LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Bettina Grün & Friedrich Leisch

Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects

Technical Report Number 024, 2008
Department of Statistics
University of Munich

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

Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects

Bettina Grün
Wirtschaftsuniversität Wien, Austria

Friedrich Leisch
Ludwig-Maximilians-Universität München, Germany

This is a pre-print of an article which has been accepted for publication in the *Journal of Classification*. The original publication will be available at `http://www.springerlink.com`.

Abstract

Unique parametrizations of models are very important for parameter interpretation and consistency of estimators. In this paper we analyze the identifiability of a general class of finite mixtures of multinomial logits with varying and fixed effects, which includes the popular multinomial logit and conditional logit models. The application of the general identifiability conditions is demonstrated on several important special cases and relations to previously established results are discussed. The main results are illustrated with a simulation study using artificial data and a marketing dataset of brand choices.

Keywords: conditional logit, finite mixture, identifiability, multinomial logit, unobserved heterogeneity

1 Introduction

Finite mixtures of multinomial and conditional logit models are part of the GLIMMIX framework (Wedel and DeSarbo, 1995) which covers finite mixtures of generalized linear models. Finite mixtures of multinomial logit models are applied in many different areas including for example medicine (Aitkin, 1999) or economics. Finite mixtures of conditional logit models are popular in marketing to investigate the brand choice of customers in dependency on marketing mix variables (Wedel and Kamakura, 2001) or in transportation research (Greene and Hensher, 2003). If individual level estimates are of interest, e.g. for market segmentation, improved estimates can be obtained using an empirical Bayes procedure instead of the point estimates (Kamakura and Wedel, 2004).

The importance of the model class is also indicated by the wide variety of software packages which can be used for estimation: GLLAMM (Rabe-Hesketh et al., 2004), Latent GOLD Choice (Vermunt and Magidson, 2003) and LIMDEP (Greene, 2002) with the add-on package NLOGIT. More limited variants are implemented in GLIMMIX (Wedel, 2002) and Mplus (Muthén and Muthén, 1998–2006).

Despite the popularity of the model class in applications a comprehensible analysis of the parameter identifiability is still missing to the authors' knowledge as only special cases have been considered. The identifiability of finite mixtures of binomial regression models has been discussed in Follmann and Lambert (1991). Their model only allows the intercept to follow a finite mixture distribution while for the other covariates fixed effects are assumed. They give sufficient identifiability conditions for mixing at the binary and the binomial level. This paper extends their results on identifiability to a more general model class of finite mixtures of multinomial logit models.

The model allows for varying and fixed effects for the coefficients of the covariates: Coefficients of *varying effects* follow a finite mixture, i.e., several groups are present in the population which have different parameters. Coefficients of *fixed effects* are constant over the whole population. This is similar to models with random effects which is also referred to as the mixed multinomial logit model (Revelt and Train, 1998). Continuous random effects are in general rarely used with a single replication per case as often the primary focus is to capture the interdependence between observations from the same individual. By contrast discrete varying effects are used to model heterogeneity in the data due to unobserved groups or clusters and therefore these models may also be applied to data without replications per case which can seriously impact on the identifiability of the models.

This paper is organized as follows: Section 2 introduces the model. Section 3 analyzes the identifiability of the model and applies the main theorem to important special cases as, e.g., finite mixtures of multinomial distributions and choice models. Section 4 illustrates the main results on artificial data with different numbers of covariate values and repetitions per individual such that the corresponding model is either identifiable or not identifiable. In Section 5 the application to marketing data where mixtures of conditional logit models are fitted is demonstrated. All computations are made in R (R Development Core Team, 2007) using package `flexmix` (Leisch 2004; Grün and Leisch 2007).

2 Model specification

Assume we have a categorical dependent variable $Y \in \{1, \ldots, K\}$, and let $\mathbb{P}(Y = k|\boldsymbol{z})$ be the probability that the dependent variable Y equals k given the covariates $\boldsymbol{z}$. Two popular regression models for these probabilities are the *multinomial logit* and *conditional logit* model, see e.g., Soofi (1992). The multinomial logit model uses a common set of predictors $\boldsymbol{z}$ for all levels of Y and choice-specific parameter vectors. The conditional logit model on the other hand allows for alternative-specific predictors $\boldsymbol{z}_k$ but uses the same parameter vector for all of them.

The combined multinomial and conditional logit model is given by

$$\mathbb{P}(Y = k|\boldsymbol{z}) \quad = \quad \frac{e^{\boldsymbol{z}'_{1,k}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_2\boldsymbol{\gamma}_{2,k}}}{\sum_{u=1}^{K} e^{\boldsymbol{z}'_{1,u}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_2\boldsymbol{\gamma}_{2,u}}}, \quad k = 1, \ldots, K$$

such that $\text{logit}[\mathbb{P}(Y = k|\boldsymbol{z})] = \boldsymbol{z}'_{1,k}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_2\boldsymbol{\gamma}_{2,k}$. $\boldsymbol{z}_{1,k}$ are the covariates for the conditional logit part and $\boldsymbol{z}_2$ the covariates for the multinomial logit part. For identifiability different contrasts can be imposed, as e.g., by defining category K as baseline and constraining $\boldsymbol{\gamma}_{2,K} = \mathbf{0}$ and $\boldsymbol{z}_{1,K} = \mathbf{0}$.

This basic model is extended to account for unobserved heterogeneity in the population by introducing varying effects which follow a finite mixture distribution. Assume there are S latent segments with prior probabilities $\pi_s > 0$, $\sum_{s=1}^{S} \pi_s = 1$. If the observation Y belongs to segment s the logit model with varying and fixed effects is given by

$$\text{logit}[\mathbb{P}(Y = k|\boldsymbol{x}, \boldsymbol{z}, s)] \quad = \quad \boldsymbol{x}'_{1,k}\boldsymbol{\beta}_1^s + \boldsymbol{x}'_2\boldsymbol{\beta}_{2,k}^s + \boldsymbol{z}'_{1,k}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_2\boldsymbol{\gamma}_{2,k}, \quad k = 1, \ldots, K.$$

$\boldsymbol{x}$ are the covariates and $\boldsymbol{\beta}$ the coefficients of the varying effects, and $\boldsymbol{z}$ the covariates and $\boldsymbol{\gamma}$ the coefficients of the fixed effects.

As will be shown below repeated measurements for some individuals are valuable information in order to ensure the identifiability of the model. Therefore, the following notation is convenient, where T denotes the set of all individuals in the population. All observations for a single individual t with equal covariate values $\boldsymbol{x}_i$ and $\boldsymbol{z}_j$ are combined. The observations are sorted such that the covariates of the varying effects for individual $t \in T$ are grouped together in the index set I_t. In addition we have a set J_i with the indices of all covariates for the fixed effects where the varying effects are equal to $\boldsymbol{x}_i$. For each individual t the unique covariate vectors are given by $(\boldsymbol{x}'_i, \boldsymbol{z}'_j)$ with $i \in I_t$ and $j \in J_i$. N_{ij} is the number of repeated observations available for this covariate vector and individual. The dependent variable $\boldsymbol{y}_{ij} \in \mathbb{N}^K$ for these unique covariate points is the vector of counts for each category, i.e. $\mathbf{1}'_K\boldsymbol{y}_{ij} = N_{ij}$, where $\mathbf{1}_K$ is the vector of K ones.

Table 1: Illustration of notation for a given individual t.

I_t	J_i	$\boldsymbol{x}_i$	z_{ij}	N_{ij}	$\boldsymbol{y}_{ij}$
1	1	1.1 0	4	7	3 2 2
1	2	1.1 0	0	3	1 2 0
1	3	1.1 0	1	1	0 0 1
1	4	1.1 0	2	5	2 1 2
2	1	2.7 1	4	1	0 0 1
2	2	2.7 1	0	1	0 1 0
2	3	2.7 1	2	6	4 0 2

For matrix notation all unique covariate points and the dependent variables are row-wise combined. Let $\boldsymbol{X}_{1,k} := (\boldsymbol{x}'_{1,k,i} : j \in J_i, i \in I_t, t \in T)$, $\boldsymbol{X}_2 := (\boldsymbol{x}'_{2,i} : j \in J_i, i \in I_t, t \in T)$, $\boldsymbol{X}_k := (\boldsymbol{X}_{1,k}, \boldsymbol{X}_2)$ and $\boldsymbol{X} := (\boldsymbol{X}_k : k = 1, \ldots, K)$ and let $\boldsymbol{Z}_{1,k}$, $\boldsymbol{Z}_2$, $\boldsymbol{Z}_k$, $\boldsymbol{Z}$ and $\boldsymbol{Y}$ be analogously defined. The mixture distribution can now be written as

$$H(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}) \quad = \quad \prod_{t \in T} \Big[\sum_{s=1}^{S} \pi_s \prod_{i \in I_t} \prod_{j \in J_i} F(\boldsymbol{y}_{ij}; N_{ij}, \boldsymbol{\theta}^s_{ij}) \Big], \tag{1}$$

where $F(\cdot; N, \boldsymbol{\theta})$ is the multinomial distribution with repetition parameter N and probability parameter vector $\boldsymbol{\theta} \in (0,1)^K$. Please note that the product over the observations for each individual t is within the sum over the finite mixture components. This signifies that the component membership is fixed for each individual.

