they are only sufficient conditions for a certain model class. The conditions can therefore only indicate if the model class contains at least one single model which is not identifiable. In addition they are hard to verify in practice as it is an NP hard problem (Hennig, 2000). In general it will be of interest if a fitted model suffers from identifiability problems. This means that it has to be checked if there exist several modes of the likelihood in the parameter space ident(Ω) given data sets sampled from the fitted mixture model. In a frequentist estimation setting bootstrap methods can be used to investigate potential identifiability problems of a fitted finite mixture model, see Grün and Leisch (2004, 2007).

4 Estimation

Finite mixture models can be either estimated within a frequentist framework, within a Bayesian framework, with moment estimators (Lindsay, 1989) or by applying graphical tools (Titterington et al, 1985). An important characteristic of the estimation method is if the number of components has to be fixed a-priori or is simultaneously estimated. In the following maximum likelihood estimation with the EM algorithm is described and a short overview on Bayesian estimation using MCMC samplers is given.

4.1 Frequentist maximum likelihood with the EM algorithm

There exist different methods for frequentist estimation of finite mixture models. The most popular is the EM algorithm (Dempster et al, 1977; McLachlan and Krishnan, 1997) which aims at determining the ML estimator for a finite mixture model with a given number of components K. The EM algorithm has the advantage that it provides a general framework for estimating different kinds of mixture models as often only the M-step has to be modified if different component specific models are used. In addition, already available tools for weighted maximum likelihood estimation can be applied. Nevertheless, there are also some known disadvantages such as slow convergence or that one might get stuck in local optima, i.e. it is in general difficult to ensure that the root corresponding to the maximum likelihood estimator was detected.

The EM algorithm uses a data augmentation scheme and is a general estimation method in the presence of missing data. In the case of finite mixture models the missing data is the latent variable $D_t \in \{0,1\}^K$ for each individual t which indicates the component membership. This means that D_{tk} equals 1 if individual t is from component k and 0 otherwise. The data is therefore

augmented with estimates of the component memberships, i.e. the estimated a-posteriori probabilities $\hat{p}_{tk}$.

For simplicity of notation it is in the following assumed that the component density function $f(\cdot|\cdot)$ takes all observations from each individual as arguments. For a sample of N individuals $\{(y_1, x_1), \ldots, (y_N, x_N)\}$ the EM-algorithm is given by:

E-step: Given the current parameter estimates $\Theta^{(j)}$ in the j-th iteration, replace the missing data D_{tk} by the estimated a-posteriori probabilities

$$\hat{p}_{tk} = \frac{\pi_k^{(j)} f(y_t|x_t, \theta_k^{(j)})}{\sum_{l=1}^{K} \pi_l^{(j)} f(y_t|x_t, \theta_l^{(j)})}.$$

M-step: Given the estimates for the a-posteriori probabilities $\hat{p}_{tk}$ (which are functions of $\Theta^{(j)}$), obtain new estimates $\Theta^{(j+1)}$ of the parameters by maximizing

$$Q(\Theta^{(j+1)}|\Theta^{(j)}) = Q_1(\theta^{(j+1)}|\Theta^{(j)}) + Q_2(\pi^{(j+1)}|\Theta^{(j)}),$$

under the restriction for the component weights given in Equation (1) and where

$$Q_1(\theta^{(j+1)}|\Theta^{(j)}) = \sum_{t=1}^{N}\sum_{k=1}^{K} \hat{p}_{tk} \log(f(y_t|x_t, \theta_k^{(j+1)}))$$

and

$$Q_2(\pi^{(j+1)}|\Theta^{(j)}) = \sum_{t=1}^{N}\sum_{k=1}^{K} \hat{p}_{tk} \log(\pi_k^{(j+1)}).$$

Q_1 and Q_2 can be maximized separately. The maximization of Q_1 gives new estimates $\theta^{(j+1)}$ and the maximization of Q_2 gives $(\pi_k^{(j+1)})_{k=1,\ldots,K}$. Q_1 is maximized using weighted ML estimation of GLMs and the parameter estimates $\pi_k^{(j+1)}$ which maximize Q_2 are given by

$$\pi_k^{(j+1)} = \frac{1}{N}\sum_{t=1}^{N} \hat{p}_{tk} \quad \forall k = 1, \ldots, K.$$

