Kneib, Baumgartner, Steiner:

Semiparametric Multinomial Logit Models for Analysing Consumer Choice Behaviour

Sonderforschungsbereich 386, Paper 501 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Semiparametric Multinomial Logit Models for Analysing Consumer Choice Behaviour

Thomas Kneib*
Department of Statistics
Ludwig-Maximilians-University Munich
thomas.kneib@stat.uni-muenchen.de

Bernhard Baumgartner
University of Regensburg
bernhard.baumgartner@wiwi.uni-regensburg.de

Winfried J. Steiner
University of Regensburg
winfried.steiner@wiwi.uni-regensburg.de

Abstract

The multinomial logit model (MNL) is one of the most frequently used statistical models in marketing applications. It allows to relate an unordered categorical response variable, for example representing the choice of a brand, to a vector of covariates such as the price of the brand or variables characterising the consumer. In its classical form, all covariates enter in strictly parametric, linear form into the utility function of the MNL model. In this paper, we introduce semiparametric extensions, where smooth effects of continuous covariates are modelled by penalised splines. A mixed model representation of these penalised splines is employed to obtain estimates of the corresponding smoothing parameters, leading to a fully automated estimation procedure. To validate semiparametric models against parametric models, we utilise proper scoring rules and compare parametric and semiparametric approaches for a number of brand choice data sets.

Key words: mixed models, multinomial logit model, brand choice, penalised splines, proper scoring rules, semiparametric regression

*We thank Cornelia Oberhauser for computational assistance regarding both the development of the presented methodology and the application to the brand choice data sets. The first author received financial support from the German Science Foundation, Collaborative Research Center 386 "Statistical Analysis of Discrete Structures".

1 Introduction

Since its introduction to marketing by McFadden (1974, 1980), the multinomial logit model (MNL) has become one of the most popular models in marketing applications. In particular, it has regularly been used to identify and quantify the influence of price, promotional activities and consumer loyalty on brand choice (e.g. Guadagni and Little, 1983, or Ailawadi et al., 1999). In nearly all empirical applications of the MNL model, the consumer's utility associated with a brand chosen has been assumed to be a linear function of explanatory variables, although possible nonlinear effects have been suggested by several marketing theories, as briefly described in the following.

According to *adaptation level theory* (Helson, 1964), an individual's perception of a new stimulus is formed relative to a reference value (an adaptation level) formed through experience (e.g., Blattberg and Neslin, 1990). Applied to brand choice problems, this theory suggests that consumers compare observed prices to internal reference prices, determined by previous prices to which the individual has been exposed. Some authors have therefore included additional covariates in brand choice models reflecting the discrepancy between observed prices and reference prices (e.g. Kalwani et al., 1990, or Kalyanaram and Little, 1994). Sherif and Hovland (1961) postulate the existence of a region of indifference or "latitude of acceptance" surrounding the reference point. The difference between the observed price and the reference point is assumed to be underestimated (assimilated) if the price lies within this region, while larger discrepancies are overestimated (contrasted) and therefore perceived larger than they really are. If this theory (known as *assimilation-contrast theory*) holds, we would expect a consumer's utility to be a sigmoid function of price difference. Observed prices above the reference price are perceived as "losses", while observed prices below the reference price are considered as "gains". Kahnemann and Tversky (1979) observed that individuals tend to overstate losses as compared to gains. This bias in perception will result in a steeper function when consumers observe a price above their reference price.

In addition to these reference price theories supporting non-linear price response, there are further arguments to assume nonlinear effects of marketing variables in brand choice models. For example, we may expect a saturation effect on a consumer's utility for increasing or decreasing levels of brand attributes other than price, if considered in the model. For other important predictors of brand choice, there is at least an exploratory justification to look for possible non-linear effects. For example, there is no reason to expect a strictly linear effect of a consumer's brand loyalty on her/his brand utility so that a more flexible model, allowing for non-linear influences, is a valuable tool to validate the linearity assumption.

One possibility to overcome strict linearity for the consumer's utility function are nonlinear transformations of predictor variables which still yield utility functions linear in parameters. Krishnamurthi and Raj (1988) used a logarithmic transformation of prices, while Tellis (1988) compared models with logarithms and quadratic functions of advertising exposure as well as interaction

terms between ad exposure and loyalty. Piecewise linear utility functions have been applied by, for example, Kalyanaram and Little (1994), Ben-Akiva and Lerman (1985) or Wedel and Leeflang (1998). The main advantage of nonlinear transformations or piecewise linear functions is that the parameters can still be estimated by ordinary maximum likelihood using standard software. However, such approaches have the fundamental drawback that the functional form is predetermined by the choice of possible transformations or by boundaries (knots) between linear pieces, respectively. For piecewise linear functions, extensive search algorithms may be used to determine the number and location of knots instead of fixing them prior to estimation (see, e.g. Steinberger, 2001).

Abe (1998, 1999) introduced a framework to estimate semiparametric utility functions within the MNL. With this methodology, called GAM-MNL, splines or kernels can be used to model nonlinear influences on the response variable. While this model requires no a priori assumption about the functional form, the amount of smoothness for each function must be fixed before estimation. Hence, in case of several metric covariates, the selection of an optimal model has to be achieved by a high-dimensional grid search-type algorithm based on a number of predetermined smoothing parameter values.