For the probability parameter vectors it holds that

$$\text{logit}[\theta^s_{k,ij}] \quad = \quad \boldsymbol{x}'_{1,k,i}\boldsymbol{\beta}^s_1 + \boldsymbol{x}'_{2,i}\boldsymbol{\beta}^s_{2,k} + \boldsymbol{z}'_{1,k,j}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_{2,j}\boldsymbol{\gamma}_{2,k}.$$

The total parameter vector $\boldsymbol{\Theta}$ is equal to $((\pi_s, \boldsymbol{\beta}_s)_{s=1,\ldots,S}, \boldsymbol{\gamma})$ where $\boldsymbol{\beta}_s = (\boldsymbol{\beta}^s_1, (\boldsymbol{\beta}^s_{2,k})_{k=1,\ldots,K})$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, (\boldsymbol{\gamma}_{2,k})_{k=1,\ldots,K})$. It is assumed that $\Theta \in \Omega$ where Ω defines the parameter space given by $\Sigma_S \times \mathbb{R}^{\sum_{s=1}^{S} |\boldsymbol{\beta}_s| + |\boldsymbol{\gamma}|}$ where Σ_S is the unit simplex of dimension S and $|\cdot|$ gives the length of a vector.

The notation is illustrated by the following example where the data matrix for a given individual t is presented.

Example 1. *Let the dependent categorical variable have three different categories. The covariates $\boldsymbol{x}$ of the random effects consist of a numeric variable and a binary variable, whereas the covariate $\boldsymbol{z}$ of the fixed effects is a categorical variable with 4 categories. For simplicity of presentation these variables are all for a multinomial logit model. Assume that for individual t 24 trinomial outcomes are observed at 7 different covariate values. For example, when $\boldsymbol{x} = (1.1, 0)$ and $z = 0$ 3 trinomial outcomes are observed, a "1" and two "2"s. The varying covariate $\boldsymbol{x}$ assumes two values $(1.1, 0)$ and $(2.7, 1)$. These have, respectively, 16, and 8 replicates where we allow different z values. The corresponding data matrix is given in Table 1.*

3 Identifiability

Several kinds of identifiability problems arise for finite mixture models, including so-called *trivial* problems and *generic* problems (Frühwirth-Schnatter, 2006). Trivial identifiability problems refer to problems which occur due to empty components, due to components with the same parameters and due to the invariability of the likelihood to permutations of the segments. These problems can be avoided by restricting the feasible parameter space Ω to $\tilde{\Omega} \subset \Omega$ where for all $\Theta \in \tilde{\Omega}$:

- $\pi_s > 0$ for all $s = 1, \ldots, S$,
- $\boldsymbol{\beta}_s \neq \boldsymbol{\beta}_t$ for all $s \neq t; s, t \in \{1, \ldots, S\}$, and
- imposing a suitable ordering constraint, e.g., $\pi_s < \pi_t$, for all $1 \leq s < t \leq S$ if all segments have pairwise different sizes.

With respect to generic identifiability it has been shown that mixtures of binomial distributions (Teicher, 1963; Blischke, 1964; Titterington et al., 1985) are identifiable if the condition $N \geq 2S-1$ is fulfilled where N denotes the number of repetitions for a given individual. This constraint is necessary and sufficient for the model class of all mixtures with a maximum of S segments. Lindsay (1995, p. 48, Prop. 6) obtains the same result for the more general class of mixtures of discrete exponential family densities with $N + 1$ points of support and the same condition applies for mixtures of multinomial distributions (Grün 2002; Elmore and Wang 2003).

The identifiability of mixtures of Gaussian regression models is analyzed in Hennig (2000). The results indicate that requiring a covariate matrix of full rank is not sufficient. Contrarily, it is necessary to check a coverage condition in order to ensure identifiability. We extend Hennig's work to mixtures of multinomial and conditional logit models. We also allow varying and fixed effects for the coefficients and repeated observations where the segment membership is fixed. Supplementary results for finite mixtures of Gaussian regression models with two components where only local identifiability is considered are derived in Meijer and Ypma (2008).

Please note that the sufficient identifiability conditions imply that any mixture distribution function from the specified model class can be uniquely parameterized, i.e. the parameters can be uniquely determined given infinitely many observations. By contrast, if a mixture distribution is not identifiable the parameters can still not be uniquely determined even if an infinite amount of data is available.

3.1 Conditional logit

We first present sufficient conditions for identifiability of the conditional logit model with varying and fixed effects, results for the combined model are then derived in a second theorem.

Theorem 1. *The model defined by* (1) *where*

$$\ln\left[\frac{\theta^s_{k,ij}}{\theta^s_{K,ij}}\right] = \boldsymbol{x}'_{k,i}\boldsymbol{\beta}^s + \boldsymbol{z}'_{k,j}\boldsymbol{\gamma}$$

is identifiable if the following conditions are fulfilled:

1. (a) *for all* $k \in \{1, \ldots, K-1\}$ *there exists an* $\tilde{I}_k$ *which is not empty and a subset of* $\bigcup_{t \in T} I_t$ *and for which it holds that*

$$\sum_{i \in E_{k,i^*}} \sum_{j \in J_i} N_{ij} \geq 2S - 1 \qquad \text{for all } i^* \in \tilde{I}_k.$$

E_{k,i^*} is given by $\{i \in I_{t(i^*)} : \boldsymbol{x}_{k,i} = \boldsymbol{x}_{k,i^*}\}$. $I_{t(i^*)}$ is defined as the index set of all observations for the individual t with covariate vector $\boldsymbol{x}_{k,i^*}$.

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

$$q^* := \min\Bigg\{q : \text{ for all } i^* \in \bigcup_{k=1}^{K-1} \tilde{I}_k : \text{ there exists an } H_j \in \{H_1, \ldots, H_q\} \text{ with } \{\boldsymbol{x}_{k,i} : i \in I_{t(i^*)} \cap \tilde{I}_k, k = 1, \ldots, K-1\} \subseteq H_j \wedge H_j \in \mathcal{H}_U\Bigg\},$$

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

2. *$rk(\boldsymbol{X}, \boldsymbol{Z}) = U + V$ where $rk(\cdot)$ determines the rank of a matrix.*

3. *$\boldsymbol{x}_{K,i} = \boldsymbol{0}$ and $\boldsymbol{z}_{K,j} = \boldsymbol{0}$ for all $j \in J_i, i \in I_t$ and $t \in T$.*

The proof is given in the Appendix. Condition (1) guarantees that no intra-component label switching is possible. Intra-component label switching refers to the identifiability problem where even if the labels are fixed in one covariate point according to some ordering constraint, the labels may switch in other covariate points for the different possible parameterizations of the model.

As the segment membership is fixed for each individual only those hyperplanes are feasible where the covariate points from the same individual lie on the same hyperplane. Condition (1a) implies that there exists a $t \in T$ with at least $2S - 1$ observations. For these observations the covariates for the varying effects have to be constant, but they can vary for the fixed effects. The inclusion of the set E_{k,i^*} is possible, because the covariates are allowed to change in the other categories of the multinomial distribution. Condition (1b) corresponds to the coverage condition in Hennig (2000) for mixtures of Gaussian regressions which ensures that no intra-component label switching is possible. While in Hennig (2000) only the case of one repetition per individual is considered we generalize the condition for the case where repeated observations per individual are available. The coverage condition implies that the maximum number of segments for the mixture has to be smaller than the minimum number of feasible hyperplanes which are necessary to cover the covariate points for all $k = 1, \ldots, K$ where enough repetitions are available to guarantee marginal identifiability. Hyperplanes are feasible if (1) they go through the origin and (2) they cover all observations from the same individual which are marginally identifiable. This is a stronger condition than to have full rank of the corresponding covariate matrix.