Before each M-step the average component sizes (over the given data points) are checked and components which are smaller than a given (relatively) small size are omitted in order to avoid too small components where fitting problems might arise. This strategy has also been recommended for the a variant of the EM algorithm, the stochastic EM (SEM; Celeux and Diebolt, 1988), in order to determine the number of components. For the

SEM algorithm an additional step between the E- and M-step is performed where estimates for D_{kt} are determined by drawing a sample from the multinomial distribution implied by the posteriors for each observations and these estimates are then used as weights in the M-step. If the algorithm is started with too many components they will be omitted during the estimation process. The algorithm is stopped if the relative change in the likelihood is smaller than a pre-specified ϵ or the maximum number of iterations is reached.

It has been shown that the values of the likelihood are monotonically increased during the EM algorithm. This ensures the convergence of the EM algorithm if the likelihood is bounded. Unboundedness of the likelihood, however, might occur at the edge of the parameter space, e.g., if the variance of one component tends to zero for mixtures of Gaussian distributions. As even in the case of boundedness only the detection of a local maximum can be guaranteed, it is in general recommended to repeat the EM algorithm with different initializations and to choose as final solution the one with the maximum likelihood. Different initialization strategies for the EM algorithm have been proposed, as its convergence to the optimal solution depends on the initialization.

4.2 Bayesian MCMC sampling

Estimation within a Bayesian framework has become popular with the advent of MCMC methods, an overview on the different sampling approaches is given in Frühwirth-Schnatter (2006, chap. 3). Gibbs sampling is the most commonly used approach and it is done by augmenting the data with the unobservable variable of class membership similar to the EM algorithm. A drawback of the Gibbs sampler is that it might fail to escape the attraction area of one mode and therefore does not explore the entire parameter space. It was therefore suggested to use Metropolis-Hastings sampling schemes. Alternatively, the permutation sampler may be used.

5 Application

Three different applications of finite mixtures of regressions are presented. As the main purpose is to illustrate the application of the model class data sets are chosen which can be easily visualized in order to facilitate the understanding of the fitted models. In two cases ("Aphids" and "Movies" data set) the presence of latent groups is assumed and clustering the observations is one of the modelling aim. The difference between the two application however is that for the "Aphids" data set the presence of two separate groups with different regression coefficients can already be visually distinguished while for the "Movies" data set no separate groups can be observed even though considerable heterogeneity in the regression coefficients is present between the observations. If a mixed-effects model was fitted to the "Movies" data set this

heterogeneity would be modelled through an a-priori specified distribution. The advantage of finite mixtures in this application are that (1) it is not required to specify the distribution for modelling heterogeneity in regression coefficients a-priori and (2) the components allow to easily inspect the range of heterogeneity present in the data. For the third data set ("Fabric faults") a random intercept model is assumed in order to account for overdispersion in the data.

5.1 Infection of tobacco plants

A finite mixture of binomial logit models is fitted to the "Aphids" data set from Section 1. The model is given by

$$h(\texttt{n.inf}|\texttt{n.aphids}, \Theta) = \sum_{k=1}^{K} \pi_k f_{\text{Bi}}(\texttt{n.inf}|\pi_k(\texttt{n.aphids}), 69),$$

where $f_{\text{Bi}}(\cdot|\pi, T)$ denotes the binomial distribution with success probability π and repetition parameter T which is in this application given by 69. `n.inf` is the number of infected plants and `n.aphids` the number of released aphids. The component specific mean value is given by

$$\text{logit}(\pi_k(\texttt{n.aphids})) = \beta_{k1} + \texttt{n.aphids}\beta_{k2}.$$

Figure 1 suggests that the number of components $K = 2$. In addition to the visual inspection the number of components can be selected by fitting mixtures with different number of components to the data and determine the model with the minimum BIC. The BIC values for the mixtures with components 1 to 5 are 424.04, 274.92, 284.18, 295.5 and 305.83 where each of the mixtures is the best result of 5 different runs with random initialization to avoid local optima. This criterion hence confirms the results of the visual inspection. The fitted regression lines for each of the components together with the data are given in Figure 3. The relative sizes π_k of the 2 components are 0.54 and 0.46.