We propose a semiparametric extension of the MNL that combines flexibility in terms of non-linear effects of covariates with automatic smoothing parameter selection. The strictly linear predictor is replaced by an additive predictor consisting of several smooth functions of continuous covariates which are parsimoniously represented by penalised splines. Estimation of smoothing parameters is based on a mixed model representation of penalised splines. In Gaussian and univariate exponential family regression models, the idea of representing penalisation approaches as mixed models has gained considerable attention in recent years (e.g. Ruppert, Wand and Carroll, 2003, Fahrmeir, Kneib and Lang, 2004, or Kauermann, 2006). The advantage is that smoothing parameters in the original model formulation transform into variance components in the mixed model, and mixed model methodology can be applied for their marginal likelihood estimation. Kneib and Fahrmeir (2006) extended mixed model based inference to categorical regression models but focused on cumulative regression models for ordered responses. Furthermore, they did not allow for category-specific covariates in the MNL which are clearly needed in our application. Our approach overcomes these limitations and additionally allows for changing choice sets, due to non-availability of some of the alternatives. The approach is implemented in the software package BayesX, which is available free of charge from `http://www.stat.uni-muenchen.de/~bayesx`.

When replacing simpler models with more flexible, complicated ones, the latter should be justified not only by theoretical considerations but also by assessing their performance. We will therefore evaluate the predictive performance of the semiparametric models compared to the usual linear versions based on a cross validation approach. The notion of proper scoring rules (see Gneiting and Raftery, 2005) will give us the possibility to select sensible scores for the evaluation of predictions. We will therefore give a short review about proper scoring rules and compare several scores in our application.

The rest of the paper is organised as follows: Section 2 recalls some details of multinomial logit models and introduces semiparametric extensions. Section 3 describes the mixed model representation and the resulting estimation scheme, while Section 4 focuses on model assessment based on proper scoring rules. Finally, Section 5 presents results of semiparametric models for several applications of consumer choice behaviour analysis and the concluding Section 6 comments on directions of future research.

2 Semiparametric Multinomial Logit Models

2.1 Multinomial Logit Models

When analysing consumer choice behaviour, the aim is to relate the response variable 'brand choice' to covariates. Hence, from a statistical perspective the response variable is given by a nominal categorical variable $Y \in \{1, \ldots, k\}$ with unordered categories. The most prominent regression model for this situation is the multinomial logit model (MNL, McFadden, 1974, Fahrmeir and Tutz, 2001, Ch. 3) which can be derived from considering latent (unobservable) utilities of the form

$$l^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta + \varepsilon^{(r)} = \eta^{(r)} + \varepsilon^{(r)}, \quad r = 1, \ldots, k, \tag{1}$$

where u and $w^{(r)}$ are global and category-specific covariates, respectively, and $\alpha^{(r)}$ and δ denote the corresponding regression coefficients. In an MNL the error terms $\varepsilon^{(r)}$ are independent across the categories and assumed to be standard extreme value distributed.

In our applications, the latent utilities will represent the gain associated with the choice of a certain brand. Hence, assuming rational behaviour, an individual chooses the brand that maximises her/his utility, i.e.

$$Y = r \quad \Leftrightarrow \quad l^{(r)} = \max_{s=1,\ldots,k} l^{(s)}.$$

Under the MNL it follows that the probability for observing $Y = r$ is given by

$$\pi^{(r)} = P(Y = r) \propto \exp(\eta^{(r)}).$$

Since the probabilities have to sum up to one, appropriate identifiability restrictions have to be imposed on redundant parameters. Without loss of generality, we choose the last category as reference category and assume $\alpha^{(k)} = 0$ and $w^{(k)} = 0$. The latter can be achieved by simply redefining $w^{(r)}$ as $w^{(r)} - w^{(k)}$, i.e. we only consider contrasts to the reference category. This finally leads to the model

$$\pi^{(r)} = \frac{\exp(\eta^{(r)})}{1 + \sum_{s=1}^{k-1} \exp(\eta^{(s)})}, \quad r = 1, \ldots, k-1, \tag{2}$$

and

$$\pi^{(k)} = 1 - \pi^{(1)} - \ldots - \pi^{(k-1)}.$$

The standard model formulation requires that all categories are available for all observations, i.e. $\pi_i^{(r)} > 0$ for $i = 1, \ldots, n$ and $r = 1, \ldots, k$. However, in our applications we know that for some of the observations some of the brands are not available and therefore $\pi_i^{(r)} = 0$ for some i and r. From a theoretical perspective it is easy to fix this problem by simply setting some of the probabilities in (2) to zero. In practice, we can achieve this either by introducing offset terms that effectively set some of the probabilities to zero or by excluding some of the probabilities from (2) using availability indicators. While the former approach is simpler to implement, the latter has the advantage of being more exact. We compared both approaches in our applications and found almost no differences. The results presented in Section 5 are based on the second strategy.