Condition (2) and (3) correspond to conditions which are necessary for a model without varying effects in order to uniquely determine the coefficients. Condition (2) also ensures that the partition between fixed and varying effects is unique.

3.2 Multinomial and conditional logit

The following theorem gives sufficient identifiability conditions for the combined model presented in Section 2. The proof is straight-forward given Theorem 1 and using that the multinomial logit part can be transformed to a conditional logit model (Agresti, 1990, pp. 316–317).

Theorem 2. *The model defined by* (1) *where*

$$\ln\left[\frac{\theta^s_{k,ij}}{\theta^s_{K,ij}}\right] = \boldsymbol{x}'_{1,k,i}\boldsymbol{\beta}^s_1 + \boldsymbol{x}'_{2,i}\boldsymbol{\beta}^s_{2,k} + \boldsymbol{z}'_{1,k,j}\boldsymbol{\gamma}_1 + \boldsymbol{z}'_{2,j}\boldsymbol{\gamma}_{2,k}$$

is identifiable if the following conditions are fulfilled:

1. *(a) for all $k \in \{1, \ldots, K-1\}$ there exists an $\tilde{I}_k$ which is not empty and a subset of $\bigcup_{t\in T} I_t$ and for which it holds that*

$$\sum_{i \in E_{k,i^*}} \sum_{j \in J_i} N_{ij} \geq 2S - 1 \qquad \text{for all } i^* \in \tilde{I}_k.$$

E_{k,i^*} is given by $\{i \in I_{t(i^*)} : \boldsymbol{x}_{1,k,i} = \boldsymbol{x}_{1,k,i^*} \wedge \boldsymbol{x}_{2,i} = \boldsymbol{x}_{2,i^*}\}$.

(b) $q^* > S$ with

$$q^* := \min\Bigg\{q : \text{ for all } i^* \in \bigcup_{k=1}^{K-1} \tilde{I}_k : \text{ there exists an } H_j \in \{H_1, \ldots, H_q\} \text{ with } \{(\boldsymbol{x}'_{1,k,i}, \boldsymbol{x}'_{2,i}) : i \in I_{t(i^*)} \cap \tilde{I}_k, k = 1, \ldots, K-1\} \subseteq H_j \wedge H_j \in \mathcal{H}_U\Bigg\},$$

where $\mathcal{H}_U$ is the set of $H(\alpha) := \{\boldsymbol{x} \in \mathbb{R}^{U_1+U_2} : \boldsymbol{\alpha}'\boldsymbol{x} = \boldsymbol{0}\}$ where $\boldsymbol{\alpha} \neq \boldsymbol{0}$.

2. *$rk(\boldsymbol{X}, \boldsymbol{Z}) = U + V$*

3. *$\boldsymbol{x}_{1,K,i} = \boldsymbol{0}$ and $\boldsymbol{z}_{1,K,j} = \boldsymbol{0}$ for all $j \in J_i, i \in I_t$ and $t \in T$, and $\boldsymbol{\beta}_{2,K} = \boldsymbol{\gamma}_{2,K} = \boldsymbol{0}$.*

3.3 Special cases

In the following we illustrate which sufficient identifiability constraints can be derived from Theorem 2 for important special cases.

Mixtures of multinomial distributions: The simplest case are mixtures of multinomial distributions without a regression part where only a segment specific intercept needs to be estimated such that $x_{2,i} \equiv 1$ for all $i \in I$ and $U_1 = V = 0$:

$$\ln\left[\frac{\mathbb{P}(Y_r = k)}{\mathbb{P}(Y_r = K)}\right] \quad = \quad \boldsymbol{\beta}_{2,k}^s \qquad \text{for all } k = 1, \ldots, K.$$

Condition (1a) ensures that the number of observations N is for at least one individual larger or equal to $2S-1$. Hence we have the same results as in Grün (2002) and Elmore and Wang (2003): The class of mixtures of multinomial distributions with a maximum of S segments is identifiable if $N \geq 2S-1$ where again N denotes the number of repetitions for a given subject.

Model in Follmann and Lambert (1991): Theorem 2 generalizes the first set of sufficient conditions in Follmann and Lambert (1991). They considered mixtures of binomial logit distributions where only the intercept followed a finite mixture distribution and all other coefficients were constant. Hence for our model this signifies $K = 2$, $x_{1,1,i} \equiv 1$ for all $i \in I$, $U_2 = 0$ and V arbitrary.

As the multinomial and conditional logit model are equivalent in the binomial case the model is given by

$$\ln\left[\frac{\mathbb{P}(Y_r = 1|\boldsymbol{z}_r)}{\mathbb{P}(Y_r = 0|\boldsymbol{z}_r)}\right] \quad = \quad \boldsymbol{\beta}_1^s + \boldsymbol{z}_r'\boldsymbol{\gamma}.$$

The conditions in Follmann and Lambert (1991) are:

- There exists an $i \in I$ with a $j \in J_i$ such that $N_{ij} \geq 2S-1$, and
- $\mathrm{rk}(\mathbf{1}, \boldsymbol{Z}) = 1 + V$.

For condition (1a) to be fulfilled the number of repetitions N for a given individual has to be at least $2S-1$. In contrast to Follmann and Lambert (1991) the covariates for the fixed effects are allowed to vary, i.e. we require only that there exists an $i \in I$ such that $\sum_{j \in J_i} N_{ij} \geq 2S-1$. The other condition which has to be checked is condition (2) which corresponds to the rank condition of Follmann and Lambert (1991). Our conditions are less restricting as we take repeated observations for individuals with different covariate points into account.

Choice models: Conditional logit models are often applied as choice models in marketing research based on random utility theory (McFadden, 1974). Kamakura and Russell (1989) estimated a finite mixture of conditional logit models assuming that the price elasticity of the consumers varies over the consumer population but is fixed for each consumer over the different brands. This model can be specified within our framework by setting $U_2 = 0$ and $V = 0$. The conditional logit model with only varying effects is given by

$$\ln\left[\frac{\mathbb{P}(Y_r = k|\boldsymbol{x}_{1,k,i})}{\mathbb{P}(Y_r = K|\boldsymbol{x}_{1,k,i})}\right] \quad = \quad \boldsymbol{x}_{1,k,i}'\boldsymbol{\beta}_1^s \qquad \text{for all } k = 1, \ldots, K.$$

An application of the sufficient identifiability conditions to a real dataset is demonstrated in Section 5.

4 Illustration

In the following the identifiability of mixture distributions which are induced by the same parameterization, but differ with respect to the covariate matrix and the number of repetitions is analyzed. Each mixture distribution has a binomial dependent variable with categories 0 and 1 and two regressors consisting of the intercept and a univariate variable x. The probability of observing a 1 is used as the binomial parameter and in the following referred to as choice probability. The mixture distributions have two segments which are of equal size and the regression coefficients of the two segments are given by $\beta_1 = (-2, 4)'$ and $\beta_2 = (2, -4)'$, i.e., varying effects are assumed for the intercept and the covariate x.

Each mixture distribution is either defined on 2 or 5 different covariate points x which are equidistantly spread across the interval $[0, 1]$ (i.e. $\#x \in \{2, 5\}$). The mixture distributions where only 2 covariate points are available are not identifiable. In this case intra-component label switching is possible and the regression coefficients of the second solution are given by $\beta_1^{(2)} = (-2, 0)'$ and $\beta_2^{(2)} = (2, 0)'$. The number of repetitions N is fixed over all individuals and repetitions are only available for the same covariate point. The parameter N takes the values 1 or 10 for the different mixture distributions. The condition $N \geq 2S - 1$ implies that the mixture is not identifiable for $N = 1$.

In Figure 1 the observed relative frequencies of choices of 1 are given for random samples with 100 observations from each of the mixture distributions. A balanced sampling design with an equal number of observations in each covariate point is used. The symmetry of the specified model is not entirely reflected in the observed values as the sample sizes are rather small. The solid curves are the choice probabilities for each segment of the true underlying model. For $N = 1$ the mixture is observational equivalent to a degenerate mixture with only one segment. The probabilities of the degenerate model are given by the dashed line. Following the principle of parsimonity in the model fitting process, the degenerate mixture would be selected as solution. In addition to the degenerate mixture all mixtures with two segments are observationally equivalent where (1) the aggregated marginal choice probabilities are equal to those of the true model for each covariate point and (2) the relationship between the logit of the choice probabilities and x is linear. For $N = 10$ it can be seen that intra-component label switching is possible if only two different covariate points are available, whereas the mixture is identifiable for five different covariate points. The choice probabilities of each segment of the observational equivalent mixture for two covariate points are given by the dotted lines.