As the repetition parameter T is equal to 69 the mixtures of binomial distributions are identifiable in each observation point for mixtures with up to 35 components as induced by the constraint $T \geq 2K - 1$. Given that observations are available for a range of different `n.aphids` values generic identifiability is guaranteed for the fitted mixtures with up to 5 components.

The suitability of the fitted mixture to induce a clustering of the data can be assessed by investigating the a-posteriori probabilities. If for each observation the maximum a-posteriori probability over all components is high the observations can be with a high confidence assigned to one of the components and hence a partitioning of the observations into K groups can be reasonably done using the fitted mixture model. For the "Aphids" data set the maximum a-posteriori probabilities have a mean of 0.98 with a standard deviation of 0.05 and a median of 1.00. This indicates that for each observation

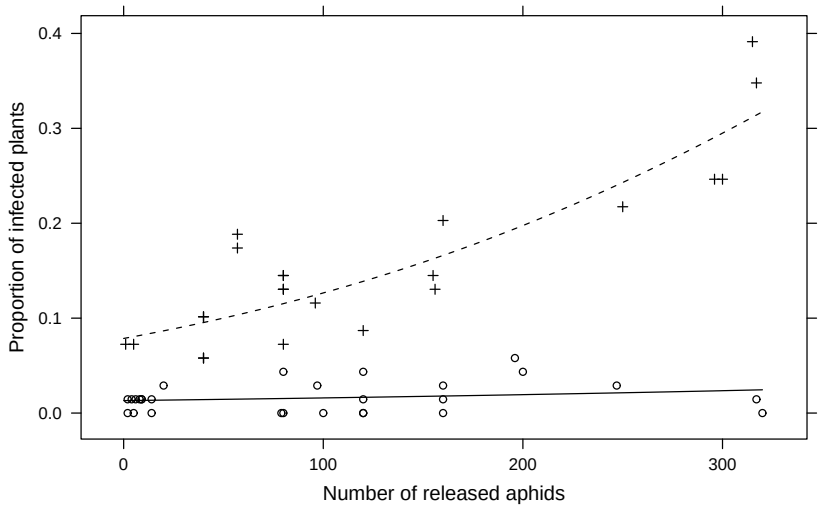

Fig. 3. "Aphids" data set with fitted regression lines for each component. The observations are plotted in different symbols according to the assignment to the component with the maximum a-posteriori probability.

(`n.inf`, `n.aphids`) it can be with high confidence decided to which component it belongs. This also means that the two components are strongly separated and in fact constitute two different regimes.

From a practitioner's point of view further investigations are important to identify reasons why and when the two different regimes emerge. One possible explanation is that some batches of aphids consisted of insects that had passed their "maiden" phase. Low or zero levels of transmission of the virus are observed in this case because after the maiden phase the aphids tend to settle on the first plant they encounter.

5.2 Market share patterns of movies

Finite mixtures of Gaussian regression models have been previously fitted to market share data of movies at the box office and theatres in the USA to investigate different patterns of decay (Jedidi et al, 1998). The box office and theaters data for 407 movies playing between May 5, 2000 and December 7, 2001 were collected from a popular website of movie records (`www.the-number.com`), see Krider et al (2005). The gross box-office takings for the 40 most popular movies for each weekend in the time period are recorded and transformed into market shares to account for the difference in volume between weekends. The market share is used as dependent variable and the number of weeks since release as covariate. For the data analysis the data is restricted to the first 20 weeks after release of a movie. This reduces

the number of movies in the data set to 394. On average 8 observations are available for each movie which gives a total of 3149 observations.

The data is given in Figure 4. Each line represents a movie and its development of market share over the weeks after release. Most of the movies have a decline in market share over the weeks, but there also some films where an increase in market share over the first weekends can be observed. Due to this opposite trends and also due to the differences in decay for the movies loosing market shares the overlap in market shares between the movies is high which renders it impossible to discern different patterns of decay.