To account for possibly non-linear covariate effects influencing consumer choice behaviour, we extend the parametric utility model (1) to a semiparametric model, i.e. we replace the parametric predictor with a semiparametric predictor of the form

$$\eta^{(r)} = u'\alpha^{(r)} + w^{(r)\prime}\delta + f_1^{(r)}(x_1) + \ldots + f_q^{(r)}(x_q) + f_{q+1}(x_{q+1}^{(r)}) + \ldots + f_p(x_p^{(r)}). \quad (3)$$

Again we have to differentiate between global covariates $x_1, \ldots, x_q$ with category-specific effects and category-specific covariates $x_{q+1}^{(r)}, \ldots, x_p^{(r)}$ with global effects. To ensure identifiability, the assumptions from above have to be extended to $f^{(k)}(x_j) = 0$, $j = 1, \ldots, q$, and $f_j(x_j^{(k)}) = 0$, $j = q+1, \ldots, p$. The latter can in principle be achieved by redefining the corresponding functions $f_j(x_j^{(r)})$ as $f_j(x_j^{(r)}) - f_j(x_j^{(k)})$. We will discuss later-on in this section how to include both conditions into the estimation framework.

2.2 Penalised Splines

To model the smooth functions in (3) we employ penalised splines as proposed by Eilers and Marx (1996) since they provide a flexible and parsimonious representation of non-linear covariate effects. To simplify notation, we will suppress the category index in the following discussion, i.e. we consider functions $f(x)$. The basic idea of penalised splines is to represent $f(x)$ as a polynomial spline of degree l based on a moderately large number of B-spline basis functions, i.e.

$$f(x) = \sum_{j=1}^{d} \beta_j B_j(x). \quad (4)$$

This, in principle, recasts the semiparametric model (3) into a large linear model, since (4) is linear in the parameters, but instead of estimating the B-spline coefficients β_j unrestricted via maximum likelihood, a penalty term is added to regularise the estimation problem. To be more specific, we want to ensure that the function estimate $\hat{f}(x)$ despite being flexible is not too wiggly. From B-spline theory (see e.g. de Boor, 1993) we know that k-th order derivatives of B-splines

depend essentially on k-th order differences of the sequence of parameters β_j, $j = 1, \ldots, d$. Therefore, if we want to ensure smooth function estimates in terms of the k-th derivative, the log-likelihood has to be augmented by a penalty term constructed based on squared k-th order differences, e.g. as

$$\frac{1}{2\tau^2} \sum_{j=2}^{d} (\beta_j - \beta_{j-1})^2$$

for first order differences or

$$\frac{1}{2\tau^2} \sum_{j=3}^{d} (\beta_j - 2\beta_{j-1} + \beta_{j-2})^2$$

for second order differences. The smoothing parameter τ^2 controls the trade-off between fidelity to the data (τ^2 large) and smoothness (τ^2 small).

In vector notation, function evaluations $f(x)$ can be represented as $f(x) = v'\beta$, where $v = (B_1(x), \ldots, B_d(x))'$ and $\beta = (\beta_1, \ldots, \beta_d)'$. Similarly, the penalty terms can be rewritten as quadratic forms $1/(2\tau^2)\beta' K \beta$, where the penalty matrix $K = D'D$ is constructed from difference matrices D of appropriate order. Applying the vector notation to the model (3) now provides us with a possibility to include the condition $f_j(x_j^{(k)}) = 0$, $j = q+1, \ldots, p$. Since $f_j(x_j^{(k)}) = v_j^{(k)\prime}\beta_j$, we can simply redefine the corresponding function as $f_j(x_j^{(r)}) = (v_j^{(r)} - v_j^{(k)})'\beta_j$. Note that this is not equivalent to using the modified covariate $x_j^{(r)} - x_j^{(k)}$ and defining $f_j(x_j^{(r)})$ as $f_j(x_j^{(r)} - x_j^{(k)})$. For functions $f_j^{(r)}(x_j) = v_j'\beta_j^{(r)}$ of globally defined covariates we simply have to set $\beta_j^{(k)} = 0$ to ensure identifiability.

3 Mixed Model Based Inference

Estimation of semiparametric multinomial logit models is based on a penalised likelihood approach, i.e.

$$l_{\text{pen}}(\alpha, \delta, \beta) = l(\alpha, \delta, \beta) - \sum_{r=1}^{k-1} \sum_{j=1}^{q} \frac{1}{2(\tau_j^{(r)})^2} \beta_j^{(r)\prime} K_j \beta_j^{(r)} - \sum_{j=q+1}^{p} \frac{1}{2\tau_j^2} \beta_j' K_j \beta_j \quad (5)$$

has to be maximised with respect to the regression coefficients $\alpha = (\alpha^{(1)\prime}, \ldots, \alpha^{k-1\prime})'$, δ and $\beta = (\beta_1^{(1)\prime}, \ldots, \beta_q^{k-1\prime}, \beta_{q+1}', \ldots, \beta_p')'$. Note that the likelihood $l(\alpha, \delta, \beta)$ can in fact be evaluated as for usual parametric multinomial logit models since despite of the penalty term our semiparametric models are still linear in the regression coefficients. Maximisation of (5) can be achieved by augmenting the usual iteratively weighted least squares (IWLS) algorithm with appropriate penalty terms, compare e.g. Fahrmeir and Tutz (2001, Ch. 5). However, the crucial question on how to select the smoothness parameters τ^2 remains. Therefore we will now introduce a different perspective on semiparametric regression

models that allows to determine the smoothness parameters based on mixed model methodology.