100 samples with 100 observations each are drawn from each mixture distribution given by all possible combinations of N and $\#x$. To each sample a mixture with two segments is fitted using the EM algorithm. The stopping criterion is the difference in log-likelihood. Even in the degenerate case where a manifold of solutions exist and the parameter estimates do not converge the observed log-likelihood values converge as the log-likelihood has the same value for parameter estimates implying observational equivalent mixture distributions (Wu, 1983).

In order to avoid local maxima the best solution of 5 runs of the EM algorithm with different random initializations is reported. The choice of 5 repetitions seems to be reasonable as only for 3% of the fitted models the best model is not already detected in runs 1–4. In most of the cases (86%) the best model is already found in the first and second repetition, in 81% of the cases the detected global maximum value of the log-likelihood is achieved at least twice.

In Figure 2 parallel coordinate plots (Wegman, 1990) are used to investigate the estimated parameters for all combinations of N and $\#x$. For an identifiable mixture with two segments the parallel coordinate plot should contain two "bundles" corresponding to the parameters of each of the segments. The parallel coordinate plot is robustified by rescaling the parameter estimates for each dimension so that 95% of the observations are shown and that the 2.5% of observations with lowest and highest values for each dimentsion are outside the plotting region.

For $N = 1$ only one large bundle can be discerned but with a wide variability of the parameter estimates for $\#x = 5$. For $N = 10$ it depends on $\#x$ if two or four bundles can be distinguished. Intra-component label switching occurs for $\#x = 2$. For the intercept the parameter estimates

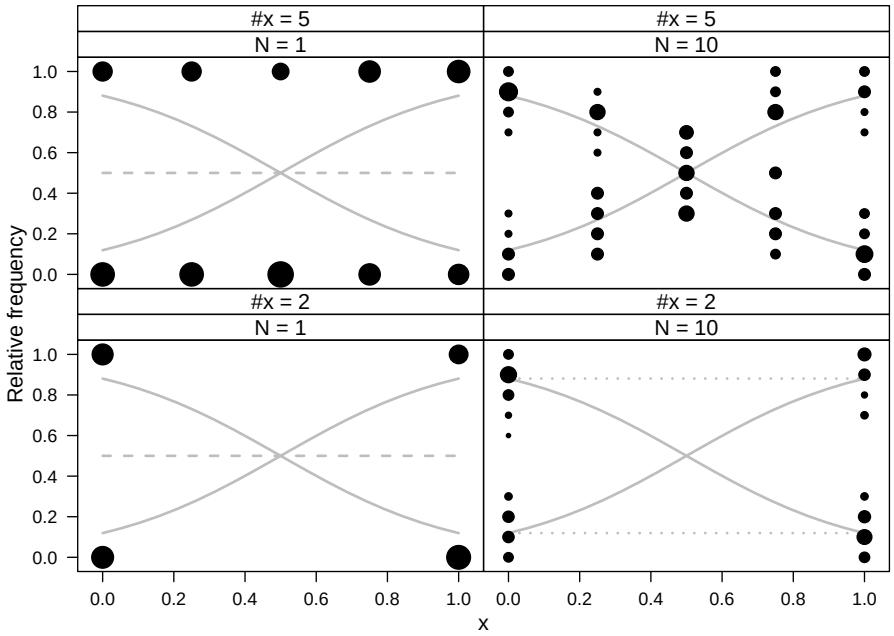

Figure 1: Observed values for the artificial example with $N = \{1, 10\}$ and where the number of different x values is 2 or 5.

cluster around two different values and the coefficients of x cluster around three different values. This induces four different bundles. For $\#x = 5$ two bundles can be distinguished which correspond to the parameter vectors for each segment of the true model. An extended version of this illustrative example is described in Grün (2006).

5 Application

A dataset on brand choice is analyzed in order to illustrate how the sufficient identifiability constraints can be used to gain insights into potential identifiability problems before and after the data analysis. These findings are compared to those derived using the parametric bootstrap to analyze the fitted finite mixture (Grün and Leisch 2004). The optical-scanner panel dataset is from Jain et al. (1994) where finite mixtures of conditional logit models are fitted. 2798 purchases of a single product category made by 300 households over the data-collection period of about two years are available. The purchases of 4 catsup brands are included: Hunt's 32, Heinz 40, Heinz 32 and Heinz 28. The available marketing mix informations are price, feature and display. The model which is preferred as the best in Jain et al. (1994) is a 4-component mixture where all coefficients are allowed to vary between the components.

Two factors are crucial in order to guarantee identifiability: (1) the number of repetitions in each covariate point for each individual and (2) the number of different covariate points for each individual and over all individuals. Identifiability problems would occur if the purchases were observed without the household information. In this case only binary observations would be available and it would therefore be possible to separate the four brands with a mixture with four components so that in each component only the purchases of one single brand are contained.

The specified model class is also not generically identifiable if no observations with different values for the marketing mix variables feature or display were available for the households. While this is in practice an unlikely problem, possibly other binary variables could be included which depend on the specific shop or region and where it is less likely that purchases with different

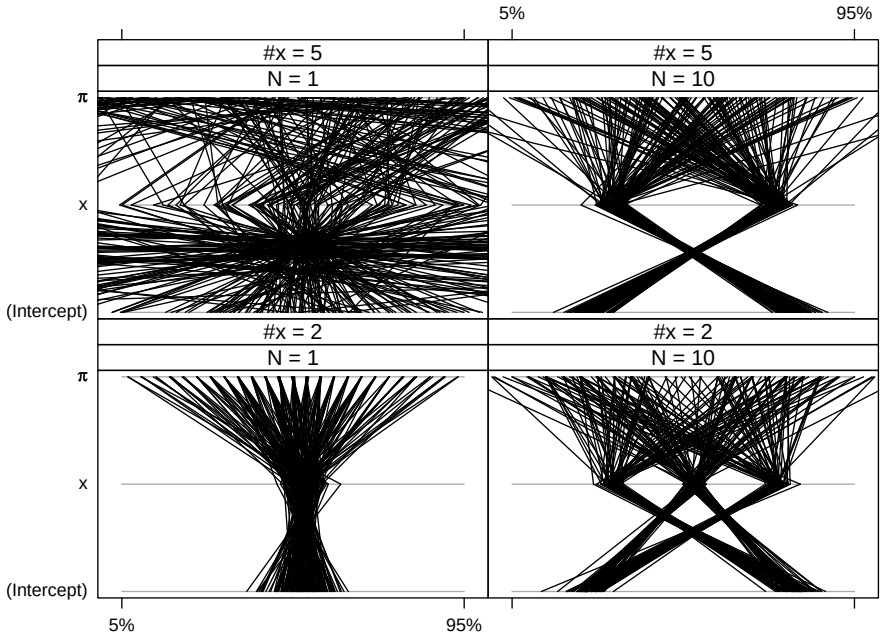

Figure 2: Parallel coordinate plots of the estimated parameters for 100 samples from the artificial example with different repetition parameters N and number of covariate points x.

values for these variables are observed for the same household. In this case identifiability problems occur for certain parameterizations from this model class. This signifies that it might again not be a problem for a specific fitted mixture. An unidentified model from this model class would for example be given by a mixture where two equally sized components are homogeneous with respect to all covariates except for this binary variable. Then the components can be arbitrarily combined between the two values of the binary variable.

An overview on the number of observations per household and covariate point for the dataset is given in Figure 3. As observations are collected over a rather long period multiple observations are available for each household with a minimum of 5, a maximum of 44 and a mean of 9.33 observations. In order to analyze the number of different covariate points with the corresponding repetitions per individual a modified model matrix is used: The continuous price variable is split into an ordered categorical variable consisting of three equally sized classes labelled "cheaper", "about the same" and "more expensive" with Hunt's 32 as baseline. For this modified model matrix observations from the same individual with similar price values are combined in order to increase the number of repetitions in each covariate point. As a monotone relationship between the associated utility and the price is assumed, the underlying mixture model generating this data is similar to the original model. Even after categorizing the price variable only one repetition per individual is available for most of the covariate points (1422 out of 3010; 47%). For 263 (9%) of the covariate points there are at least 7 repetitions per individual available which means that they are identifiable for a mixture with at most 4 components.

The availability of different covariate points for the same individual reduces the feasible hyperplanes. For all respondents there are observations for at least 5 different covariate points available with a mean of 10.03 covariate points. With respect to the number of different covariate points which are also identifiable most households (179 out of 300; 60%) have again no covariate points where this condition is fulfilled. However, there are 44 households (15%) with two covariate points, 29 households (10%) with three covariate points and 48 households (16%) with four and more covariate points.