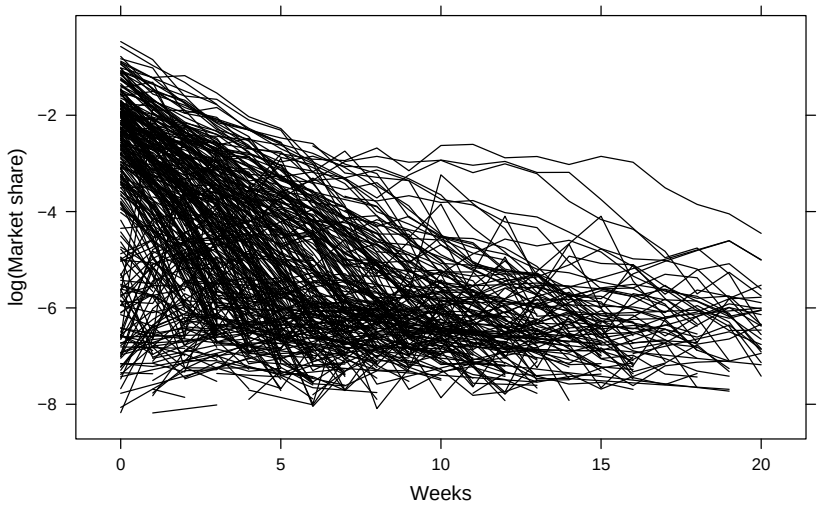

Fig. 4. Market share patterns of the "Movies" data set.

As most movies exhibit an exponential decay in market share the following mixture model is used to describe the data

$$h(\texttt{share}|\texttt{week}, \Theta) = \sum_{k=1}^{K} \pi_k f_N(\log(\texttt{share})|\mu_k(\texttt{week}), \sigma_k^2),$$

with the mean given by

$$\mu_k = \beta_{1k} + \texttt{week}\beta_{2k}.$$

As it is assumed that the component membership is fixed over the weeks for the movies, the information which observations are from the same movie is included in the estimation process.

Using an exponential decay model signifies that movies with a rise in market share at the beginning and a decline afterwards can only be approximated

through a straight line which is still reasonable considering the small recorded time interval of 20 weeks. In addition we restrict the feasible mixtures to those where all component weights are at least 0.1, i.e. each component represents 39 movies or more.

Finite mixtures with 1 to 10 components are fitted and for each number of components the EM algorithm is repeated 10 times with random initialization in order to insure that the global optimum is detected. The BIC criterion is again used to determine the optimal number of components. The BIC suggests 5 components. However, it has to be noted that even though mixtures with up to 10 components are initially specified the EM algorithm did not converge to a mixture with more than 5 components as components with a weight of less than 0.1 are omitted during the run of the algorithm.

The parameter estimates are given in Table 1. C_k indicates that the parameters in this column belong to the k^{th} component. The components are sorted in decreasing order with respect to parameter β_{1k}. The predicted mean values of market share for each component are depicted in Figure 5. The numbers indicate the component to which the line corresponds. For ease of comparison of the fitted parameters between the components they are plotted together with approximate 95% confidence intervals in Figure 6.

Parameter	C_1	C_2	C_3	C_4	C_5
π	0.15	0.17	0.23	0.13	0.32
β_{1k}	-1.99	-2.39	-2.95	-4.73	-6.49
β_{2k}	-0.29	-0.42	-0.61	-0.03	-0.01
σ	0.80	0.66	0.74	1.21	0.62

Table 1. Estimated parameters for the mixture with 5 components fitted to the "Movies" data set.

Comparing the intercepts given by β_{1k} indicates that there are three components with higher market shares at the release weekend. Components 1, 2, and 3 start with market shares of around 8.7%. The other two components achieve only market shares of 0.9% and 0.2% respectively on their release weekend. With respect to β_{2k}, which indicates the long-term success of a movie, component 3 has the strongest decline over the weeks indicating that in contrast to component 1 and 2 it is not able to stay on a high market share level for a longer time period. Component 1 seems to consist of the successful films which are also highly promoted leading to high market shares at the beginning and a slow decay over the weeks. Component 4 and 5 both have insignificant decay coefficients which indicates that they stay at about the same low level of market share during the first 20 weeks after release.