For the sake of simplicity we will again restrict ourselves to one nonparametric function $f(x)$ and drop the category indices. As we have seen, $f(x)$ can be written as a linear combination of scaled basis functions, i.e. $f(x) = v'\beta$. The vector of regression coefficients is assigned a quadratic penalty term, which, from a Bayesian perspective, is equivalent to assuming that β is multivariate Gaussian distributed with density

$$p(\beta|\tau^2) \propto \left(\frac{1}{2\pi\tau^2}\right)^{\frac{\mathrm{rk}(K)}{2}} \exp\left(-\frac{1}{2\tau^2}\beta' K \beta\right). \tag{6}$$

From a frequentist perspective this corresponds to the assumption that β is a correlated vector of random effects with (6) as random effects distribution. Note however, that in contrast to classical mixed models the distribution of β is partially improper since $\mathrm{rk}(K) < \dim(\beta)$. In principle, we might now employ mixed model based estimation schemes for the determination of variance parameters such as marginal likelihood estimation. However, to obtain a valid mixed model representation, we have to reparameterise β to obtain a mixed model with proper random effects distribution. This can be achieved by decomposing the vector β as

$$\beta = X\gamma + Zb \tag{7}$$

where the dimension of b is given by the rank of the penalty matrix, i.e. $\dim(b) = \mathrm{rk}(K)$ while the dimension of γ is given by $\dim(\gamma) = \dim(\beta) - \mathrm{rk}(K)$. The goal of this decomposition is to achieve an explicit differentiation between penalised and unpenalised coefficients. Since K does not have full rank, a part of β remains unpenalised while another part is penalised. However, the penalised and the unpenalised part are fused together in a complicated way in the vector β. In contrast, after applying (7) we can explicitly differentiate between penalised coefficients (contained in the vector b) and unpenalised coefficients (contained in the vector γ). In addition, if we choose appropriate design matrices in (7) we can achieve

$$p(\gamma) \propto const \quad \text{and} \quad b \sim N(0, \tau^2 I),$$

i.e. γ represents a vector of fixed effects and b represents a vector of i.i.d. random effects with common variance τ^2. Compare Kneib and Fahrmeir (2006) for details on how to construct the design matrices X and Z.

Applying decomposition (7) to all nonparametric components in our semiparametric MNL, yields a mixed model representation where all random effects are proper. Now we can apply mixed model methodology to estimate not only the regression coefficients but also to automatically determine the smoothness parameters. We utilised a marginal likelihood approach which can be interpreted as a generalisation of the well-known restricted maximum likelihood approach to situations with nonnormal responses. This involves a Laplace approximation to the likelihood of the multinomial logit model, leading to a multivariate working linear model with working observations and working weights as in the IWLS

algorithm for the determination of the regression coefficients (compare Kneib and Fahrmeir, 2006, and Fahrmeir and Tutz, 2001, Ch. 3 for details).

Finally, our estimation scheme can be summarized as follows: Iteratively update the regression coefficients for given variances using a penalised version of IWLS and the variances based on a working linear mixed model given the current estimates of the regression coefficients. Upon convergence, the algorithm returns the penalised maximum likelihood estimates for the regression coefficients and marginal likelihood estimates for the smoothness parameters.

4 Predictions and Proper Scoring Rules

To evaluate the performance of the proposed semiparametric MNL in our applications, we will employ a prediction oriented approach. Therefore we divided each of the data sets into two parts, estimated the model based on the first part and predicted consumer choices for the second part. Now the question arises, how to actually evaluate the predictive performance of an estimated model. Usual approaches include computation of the hit rate (i.e. the percentage of true positive predictions) or the log-likelihood in the prediction sample. However, it is not generally clear, which of these measures is preferable or which properties a suitable measure should have.

To discuss this question more generally, we will now consider the notion of proper and strictly proper scoring rules as described in Gneiting and Raftery (2005). First of all, we have to define what we understand by a prediction. In fact, a useful prediction should not only consist of a point prediction but of a whole predictive distribution. In case of the MNL this predictive distribution is simply obtained by computing the probabilities for all k categories of the response according to the estimated model, i.e. we obtain the predictive distribution $\hat{\pi} = (\hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(k)})$. Now a scoring rule is any real-valued function $S(\hat{\pi}, r)$ that assigns a value to the event that category r is observed when $\hat{\pi}$ is the predictive distribution. Hence, a suitable score is obtained by the sum

$$S = \sum_{i=1}^{n} S(\hat{\pi}_i, r_i)$$

where we sum over all observations in the test data set, $\hat{\pi}_i$ denotes the predictive distribution derived from the estimated model for observation i and r_i is the truly observed value in the test data set.