This investigation indicates that the available data allows to uniquely identify a mixture with

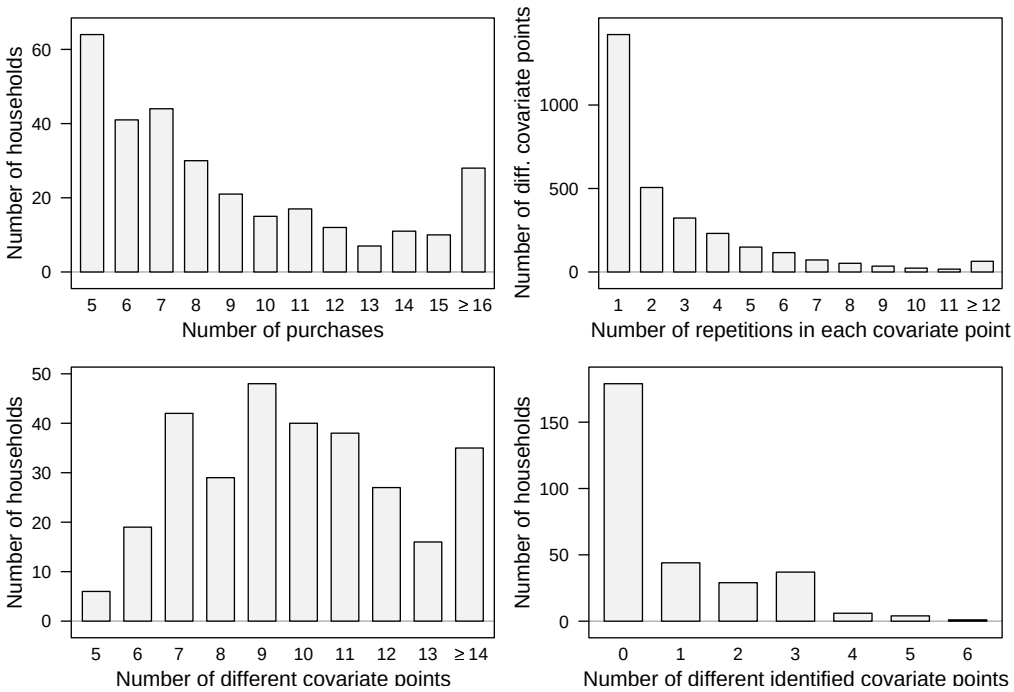

Figure 3: Information to assess the risk of identifiability problems to occur.

four components as there are several covariate points where enough repetitions are available to ensure identifiability in these points and in addition there are several households where this holds for two or more covariate points which ensures that the coverage condition is fulfilled.

This conclusion is verified using the parametric bootstrap. In addition the influence of the number of repeated observations for each household on identifiability and the occurence of local maxima is analyzed by modifying the household information of the original dataset. The observations of each real household are as evenly as possible assigned to 1, 2, 4 and 8 different "artificial" households. This simulates the situation that a dataset of the same overall size and structure is collected over shorter periods of time, such that each individual household shops less often. As we are not interested in the effects of sample size on parameter estimation here, we keep the overall sample size constant. 100 parametric bootstrap samples are drawn from the originally fitted mixture with the modified model matrix, where the household information is changed. Mixture models with 4 components are fitted to each bootstrap sample with the EM algorithm using 5 different random initializations and the log-likelihood as convergence criterion.

Again, the choice of 5 EM-repetitions seems to be reasonable, as only for 2% of the bootstrap samples the best model is not already detected in runs 1–4, and in most of the cases (82%) the best model is already found in the first or second repetition. In addition the global maximum value of the log-likelihood is detected at least two times during the 5 runs and for 95% of the bootstrap samples the global maximum value is detected either 4 or 5 times. A unique labelling of the components for all fitted models is attempted by relabelling the components in order to minimize the Euclidean distance to the parameters of the originally fitted mixture. This approach is similar to the relabelling procedure proposed in Marin et al. (2005) for Bayesian modelling where the distance to the maximum a-posteriori estimate (MAP) is minimized.

In Figure 4 the increase in variability of the parameter estimates in dependence of the decrease in household information is illustated. The top row in Figure 4 indicates that no identifiability problems are present as the estimated values cluster around the parameter estimates of the originally fitted model which are indicated by the white lines. Table 2 illustrates the difference in

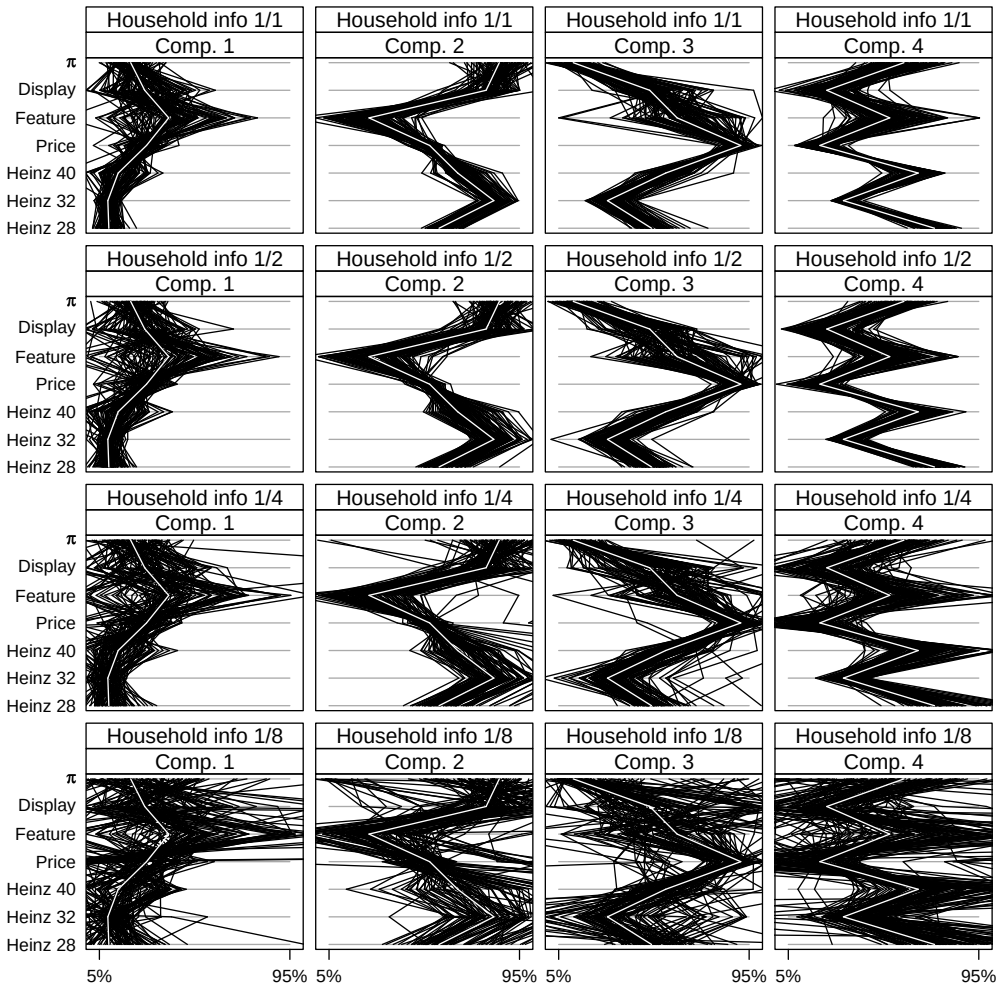

Figure 4: Parameter estimates for the 100 parametric bootstrap samples.

convergence behaviour and stability of induced clusterings. The columns indicate if the observations of each single household are assigned to 1, 2, 4, or 8 different households. The rows for "# global maximum" indicate how often the detected global maximum of the likelihood is reached during the 5 repetitions of the EM algorithm. While for 75% of the bootstrap samples where the household information is unchanged the same global maximum of the log-likelihood is detected in each run of the EM algorithm, this goes down to 3% of the bootstrap samples where the observations of one household are assigned to 8 different households. By contrast the global maximum of the log-likelihood is only detected once during the 5 runs for 67% of the samples. This suggests the presence of several local maxima and a flatter likelihood.

The Rand index corrected by chance (Hubert and Arabie, 1985) is used to analyze the stability of the clusterings induced by assigning each observation to the component with the maximum a-posteriori probability. The induced clustering of the original data using the estimated solution is compared to those predicted by the models fitted to the bootstrap samples. A decrease in stability can clearly be observed if the observations per household are reduced. This analysis indicates that with only half of the information per household the results are still comparable with respect to convergence behaviour and stability of classifications, but the performance deteriorates considerably if less information per household is available.