The a-posteriori probabilities are determined for each movie and used to assign them to the different components. Most of the films can be with high confidence assigned to one of the components. The mean of the maximum

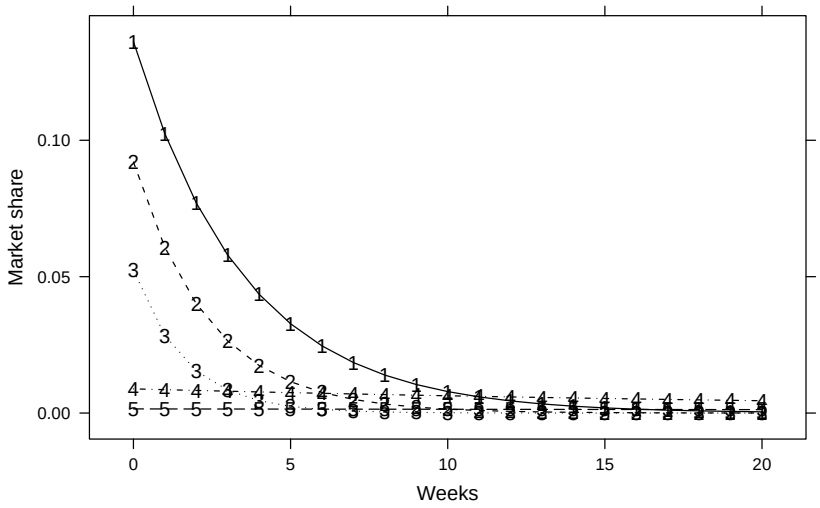

Fig. 5. Mean market share patterns of the finite mixture fitted to the "Movies" data set.

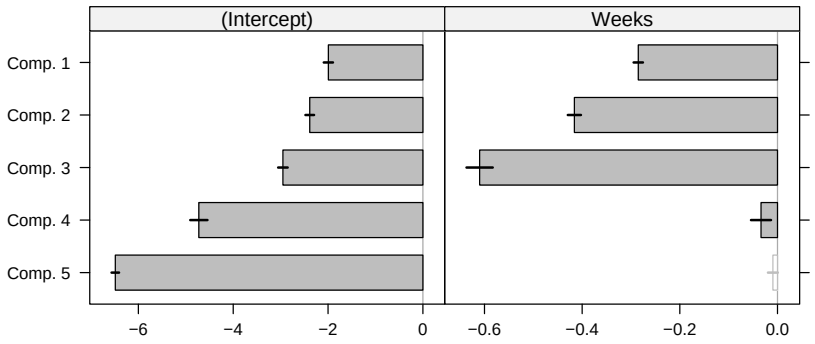

Fig. 6. Fitted regression coefficients and their approximate 95% confidence intervals for the "Movies" data set.

a-posteriori probabilities is 0.97 with a standard deviation of 0.08 and the median is 1.00. Rootograms of the posteriors of each component are given in Figure 7 (Leisch, 2004a). A rootogram is a modified version of a histogram where the square roots of the frequencies instead of the frequencies are used as heights for each bar. Please note that posteriors of less than 10^{-4} are omitted in order to ensure that the bar at zero does not dominate the plot.

The overlap of the components can be investigated by plotting the posteriors which correspond to observations assigned to a given component in a different color. If the posteriors for component 5 are highlighted it can be ob-

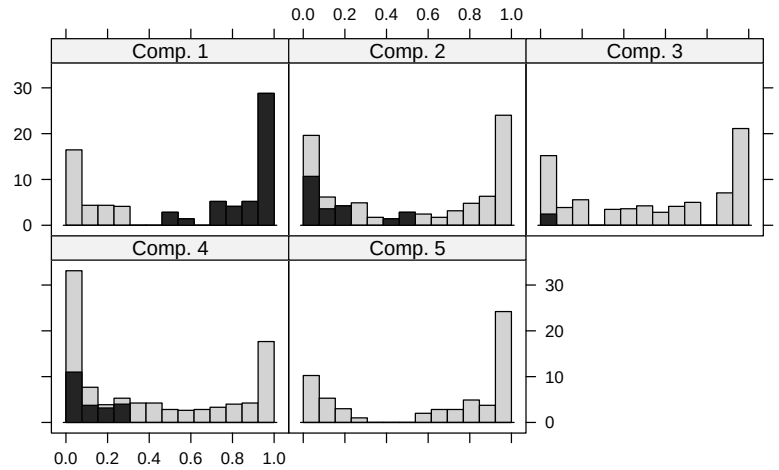

Fig. 7. Rootograms of the a-posteriori probabilities of the fitted mixture to the "Movies" data set. The posteriors of observations which are assigned to component 5 are shaded in dark grey.

served that the overlap with component 1 which consists of the most successful films is surprisingly high.