To compare different scoring rules, Gneiting and Raftery (2005) consider the expected value of the score under the true distribution π_0 and denote it by $S(\hat{\pi}, \pi_0)$. Then a scoring rule is called proper if $S(\pi_0, \pi_0) \leq S(\hat{\pi}, \pi_0)$ for any predictive distribution $\hat{\pi}$. The scoring rule is called strictly proper if equality holds if and only if $\hat{\pi} = \pi_0$. Gneiting and Raftery present theoretical results characterising (strictly) proper scoring rules and discuss a number of examples.

Let us consider first the so-called hit rate, i.e.

$$S(\hat{\pi}_i, r_i) = \begin{cases} \frac{1}{n} & \text{if } \hat{\pi}^{(r_i)} = \max\{\hat{\pi}^{(1)}, \ldots, \hat{\pi}^{(k)}\}, \\ 0 & \text{otherwise.} \end{cases}$$

Gneiting and Raftery (2005) show that this scoring rule is proper but not strictly proper. It is also quite obvious that we are only using a very small part of the information contained in the predictive distribution. The scoring rule takes on the same value regardless of the exact form of the predictive distribution as long as the category with the maximum value remains the same. In particular, the scoring rule does not take into account the variability in the predictive distribution.

A second popular scoring rule is obtained from the log-likelihood contributions and is called the logarithmic score:

$$S(\hat{\pi}, r_i) = \log(\hat{\pi}^{(r_i)}).$$

This scoring rule is strictly proper but still has the drawback that it involves only one of the probabilities of the predictive distribution. Therefore the logarithmic score is sensitive with respect to extreme observations.

Two further strictly proper scoring rules are given by the Brier score

$$S(\hat{\pi}, r_i) = -\sum_{r=1}^{k} \left(\mathbb{1}(r_i = r) - \hat{\pi}^{(r)} \right)^2$$

and the spherical score

$$S(\hat{\pi}, r_i) = \frac{\hat{\pi}^{(r_i)}}{\sqrt{\sum_{r=1}^{k} (\hat{\pi}^{(r)})^2}}.$$

Especially the Brier score is a popular choice to compare categorical regression model that circumvents the problems discussed for the hit rate and the logarithmic score.

In our applications we will utilise all four scores to compare the predictive performance of semiparametric MNL models to parametric approaches.

5 Empirical Application: Consumer Choice Behavior

5.1 Data

In our empirical application, we use panel data from three product categories. One data set includes purchases of *coffee* over a time span of 52 weeks. The sample refers to the five largest brands in this category (in terms of market share), accounting for 53% of all purchase acts, and covers a total of 49.083 purchases

from 6.407 households. We further apply the model to two US scanner panel data sets provided by A.C. Nielsen. These data sets consist of 26.820 purchase acts by 2.494 households in the category *ketchup* and 66.679 purchase acts by 2.317 households in the category *yogurt*. In the ketchup (yogurt) category, we considered the three (five) largest brands accounting for 87% (74%) of all purchases. The data provide information about the date of purchase acts and the brand chosen by the household as well as prices ("price") and promotional activities.

5.2 Specification of Loyalty and Reference Price Terms

Marketing research has revealed that consumers develop loyalities to particular brands. In brand choice models, therefore, variables reflecting brand loyalty are regularly included. To measure brand loyalty, we employed an approach introduced by Guadagni and Little (1983) where the loyalty for brand r of household i at the k-th buying occasion is recursively defined as

$$loyalty_{ik}^{(r)} = \varphi_L \, loyalty_{i,k-1}^{(r)} + (1 - \varphi_L) \, \mathbb{1}(Y_{i,k-1} = r).$$

Hence, loyalty at the k-th buying occasion is an exponentially weighted average of past purchases of the same brand, with the constant φ_L determining the persistence of loyalty.

The approach of Guadagni and Little has shown its high ability to increase model fit and prediction in a number of applications and is widely used in the marketing literature to capture brand loyalty. The smoothing constant φ_L, estimated by a grid search within the interval [0,1], is 0.75, 0.74, and 0.48 for the brand types coffee, ketchup and yogurt, respectively.

Similarly, we obtained the reference price for brand r of household i at the k-th buying occasion via (see, e.g., Kalyanaram and Little, 1994)

$$refprice_{ik}^{(r)} = \varphi_R \, refprice_{i,k-1}^{(r)} + (1 - \varphi_R) \, price_{i,k-1}^{(r)}.$$

Here, 0.57, 0.69 and 0.40 were estimated for the smoothing constant φ_R for the brand types coffee, ketchup and yogurt, respectively. For each data set, the first three purchases of each household were used for initialisation of *loyalty* and *refprice*, and were therefore excluded from estimation.

In our model, we consider both the reference price and the difference between the reference price and the observed price *(diffprice)* as covariates:

$$diffprice_{ik}^{(r)} = refprice_{ik}^{(r)} - price_{ik}^{(r)}.$$

An inclusion of price, reference price *and* price deviation would result in perfect collinearity. In accordance with most studies, we use the reference price and not the observed price, because the correlation between reference price and price deviation is considerably smaller (0.27, 0.30 and 0.14 for coffee, ketchup and yogurt) as compared to the correlation between reference price and observed price (0.64, 0.66 and 0.33 respectively).