Table 2: Convergence behaviour and stability of classification in dependence of the number of repetitions per household.

		Household info			
		1/1	1/2	1/4	1/8
# global maximum	5	75	66	32	3
	4	20	10	10	6
	3	2	8	15	10
	2	3	9	17	14
	1	0	7	26	67
corr. Rand index	Mean	0.97	0.96	0.94	0.87
	Std.Dev.	0.01	0.02	0.03	0.06

6 Conclusions and future work

Generic identifiability is defined for a given model class and it only ensures that the model can be uniquely determined given an infinite number of observations. In the finite sample case empirical identifiability problems are also of concern, where in fact the question is if the given dataset is informative enough to distinguish between different mixture models.

The sufficient identifiability constraints guarantee the identifiability of a general model class such that no two parameterizations are included which induce the same mixture model. However, if a certain model class is not identifiable this does not signify that for each finite mixture included in the model class there exist two or more different parameterizations. Even if the model class suffers from identifiability problems, a given mixture from this class might be identifiable and hence have a unique parameterization.

In the future the identifiability conditions should be extended to include concomitant variables (Dayton and Macready, 1988). This is not straight-forward if the concomitant variables and the regressor variables are not independent. The identifiability of mixture regression models where the cluster membership is not independent of the covariate values is analyzed in Hennig (2000) and special identifiability conditions apply. Furthermore, it might be interesting to investigate if the second set of sufficient identifiability conditions given in Follmann and Lambert (1991) can be generalized to mixtures of multinomial logit models.

A Proof of Theorem 1

If the model is not identifiable, there exist two different parameterizations $\boldsymbol{\Theta}$ and $\check{\boldsymbol{\Theta}}$ with at most S segments such that $H(\cdot \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}) \equiv H(\cdot \mid \boldsymbol{X}, \boldsymbol{Z}, \check{\boldsymbol{\Theta}})$, where $\boldsymbol{\Theta} = (\pi_l, \boldsymbol{\beta}_l, \boldsymbol{\gamma})_{l=1,\ldots,s}$ and $\check{\boldsymbol{\Theta}} = (\check{\pi}_m, \check{\boldsymbol{\beta}}_m, \check{\boldsymbol{\gamma}})_{m=1,\ldots,\check{s}}$

With condition (1a) we show that the binomial distributions with alternatives $\{k, K\}$ are identifiable for all $i^* \in \tilde{I}_k$ for all k (Step (a) and (b)). The covariate points where binomial identifiability was shown can be used to prove that no intra-component label switching is possible (Step (c)). This gives that the coefficients of the fixed and varying effects are identical up to arbitrary constants. The rank condition is needed to prove that the constants are equal to zero (Step (d)).

(a) We show that for all $k = 1, \ldots, K$ and for all $i^* \in \tilde{I}_k$:

$$\boldsymbol{z}'_{k,j}(\boldsymbol{\gamma} - \check{\boldsymbol{\gamma}}) = c_{k,i^*} \quad \text{for all } j \in \bigcup_{i \in E_{k,i^*}} J_i. \tag{2}$$

(b) We show that given an arbitrary $k \in \{1, \ldots, K-1\}$ it holds for all $i^* \in \tilde{I}_k$ that $\check{s}(i^*) = s(i^*)$ and that there exists a suitable ordering of the segments such that for all $l = 1, \ldots, s(i^*)$:

$$\alpha^l_{k,i^*} = \check{\alpha}^l_{k,i^*} + c_{k,i^*} \tag{3}$$

with $\alpha^l_{k,i^*} \in \{\boldsymbol{x}'_{k,i^*}\boldsymbol{\beta}_u : u = 1, \ldots, s\}$ and $\check{\alpha}^l_{k,i^*}$ analogously defined.

(c) We show with condition (1b) analogously to Hennig (2000) that $\check{s} = s$ and that for a suitable ordering of the segments it holds for suitable $\boldsymbol{\delta} \in \mathbb{R}^U$ that for all $l = 1, \ldots, s$:

$$\check{\pi}_l = \pi_l \qquad \text{and} \qquad \check{\boldsymbol{\beta}}_l = \boldsymbol{\beta}_l + \boldsymbol{\delta}. \tag{4}$$

(d) We show $\boldsymbol{\delta} = \mathbf{0}$ and $\boldsymbol{\gamma} = \check{\boldsymbol{\gamma}}$.

ad (a): The equation trivially holds for $k = K$ and the following holds for all $k = 1, \ldots, K-1$. If the mixture distributions are equivalent, this equivalence must also hold for a subset of the covariate points. Hence, we have for all $i^* \in \tilde{I}_k$:

$$\sum_{l=1}^{s} \pi_l \prod_{i \in E_{k,i^*}} \prod_{j \in J_i} F(\boldsymbol{y}_{ij}; N_{ij}, \boldsymbol{\theta}^l_{ij}) \;=\; \sum_{m=1}^{\check{s}} \check{\pi}_m \prod_{i \in E_{k,i^*}} \prod_{j \in J_i} F(\boldsymbol{y}_{ij}; N_{ij}, \check{\boldsymbol{\theta}}^m_{ij}).$$

The following holds for all $u \in E_{k,i^*}$ and $v \in J_u$ where $(\boldsymbol{y}_{ij})_{j \in J_i, i \in E_{k,i^*}}$ is given by $y_{k,ij} = \delta_{iu,jv}$ and $y_{K,ij} = N_{ij} - y_{k,ij}$. $\delta_{iu,jv}$ is the Kronecker delta, i.e. it is one if $i = u$ and $j = v$ and zero otherwise. The multinomial coefficients on both sides are cancelled and the terms which do not depend on l or m are taken out of the sums and separated on one side of the equation:

$$e^{\boldsymbol{z}'_{k,v}(\boldsymbol{\gamma} - \check{\boldsymbol{\gamma}})} = \frac{\sum_{m=1}^{\check{s}} \check{\pi}_m \left[e^{\check{\alpha}^m_{k,i^*}} \prod_{i \in E_{k,i^*}} \prod_{j \in J_i} (\sum_{h=1}^{K} e^{\check{\alpha}^m_{h,i} + \boldsymbol{z}'_{h,j}\check{\boldsymbol{\gamma}}})^{-N_{ij}} \right]}{\sum_{l=1}^{s} \pi_l \left[e^{\alpha^l_{k,i^*}} \prod_{i \in E_{k,i^*}} \prod_{j \in J_i} (\sum_{h=1}^{K} e^{\alpha^l_{h,i} + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma}})^{-N_{ij}} \right]}.$$

As the right hand side does not depend on index v the left hand side is constant for all $v \in \bigcup_{u \in E_{k,i^*}} J_u$ given i^*. This constant can be given by $e^{c_{k,i^*}}$. Hence, we have shown equation (2).

ad (b): For a given k and i^* we define $y_{k,..} := \sum_{i \in E_{k,i^*}} \sum_{j \in J_i} y_{k,ij}$. In the following we insert the dependent variable where $y_{K,ij} = N_{ij} - y_{k,ij}$. Then it holds for all $i^* \in \tilde{I}_k$:

$$\sum_{l=1}^{s} \pi_l \frac{e^{y_{k,..}\alpha^l_{k,i^*} + \sum_{i \in E_{k,i^*}} \sum_{j \in J_i} y_{k,ij} \boldsymbol{z}'_{k,j}\boldsymbol{\gamma}}}{\prod_{i \in E_{k,i^*}} \prod_{j \in J_i} \left(\sum_{h=1}^{K} e^{\alpha^l_{h,i} + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma}} \right)^{N_{ij}}} =$$

$$\sum_{m=1}^{\check{s}} \check{\pi}_m \frac{e^{y_{k,..}\check{\alpha}^m_{k,i^*} + \sum_{i \in E_{k,i^*}} \sum_{j \in J_i} y_{k,ij} \boldsymbol{z}'_{k,j}\boldsymbol{\gamma} + y_{k,..}c_{k,i^*}}}{\prod_{i \in E_{k,i^*}} \prod_{j \in J_i} \left(\sum_{h=1}^{K} e^{\check{\alpha}^m_{h,i} + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma} + c_{h,i^*}} \right)^{N_{ij}}}. \tag{5}$$

As the denominator on the left hand side only depends on i^* and l and not on j and $y_{k,ij}$, we define:

$$\lambda^l_{k,i^*} := \frac{\pi_l}{\prod_{i \in E_{k,i^*}} \prod_{j \in J_i} \left(\sum_{h=1}^{K} e^{\alpha^l_{h,i} + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma}} \right)^{N_{ij}}}.$$

$\check{\lambda}^m_{k,i^*}$ is analogously defined.