The proportions of movies assigned to each component using the maximum a-posteriori probabilities are 0.34, 0.16, 0.11, 0.23 and 0.15. The quality of the partition of the data achieved by using the fitted finite mixture model can be investigated in Figure 8 where the market share patterns of the are plotted in different panels for each cluster.

5.3 Fabric faults

The "Fabric faults" data set consists of 32 observations of number of faults in rolls of fabric of different length (Aitkin, 1996). The dependent variable is the number of faults (`n.fault`) and the covariate is the length of role in meters (`length`). The data is given in Figure 9.

As the dependent variable is a counting variable in a first step a standard GLM with Poisson distribution is fitted to the data where the logarithm of the lengths is used as independent variable. The fitted regression line is given in the left panel in Figure 10. An analysis of the model fit indicates that substantial overdispersion is present with a residual deviance of 64.54 on 30 degrees of freedom. To account for this overdispersion a random intercept model is fitted which is given by

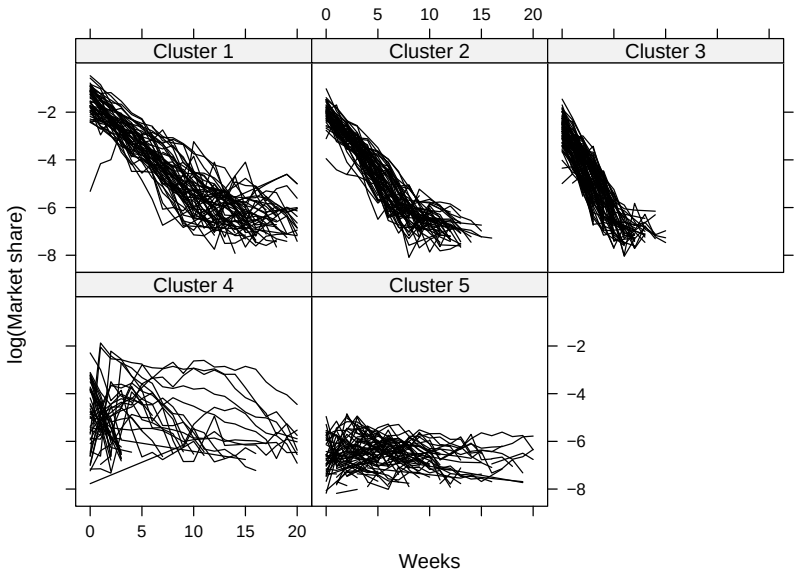

Fig. 8. Clustered market share patterns of the "Movies" data set.

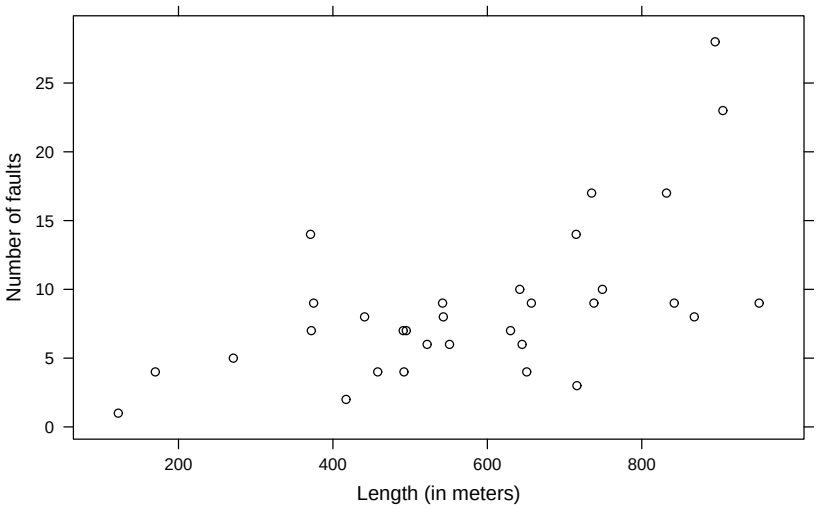

Fig. 9. "Fabric faults" data set.