For ketchup and yogurt, promotional activities are considered by using two dummy variables describing the presence (=1) or absence (=0) of advertising ("AdCode") and display activities ("Display") for each brand and purchase act. For coffee, the covariate "Promotion" covers the presence (=1) or absence (=0) of any kind of promotional activity. These dummies are included as fixed effects in our model.

Given the specifications for the reference price and loyalty terms stated above and the vector δ representing the effects of promotional dummies, we estimate the following model for a brand r (suppressing household and purchase occasion indices i and k):

$$\eta^{(r)} = \beta_0^{(r)} + w^{(r)\prime}\delta + f_1(\mathit{refprice}^{(r)}) + f_2(\mathit{diffprice}^{(r)}) + f_3(\mathit{loyalty}^{(r)})$$

5.3 Interpretation of Results

Figure 1 shows the estimated effects of *loyalty*, *refprice* and *diffprice* on a consumer's utility. As expected, a consumer's utility and therefore her/his choice probability increases with increasing loyalty. Moreover, the shape of the loyalty curve looks similar in all three product categories. Especially, we can observe a strong kink of the curves in the lower extreme of the loyalty range (around a loyalty value of approximately 0.2). This indicates that if a household buys a certain brand for the first time, the probability to purchase this brand in the future increases more strongly as compared to a household who already has bought the brand and therefore already has a higher loyalty. The marginal effect of *loyalty* further increases for loyalties above 0.95. This reflects the buying behaviour of some households who are very brand loyal and always buy the same brand.

For the effect of *refprice*, we obtain very smooth functions. As expected, the functions decrease in all three categories, reflecting a decline in brand choice probability as prices increase. However, the curves are differently shaped, ranging from a convex shape in the ketchup category and a rather linear shape in the coffee category to a concave shape in the yogurt category.

Concerning the discrepancy between observed prices and reference prices, we can not confirm theoretical suggestions as described in the introductory section. Accordingly, we would have expected an inverse s-shaped function. Especially, there is no flat region around a difference of zero (as predicted by the assimilation-contrast theory). We further recognise an unexpected decrease in utility for large gains in the coffee category as well as an unexpected increase for large losses in the ketchup category. These irregularities may be due to chance, however, as only less than one percent of all price differences (0.8% in the coffee category and 0.15% in the Ketchup category, respectively) lie within these areas. This is also reflected by the relatively wide confidence intervals for these parts of the estimated functions. The flat region for differences larger than one in the ketchup category may lead to the conclusion that large price decreases do not pay for the company. However, one must not ignore that there

is usually some collinearity between price gains and promotional variables, as promotional activities (like the use of display or advertising) are often accompanied by temporary price reductions. With regard to all observations (purchases) in the Ketchup category, only 7 percent of the brands were bought on display and only 16 percent were accompanied by advertising. When we consider only purchases with price gains larger than one, however, 15% of those purchases were made on display and 40% were accompanied by advertising.

Table 1 summarises the estimated effects for the promotional variables. In the product categories *ketchup* and *yogurt*, the promotional effects show the expected sign (i.e., if a promotion is offered for a brand, a consumer's utility is expected to increase). The negative sign of the promotional parameter in the product category *coffee* may be caused by collinearity with the variable *diffprice*. The average price difference is significantly higher for purchases of coffee under promotion (-26.92 for purchases not accompanied by a promotion versus 41.90 for purchases on promotion) accounting for an increase in utility which is three times higher than the promotional parameter.

	$\hat{\delta}_j$	$\text{sd}(\hat{\delta}_j)$	95% ci		p-value
			Coffee data		
Promotion	-0.311	0.030	-0.370	-0.252	<0.0001
			Ketchup data		
AdCode	0.898	0.051	0.798	0.998	<0.0001
Display	0.808	0.072	0.666	0.949	<0.0001
			Yogurt data		
AdCode	1.076	0.052	0.973	1.178	<0.0001
Display	0.556	0.108	0.345	0.767	<0.0001

Table 1: Estimated fixed effects.

Finally, Table 2 contains results for the four scoring rules discussed in the previous section for both the estimation and the prediction part of the data sets. Obviously, semiparametric models always lead to an improved score in the estimation data set due to their additional flexibility. However, this observation does not necessarily hold for the validation data, since more flexible models bear the risk of overfitting. Indeed, we observe this phenomenon for the coffee data, where all four scoring rules indicate a better predictive performance of the parametric model on the validation data. Hence, a purely parametric approach seems to be more suitable for this data set.

In contrast, the semiparametric MNL outperforms the parametric MNL for the ketchup data. Despite of the hit rate, all scoring rules indicate an improvement of the predictive performance when allowing for a more flexible model equation. In terms of the hit rate, both models perform equally well. Since, however, the hit rate is not strictly proper, it seems more plausible to rely on the results of the three strictly proper scoring rules.