Substituting λ_{k,i^*}^l and $\check{\lambda}_{k,i^*}^m$ into equation (5) and eliminating the equal terms on the left and right hand side gives for all $i^* \in \tilde{I}_k$:

$$\sum_{l=1}^{s} \lambda_{k,i^*}^l \left(e^{\alpha_{k,i^*}^l} \right)^{y_{k,..}} = \sum_{m=1}^{\check{s}} \check{\lambda}_{k,i^*}^m \left(e^{\check{\alpha}_{k,i^*}^m + c_{k,i^*}} \right)^{y_{k,..}} \tag{6}$$

with $y_{k,..} \in \{0, \ldots, \sum_{i \in E_{k,i^*}} \sum_{j \in J_i} N_{ij}\}$.

With condition (1a) it follows that the sum over the unique elements in equation (6) has only the trivial solution for all $i^* \in \tilde{I}_k$. This implies that equation (3) holds. This also means that the k^{th} marginal binomial distribution with alternatives $\{k, K\}$ is identifiable in point $\boldsymbol{x}_{k,i^*}$.

ad (c): We assume that there can be a $\tilde{\boldsymbol{\beta}}_l$ defined for all l such that $\tilde{\boldsymbol{X}}_{k,i}\boldsymbol{\beta}_l + \tilde{\boldsymbol{Z}}_{k,i}\boldsymbol{\gamma} = \tilde{\boldsymbol{X}}_{k,i}\tilde{\boldsymbol{\beta}}_l + \tilde{\boldsymbol{Z}}_{k,i}\check{\boldsymbol{\gamma}}$ holds for all $i \in \tilde{I}_k$ and $k \in \{1, \ldots, K-1\}$. $\tilde{\boldsymbol{X}}_{k,i} := (\boldsymbol{x}'_{k,i})_{j \in J_i}$ and $\tilde{\boldsymbol{Z}}_{k,i}$ are analogously defined. The existence of $\tilde{\boldsymbol{\beta}}_l$ is guaranteed because (1) the inverse logit function is a one-to-one mapping (due to condition (3)) and (2) all marginal binomial distribution with alternatives $\{k, K\}$ are identifiable for all $k = 1, \ldots, K-1$. These two conditions imply that for all $i \in \tilde{I}_k$:

$$\tilde{\boldsymbol{X}}_{\tilde{I}} \left(\sum_{l=1}^{s} \pi_l \boldsymbol{\beta}_l - \sum_{m=1}^{\check{s}} \check{\pi}_m \check{\boldsymbol{\beta}}_m \right) = \tilde{\boldsymbol{Z}}_{\tilde{I}} (\check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}),$$

where $\tilde{\boldsymbol{X}}_{\tilde{I}} := (\tilde{\boldsymbol{X}}_{k,i})_{i \in \tilde{I}_k, k=1,\ldots,K-1}$ and $\tilde{\boldsymbol{Z}}_{\tilde{I}}$ is analogously defined.

As because of condition (1b) $\tilde{\boldsymbol{X}}_{\tilde{I}}$ has full column rank we can define

$$\tilde{\boldsymbol{\beta}}_l := \boldsymbol{\beta}_l + \boldsymbol{\delta} \qquad \text{with} \qquad \boldsymbol{\delta} := \left(\tilde{\boldsymbol{X}}'_{\tilde{I}} \tilde{\boldsymbol{X}}_{\tilde{I}} \right)^{-1} \tilde{\boldsymbol{X}}'_{\tilde{I}} \tilde{\boldsymbol{Z}}_{\tilde{I}} (\boldsymbol{\gamma} - \check{\boldsymbol{\gamma}}).$$

We assume without loss of generality that

$$\pi_1 \neq \check{\pi}_1 \qquad \text{and} \qquad s \geq \check{s}, \tag{7}$$

where $\check{\pi}_1$ is the a-priori probability for $\tilde{\boldsymbol{\beta}}_1$ with $\check{\pi}_1 \geq 0$.

As the marginal binomial mixture distributions for $k = 1, \ldots, K-1$ with alternatives $\{k, K\}$ are identifiable for all $i \in \tilde{I}_k$, the following must hold for all $i \in \tilde{I}_k$ given an arbitrary $k \in \{1, \ldots, K-1\}$:

$$\sum_{\{l=1,\ldots,\check{s}:\, \boldsymbol{x}'_{k,i}\check{\boldsymbol{\beta}}_l = \boldsymbol{x}'_{k,i}\tilde{\boldsymbol{\beta}}_1\}} \check{\pi}_l = \sum_{\{h=1,\ldots,s:\, \boldsymbol{x}'_{k,i}\boldsymbol{\beta}_h = \boldsymbol{x}'_{k,i}\boldsymbol{\beta}_1\}} \pi_h. \tag{8}$$

The assumption $S < q^*$ is in contradiction to the existence of some $\tilde{\boldsymbol{\beta}} \in \{\tilde{\boldsymbol{\beta}}_u : u = 1, \ldots, s\}$ such that there exists an $l \in \{1, \ldots, s\}$ with

$$\tilde{\boldsymbol{\beta}} \neq \check{\boldsymbol{\beta}}_l \qquad \wedge \qquad \boldsymbol{x}'_{k,i}\tilde{\boldsymbol{\beta}} = \boldsymbol{x}'_{k,i}\check{\boldsymbol{\beta}}_l \quad \text{for all } k \in \{1, \ldots, K-1\} \wedge i \in I_{t(i^*)} \cap \tilde{I}_k,$$

because then $q^* \leq s \leq S$ would hold.

Thus it holds for all $\tilde{\boldsymbol{\beta}}_l$ $l = 1, \ldots, s$— and in particular for $\tilde{\boldsymbol{\beta}}_1$—that there exists a $k^* = k(\tilde{\boldsymbol{\beta}}_l)$ and $i^* = i(\tilde{\boldsymbol{\beta}}_l) \in \tilde{I}_k$ such that for all $\tilde{\boldsymbol{\beta}} \in \{\check{\boldsymbol{\beta}}_m : m = 1, \ldots, \check{s}\}$:

$$\boldsymbol{x}'_{k^*,i^*}\tilde{\boldsymbol{\beta}}_l = \boldsymbol{x}'_{k^*,i^*}\check{\boldsymbol{\beta}} \qquad \Rightarrow \qquad \tilde{\boldsymbol{\beta}}_l = \check{\boldsymbol{\beta}}.$$

Considering the marginal mixture distribution for $k^* := k(\tilde{\boldsymbol{\beta}}_1)$ and $i^* := i(\tilde{\boldsymbol{\beta}}_1)$, we have for all $l \in \{1, \ldots, \check{s}\}$:

$$\check{\boldsymbol{\beta}}_l \neq \tilde{\boldsymbol{\beta}}_1 \qquad \Rightarrow \qquad \boldsymbol{x}'_{k^*,i^*}\check{\boldsymbol{\beta}}_l \neq \boldsymbol{x}'_{k^*,i^*}\tilde{\boldsymbol{\beta}}_1. \tag{9}$$

Thus, using condition (8),

$$\check{\pi}_1 = \sum_{\{h=1,\ldots,s:\, \boldsymbol{x}'_{k^*,i^*}\boldsymbol{\beta}_h = \boldsymbol{x}'_{k^*,i^*}\boldsymbol{\beta}_1\}} \pi_h$$

implying $\check{\pi}_1 > 0$.

Because of (7) — $\pi_1 \neq \check{\pi}_1$ — it must hold that there exists a $h \in \{2, \ldots, s\}$ with

$$\boldsymbol{\beta}_h \neq \boldsymbol{\beta}_1 \qquad \wedge \qquad \boldsymbol{x}'_{k^*,i^*}\boldsymbol{\beta}_h = \boldsymbol{x}'_{k^*,i^*}\boldsymbol{\beta}_1. \tag{10}$$

Without loss of generality one can assume that this h equals 2.