$$h(\texttt{n.fault}|\texttt{length}, \Theta) = \sum_{k=1}^{K} \pi_k f_{\text{Poi}}(\texttt{n.fault}|\lambda_k(\texttt{length})).$$

where $f_{\text{Poi}}(\cdot|\lambda)$ denotes the Poisson distribution with mean λ. The mean λ_k is in the random intercept model given by

$$\log(\lambda_k) = \beta_{1k} + \log(\texttt{length})\beta_2.$$

Please note that the coefficient of the covariate does not have an index k which means that it is constant over the components.

Again the optimal number of components is selected using the BIC criterion after fitting the model with the EM algorithm for different number of components ranging from 1 to 5 and 5 repeated fittings with random initialization and the number of components fixed. The BIC values are 194.77, 186.53, 193.46, 200.39 and 207.32 and consequently the mixture with 2 components is selected. The resulting regression lines for each of the components separately are the dashed lines in the right panel of Figure 10. The full line represents the fitted regression line of the random intercept model to the complete data set. The plotting symbols of the observations in the right panel are according to an assignment of the observations to the two components given the maximum a-posteriori probabilities.

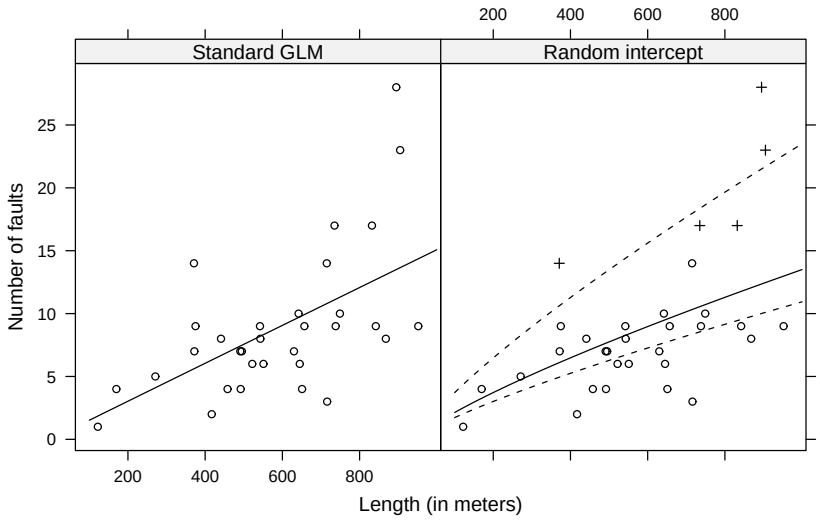

Fig. 10. Fitted regression lines to the "Fabric faults" data set for the standard GLM and a random intercept model with 2 components.

6 Conclusion and outlook

Finite mixtures of GLMs are an important statistical modelling technique which is an obvious extension of standard GLMs. They relax the assumption

of homogeneity of parameters, but do not require to a-priori specify and fix the distribution which accounts for the heterogeneity in parameters as in mixed-effects models. In addition this flexible model class contains important special cases such as zero-inflated or random intercepts models.

The model class of finite mixtures of GLMs can be easily specified within the finite mixture model framework and the modification of existing estimation methods is often straight-forward in order to be able to fit the models. For the EM algorithm it is only necessary to adapt the M-step by determining the weighted ML estimator for the component specific model. Different problems in model fitting and diagnostics than in standard mixtures of distributions however might be encountered due to trivial and generic identifiability problems.

Further extensions of finite mixtures are possible for the regression case. Instead of using GLMs as component specific models generalized additive models can be used which allow to relax the assumption that the functional relationship between covariates and dependent variable is a-priori known. Another possibility is to relax the assumption of homogeneity within the components and fit a mixed-effects model in each component.

In the future model identification and diagnostics need further investigation in the regression case for finite mixtures. The performance of newly proposed methods such as a new model selection criterion for mixtures of regression models (Naik et al, 2007) needs for example to be validated in real applications on different empirical data sets. In addition new visualization techniques which enable the researcher to easily explore the characteristics of a fitted model and compare competing models would be a valuable enhancement of the finite mixture modelling toolbox.