Results for the yogurt data are somewhere in between the coffee and the ketchup data results. The hit rate as well as the logarithmic score prefer the parametric model while the Brier score and the spherical score assign improved performance to the semiparametric MNL. Again, since the hit rate and the logarithmic score have theoretical drawbacks as discussed in Section 4, we favour the Brier and the spherical score, indicating the need for a flexible model.

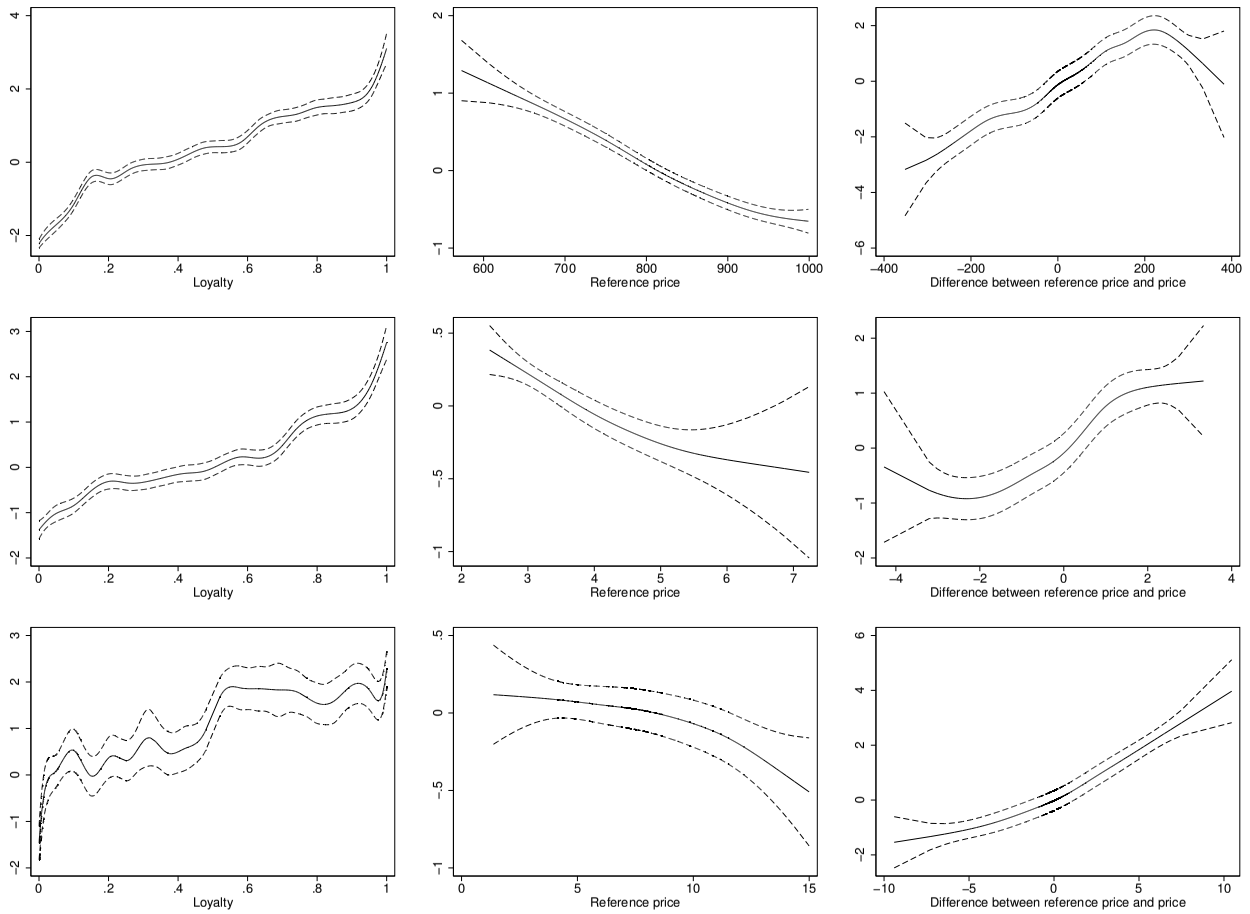

Figure 1: Estimated nonparametric effects for the coffee data (first row), the ketchup data (second row) and the yogurt data (third row) together with 95% pointwise confidence intervals.

	Coffee		Ketchup		Yogurt	
	parametric	semiparametric	parametric	semiparametric	parametric	semiparametric
Hit rate (estimation)	0.70	0.70	0.79	0.79	0.82	0.83
Hit rate (prediction)	0.70	0.66	0.78	0.78	0.83	0.81
Logarithmic Score (estimation)	-13816.90	-13491.45	-5146.61	-5024.40	-8502.60	-7923.95
Logarithmic Score (prediction)	-13955.80	-15682.87	-5297.58	-5225.32	-24061.49	-26588.13
Brier Score (estimation)	-6912.34	-6789.38	-5192.60	-5222.72	-4261.52	-4044.11
Brier Score (prediction)	-6930.30	-7646.83	-2990.25	-2962.46	-12919.96	-12416.39
Spherical Score (estimation)	12102.09	12181.02	6455.08	6450.31	12678.38	12798.71
Spherical Score (prediction)	12093.57	11588.11	7688.11	7701.58	37965.96	38231.85

Table 2: Scores evaluated for the estimation and the prediction sample.