Consider $\boldsymbol{x}_{k,i} = \boldsymbol{x}_{k(\tilde{\boldsymbol{\beta}}_2),i(\tilde{\boldsymbol{\beta}}_2)}$ and apply the arguments above again to get that there exists an l with $\check{\boldsymbol{\beta}}_l = \tilde{\boldsymbol{\beta}}_2$. This leads to a contradiction between (9) and (10). Hence we have shown equation (4).

ad (d): As equality of distributions implies equality of means we have

$$\sum_{l=1}^{s} \pi_l \frac{e^{\boldsymbol{x}'_{k,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{k,j}\boldsymbol{\gamma}}}{\sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma}}} = \sum_{l=1}^{s} \pi_l \frac{e^{\boldsymbol{x}'_{k,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{k,j}\check{\boldsymbol{\gamma}} + \boldsymbol{x}'_{k,i}\boldsymbol{\delta}}}{\sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{h,j}\check{\boldsymbol{\gamma}} + \boldsymbol{x}'_{h,i}\boldsymbol{\delta}}}$$

for all $k = 1, \ldots, K$, $j \in J_i$ and $i \in I$. The equation can be transformed to:

$$\sum_{l=1}^{s} \pi_l e^{\boldsymbol{x}'_{k,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{k,j}\boldsymbol{\gamma}} \left[\left(\sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma}} \right)^{-1} - \left(\sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{h,j}\boldsymbol{\gamma} + (\boldsymbol{x}_{h,i} - \boldsymbol{x}_{k,i})'\boldsymbol{\delta} + (\boldsymbol{z}_{h,j} - \boldsymbol{z}_{k,j})'\boldsymbol{\vartheta}} \right)^{-1} \right] = 0$$

for all $k = 1, \ldots, K$, $j \in J_i$ and $i \in I$ with $\boldsymbol{\vartheta} := \check{\boldsymbol{\gamma}} - \boldsymbol{\gamma}$.

For every $j \in J_i$ and $i \in I$ there can be a $\tilde{u}_{ij}$ defined with $\tilde{u}_{ij} = \arg\max_{k=1,\ldots,K}\{\boldsymbol{x}'_{k,i}\boldsymbol{\delta} + \boldsymbol{z}'_{k,j}\boldsymbol{\vartheta}\}$. We will contradict the assumption that there exists a $k \in \{1, \ldots, K\}$ such that

$$\boldsymbol{x}'_{k,i}\boldsymbol{\delta} + \boldsymbol{z}'_{k,j}\boldsymbol{\vartheta} \neq \mathbf{0}. \tag{11}$$

This assumption together with the normalization condition (3), which implies $\boldsymbol{x}'_{K,i}\boldsymbol{\delta} + \boldsymbol{z}'_{K,j}\boldsymbol{\vartheta} = 0$, gives that there exists a $\tilde{v}_{ij} \in \{1, \ldots, K\}$ for which $\boldsymbol{x}'_{\tilde{u}_{ij},i}\boldsymbol{\delta} + \boldsymbol{z}'_{\tilde{u}_{ij},j}\boldsymbol{\vartheta} > \boldsymbol{x}'_{\tilde{v}_{ij},i}\boldsymbol{\delta} + \boldsymbol{z}'_{\tilde{v}_{ij},j}\boldsymbol{\vartheta}$ holds. Therefore we get $\sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{k,j}\boldsymbol{\gamma}} > \sum_{h=1}^{K} e^{\boldsymbol{x}'_{h,i}\boldsymbol{\beta}_l + \boldsymbol{z}'_{k,j}\boldsymbol{\gamma} + (\boldsymbol{x}_{h,i} - \boldsymbol{x}_{\tilde{u}_{ij},i})'\boldsymbol{\delta} + (\boldsymbol{z}_{h,j} - \boldsymbol{z}_{\tilde{u}_{ij},j})'\boldsymbol{\vartheta}}$ for all $l = 1, \ldots, s$. This leads to a contradiction of assumption (11), because a linear combination of negative numbers using only positive coefficients cannot give 0. This means that $\boldsymbol{x}'_{k,i}\boldsymbol{\delta} - \boldsymbol{z}'_{k,j}\boldsymbol{\vartheta} = \mathbf{0}$ for all $k = 1, \ldots, K$, $j \in J_i$ and $i \in I$. Because of condition (2) $\boldsymbol{\delta} = \boldsymbol{\vartheta} = \mathbf{0}$ follows. Hence we get $\check{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$ and $\check{\boldsymbol{\beta}}_l = \boldsymbol{\beta}_l$ for all $l = 1, \ldots, s$ and $k = 1, \ldots, K$.

Acknowledgements

We thank the referees for valuable suggestions which have helped to improve the manuscript. This research was supported by the Austrian Academy of Sciences (ÖAW) through a DOC-FFORTE scholarship and the Austrian Science Foundation (FWF) under grants P17382 and T351.

References

Alan Agresti. *Categorical Data Analysis*. Wiley, first edition, 1990.

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

W. R. Blischke. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59(306):510–528, June 1964.

C. Mitchell Dayton and George B. Macready. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401):173–178, March 1988.

Ryan T. Elmore and Shaoli Wang. Identifiability and estimation in finite mixture models with multinomial components. Technical Report 03-04, Department of Statistics, Pennsylvania State University, April 2003.

Dean A. Follmann and Diane Lambert. Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference*, 27:375–381, 1991.

Sylvia Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, 2006. ISBN 0-387-32909-9.

William H. Greene. *LIMDEP econometric modeling guide: Version 8.0*. Econometric Software, Plainview, NY, 2002.

William H. Greene and David A. Hensher. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B*, 37(8):681–698, September 2003.

Bettina Grün. *Identification and Estimation of Finite Mixture Models*. PhD thesis, Institut für Statistik und Wahrscheinlichkeitstheorie, Technische Universität Wien, September 2006. Friedrich Leisch, advisor.

Bettina Grün. Identifizierbarkeit von multinomialen Mischmodellen. Master's thesis, Technische Universität Wien, 2002. Kurt Hornik and Friedrich Leisch, advisors.

Bettina Grün and Friedrich Leisch. Bootstrapping finite mixture models. In Jaromir Antoch, editor, *Compstat 2004 — Proceedings in Computational Statistics*, pages 1115–1122. Physica Verlag, Heidelberg, 2004. ISBN 3-7908-1554-3.

Bettina Grün and Friedrich Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, pages 5247–5252, 2007.

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

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

Dipak C. Jain, Naufel J. Vilcassim, and Pradeep K. Chintagunta. A random-coefficients logit brand-choice model applied to panel data. *Journal of Business & Economic Statistics*, 12(3): 317–328, July 1994.

Wagner A. Kamakura and Gary J. Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, 26:379–390, November 1989.

Wagner A. Kamakura and Michel Wedel. An empirical Bayes procedure for improving individual-level estimates and predictions from finite mixtures of multinomial logit models. *Journal of Business and Economic Statistics*, 22(1):121–125, 2004.

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

Bruce G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. The Institute for Mathematical Statistics, Hayward, California, 1995.

Jean-Michel Marin, Kerrie Mengersen, and Christian P. Robert. Bayesian modelling and inference on mixtures of distributions. In Dipak Dey and C.R. Rao, editors, *Bayesian Thinking, Modeling and Computation*, volume 25 of *Handbook of Statistics*, chapter 16, pages 459–507. North–Holland, Amsterdam, 2005.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

Erik Meijer and Jelmer Y. Ypma. A simple identification proof for a mixture of two univariate normal distributions. *Journal of Classification*, 2008.

Linda K. Muthén and Bengt O. Muthén. *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, fourth edition, 1998–2006.

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

S. Rabe-Hesketh, A. Skrondal, and A. Pickles. GLLAMM manual. Working Paper Series 160, U.C. Berkeley Division of Biostatistics, 2004.

David Revelt and Kenneth Train. Mixed logit with repeated choices: Households' choices of appliance efficiency level. *The Review of Economics and Statistics*, 80(4):647–657, 1998.

Ehsan S. Soofi. A generalizable formulation of conditional logit with diagnostics. *Journal of the American Statistical Association*, 87(419):812–816, 1992.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 34:1265–1269, 1963.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

Jeroen K. Vermunt and Jay Magidson. *Latent GOLD Choice User's Guide*. Statistical Innovations Inc., Boston, 2003.

Michel Wedel. *GLIMMIX—A program for estimation of latent class mixture and mixture regression models, Version 3.0*. ProGAMMA bv, Groningen, The Netherlands, 2002.

Michel Wedel and Wagner S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12:21–55, 1995.

Michel Wedel and Wagner A. Kamakura. *Market Segmentation — Conceptual and Methodological Foundations*. Kluwer Academic Publishers, second edition, 2001. ISBN 0-7923-8635-3.

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

C. F. Jeff Wu. On the converence properties of the EM algorithm. *The Annals of Statistics*, 11 (1):95–103, 1983.