Acknowledgement. This research was supported by the Austrian Science Foundation (FWF) under grant P17382.

References

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. Statistics and Computing 6:251–262

Aitkin M (1999) Meta-analysis by random effect modelling in generalized linear models. Statistics in Medicine 18(17–18):2343–2351

Böhning D, Dietz E, Schlattmann P, Mendonça L, Kirchner U (1999) The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society A 162(2):195–209

Boiteau G, Singh M, Singh RP, Tai GCC, Turner TR (1998) Rate of spread of pvy-n by alate myzus persicae (sulzer) from infected to healthy plants under laboratory conditions. Potato Research 41(4):335–344

Celeux G, Diebolt J (1988) A random imputation principle: The stochastic EM algorithm. Rapports de Recherche 901, INRIA

Dasgupta A, Raftery AE (1998) Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association 93(441):294–302

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39:1–38

Follmann DA, Lambert D (1989) Generalizing logistic regression by nonparametric mixing. Journal of the American Statistical Association 84(405):295–300

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer Series in Statistics, Springer, New York

Grün B (2006) Identification and estimation of finite mixture models. PhD thesis, Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, friedrich Leisch, advisor

Grün B, Leisch F (2004) Bootstrapping finite mixture models. In: Antoch J (ed) Compstat 2004 — Proceedings in Computational Statistics, Physica Verlag, Heidelberg, pp 1115–1122

Grün B, Leisch F (2006) Fitting finite mixtures of linear regression models with varying & fixed effects in R. In: Rizzi A, Vichi M (eds) Compstat 2006—Proceedings in Computational Statistics, Physica Verlag, Heidelberg, Germany, pp 853–860

Grün B, Leisch F (2007) Flexmix 2.0: Finite mixtures with concomitant variables and varying and fixed effects. Submitted for publication

Grün B, Leisch F (2007) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects, unpublished manuscript

Grün B, Leisch F (2007) Testing for genuine multimodality in finite mixture models: Application to linear regression models. In: Decker R, Lenz HJ (eds) Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation, Springer-Verlag, Studies in Classification, Data Analysis, and Knowledge Organization, vol 33, pp 209–216

Hennig C (2000) Identifiability of models for clusterwise linear regression. Journal of Classification 17(2):273–296

Jedidi K, Krider RE, Weinberg CB (1998) Clustering at the movies. Marketing Letters 9(4):393–405

Krider RE, Li T, Liu Y, Weinberg CB (2005) The lead-lag puzzle of demand and distribution: A graphical method applied to movies. Marketing Science 24(4):635–645

Leisch F (2004a) Exploring the structure of mixture model components. In: Antoch J (ed) Compstat 2004 — Proceedings in Computational Statistics, Physica Verlag, Heidelberg, pp 1405–1412

Leisch F (2004b) FlexMix: A general framework for finite mixture models and latent class regression in R. Journal of Statistical Software 11(8), URL `http://www.jstatsoft.org/v11/i08/`

Lindsay BG (1989) Moment matrices: Applications in mixtures. The Annals of Statistics 17(2):722–740

McCullagh P, Nelder JA (1989) Generalized Linear Models, 2nd edn. Chapman and Hall

McLachlan GJ, Krishnan T (1997) The EM Algorithm and Extensions, 1st edn. John Wiley and Sons Inc.

Naik PA, Shi P, Tsai CL (2007) Extending the Akaike information criterion to mixture regression models. Journal of the American Statistical Association 102(477):244–254

Newcomb S (1886) A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics 8:343–366

Pearson K (1894) Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society A 185:71–110

R Development Core Team (2007) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`

Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26(2):195–239

Titterington DM, Smith AFM, Makov UE (1985) Statistical Analysis of Finite Mixture Distributions. Wiley

Wang P, Puterman ML (1998) Mixed logistic regression models. Journal of Agricultural, Biological, and Environmental Statistics 3(2):175–200

Wang P, Puterman ML, Cockburn IM, Le ND (1996) Mixed Poisson regression models with covariate dependent rates. Biometrics 52:381–400

Wedel M, Kamakura WA (2001) Market Segmentation — Conceptual and Methodological Foundations, 2nd edn. Kluwer Academic Publishers