6 Discussion

Adequate modelling of real-life phenomena frequently requires more flexible models than the usual regression models relying on parametric, linear predictors. In this paper we have described such a flexible regression model for the case of categorical, unordered response variables and successfully applied it to the analysis of consumer choice behaviour. A fully automated inferential scheme has been outlined, allowing for the joint determination of all model parameters without the need of subjective judgements. The notion of (strictly) proper scoring rules provided a useful tool to validate the (predictive) performance of the flexible models.

While our estimation scheme can be considered a frequentist approach based on a penalised likelihood, it is also conceptually equivalent to an empirical Bayes approach relying on posterior modes. Hence, it would be worthwhile to compare it to its fully Bayesian counterpart with estimation relying on Markov chain Monte Carlo (MCMC) simulation techniques. Brezger and Lang (2006) describe Bayesian semiparametric multinomial logit models but their approach does not allow for category-specific covariates and varying choice sets.

Along the lines of Brezger and Lang (2006) and Kneib and Fahrmeir (2006) our semiparametric MNL can also be extended to include further model components if required by the application at hand. For example, spatial effects can be based on similar penalisation approaches leading to geoadditive MNL models.

A further direction of future research is the inclusion of monotonicity or, more general, concurvity constraints. For example, we might want to restrict price effects to monotonic functions leading to a stabilisation of the estimation procedure in the tails of the price distribution where the number of observations is low. Brezger and Steiner (2007) describe a Bayesian approach to monotonic regression for the analysis of price response functions. However, their estimation procedure relies heavily on MCMC simulations and can not be straightforwardly adapted to our approach.

References

Abe, M. (1998). Measuring consumer nonlinear brand choice response to price. *Journal of Retailing* **74** 541–568.

Abe, M. (1999). A generalized additive model for discrete-choice data. *Journal of Business and Economic Statistics* **17** 271–284.

Ailawadi, K.L., Gedenk, K., Neslin, S.A. (1999). Heterogeneity and Purchase Event Feedback in Choice Models: An Empirical Comparison with Implications for Model Building. *International Journal of Research in Marketing* **16** 177–198.

Ben Akiva, M., Lerman, S. L (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* The MIT Press, Cambridge, Massachusetts.

Blattberg, R.C., Neslin, S.A. (1990). *Sales Promotion: Concepts, Methods and Strategies.* Prentice-Hall, Englewood Cliffs, N.J..

Brezger, A., Lang, S. (2006). Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* **50** 967–991.

Brezger, A., Steiner, W. (2007). Monotonic Spline Regression to Estimate Promotional Price Effects: A Comparison to Benchmark Parametric Models. *Journal of Business and Economic Statistics*, to appear.

Eilers, P. H. C., Marx, B. D. (1996). Flexible smoothing using B-splines and penalties (with comments and rejoinder). *Statistical Science* **11** 89–121.

Fahrmeir, L., Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models.* Springer, New York.

Fahrmeir, L., Kneib, T., Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica* **14** 731–761.

Gneiting, T., Raftery, A. E. (2005). Strictly proper scoring rules, prediction, and estimation. Technical Report 463R, Department of Statistics, University of Washington.

Guidagni, P. M., Little, J. D. C. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science* **11** 372–385.

Helson, H. (1964). *Adaptation-Level Theory.* Harper & Row, New-York, London.

Kahnemann, D., Tversky, A. (1979). A Prospect Theory: An Analysis of Decisions under Risk. *Econometrica* **47** 263–291.

Kalwani, M.U., Yim, C.K., Rinne, H.J., Sugita Y. (1990). A price expectations model of customer brand choice. *Journal of Marketing Research* **27** 251–262.

Kalyanaram, G., Little, J.D.C. (1994). An empirical analysis of latitude of price acceptance in consumer package goods. *Journal of Consumer Research* **21** 408–418.

Kauermann, G. (2006). Nonparametric models and their estimation. *Allgemeines Statistisches Archiv* **90** 135–150.

Kneib, T., Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics* **62** 109–118.

Krishnamurthi, L., Raj, S.P. (1988). A model of brand choice and purchase quantity price sensitivities. *Marketing Science* **7** 1–20.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics* Zarembka, P. (ed) 105–142.Academic Press, New-York, London

McFadden, D. (1980). Econometric models for probabilistic choice among Products. *Journal of Business* **53** 13–34.

Sherif, M., Hovland, C.I. (1961). *Social Judgement.* Yale University Press, New Haven, London.

Steinberger, M. (2001). *Multinomiale Logitmodelle mit linearen Splines zur Analyse der Markenwahl.* Peter Lang, Frankfurt am Main, Berlin.

Tellis, G.J. (1988). Advertising exposure, loyalty and brand purchase: a two-stage model of choice. *Journal of marketing research* **25** 134–144.

Wedel, M., Leeflang, P.S.H. (1998). A model for the effects of psychological pricing in Gabor-Granger price studies. *Journal of Economic Psychology* **19** 237–260.