CIRJE-F-193

Investigating the Competitive Assumption of Multinomial Logit Models of Brand Choice by Nonparametric Modeling

Makoto Abe
The University of Tokyo

Yasemin Boztug
Lutz Hildebrandt
Humboldt University of Berlin

February 2003

Investigating the Competitive Assumption of Multinomial Logit Models of Brand Choice by Nonparametric Modeling

Makoto Abe[1], Yasemin Boztuğ[2] and Lutz Hildebrandt[2]

[1] University of Tokyo, Faculty of Economics, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2] Humboldt University of Berlin, Institute of Marketing, Spandauer Str. 1, 10178 Berlin, Germany

Summary

The Multinomial Logit (MNL) model is still the only viable option to study nonlinear responsiveness of utility to covariates nonparametrically. This research investigates whether MNL structure of inter-brand competition is a reasonable assumption, so that when the utility function is estimated nonparametrically, the IIA assumption does not bias the result. For this purpose, the authors compare the performance of two comparable nonparametric choice models that differ in one aspect: one assumes MNL competitive structure and the other infers the pattern of brands' competition nonparametrically from data.

Keywords: nonparametric method, generalized additive models, brand choice, IIA, multinomial logit model, scanner panel data

1 Introduction

A multinomial logit (MNL) model of a qualitative response variable characterizes a choice from discrete (nominal) alternatives by a decision maker as a function of attributes associated with each alternative as well as the characteristics of the individual. Because of its analytical and computational tractability, the model has been applied extensively to discrete choice processes in such fields as econometrics (McFadden 1974, Manski and McFadden 1981), transportation (Ben–Akiva and Lerman 1985), and marketing (Guadagni and Little 1983) with great success. In marketing the advance of bar–code scanner technology has allowed large–scale household purchase records to be collected with ease and accuracy. Major marketing research firms such as A. C. Nielsen and IRI routinely collect purchase information from thousands of participating panelists. These panel data, incorporated with information provided by the store such as price and promotional activities of the competing products, offer a rather complete picture of consumers' purchase environments. The availability of such data has caused a revolution in the modeling of consumer brand choice. The application of MNL models to analyze data is now a part of daily operation in commercial firms.

The analytical tractability and ease of estimation have resulted in various extensions of a MNL model to relax its assumptions. One can accommodate consumer heterogeneity by means of latent class modeling (Kamakura and Russell 1989, Fusan and Srinivasan 1993) and individual–specific parameters (Rossi and Allenby 1993). Another extension models the nonlinear relationship of utility with covariates using a nonparametric utility function (Abe 1999). Both of these extensions have taken advantage of computational ease in estimating MNL parameters through the maximum likelihood method.

A major restriction of a MNL model is its competitive structure, reffered to as the independence from irrelevant alternatives (IIA) property, which arises from the i.i.d. error distribution. To overcome this restriction, two approaches were pursued. One is to further extend a MNL model to a nested logit model (Ben–Akiva and Lerman 1985) by recognizing its computational advantage. Its weakness is that the hierarchical, tree–like structure of a choice process, characterized by the competitive relationship of brands, must be specified a priori. Because the specification would introduce a certain degree of subjectivity, the model is not particularly suitable for deriving a competitive structure from data.

The other is to use a multinomial probit model that can relax the i.i.d. error distribution by assuming a non–diagnonal variance–covariance matrix for the stochastic component. Probit can accommodate the relaxation of the i.i.d. error distribution in a much more straightforward manner than logit's correlated multivariate normal distribution. However, because estimation of probit involves numerical integration of higher orders, the computational burden

prohibited its practical use until a simulation method was proposed recently (McFadden 1989, McCulloch and Rossi 1994). This simulation method is quite powerful in that it can estimate probit models that capture (1) competitive structure through a flexible form of the stochastic component and (2) consumer heterogeneity by permitting individual–specific parameter estimates (Rossi, McCulloch and Allenby 1996).

Among the three issues, (1) heterogeneity, (2) non–IIA, (3) nonparametric utility, a multinomial probit model combined with the simulation method can address the first two, but the last issue of a nonparametric utility function seems rather out of the question because of the amount of computation that involves tens of thousands of sequential random draws from a distribution for parameter estimate to stabilize.[1] Therefore, to investigate nonlinear relationship of utility with covariates, we still need to rely on a MNL model. The heterogeneity issue in MNL models can be addressed by either incorporating individual–specific covariates, a latent class modeling or individual–specific parameter models. But because the non–IIA extension cannot be addressed directly through the MNL framework, it is important that the IIA assumption is met when applying a nonparametric MNL model.

Previous studies in brand choice found that, as long as the response function of covariates is kept nonparametric, much of the benefit of fully nonparametric method modeling, both deterministic and random (noise/uncertainty) components of utility nonparametrically can be realized even if a parametric distributional assumption is imposed on the random component (Abe 1999, Briesch, Chintagunta and Matzkin 1997). These findings implicitly support the robustness of a nonparametric MNL model when investigating the nonlinear relationship of utility with covariates.

The objective of this manuscript is to investigate explicitly whether the MNL structure of inter–brand competition is a reasonable assumption, so that when the utility function is estimated nonparametrically, the IIA assumption does not bias the result. For this purpose, we compare the performance of two comparable nonparametric choice models that differ in one aspect: one assumes MNL competitive structure and the other infers the pattern of brands' competition nonparametrically from data.

The former model is a nonparametric MNL model proposed by Abe (1999), whose parametric counterpart is an ubiquitous MNL model with a linear–in–parameters deterministic component of utility. In the nonparametric version, the deterministic component is assumed to be additive in a one–dimensional nonparametric function of each covariate, and the difference of the random components of two brands has a logistic distribution.

[1] In fact from this very reason, the variance–covariance matrix of the stochastic component is currently limited to be only diagonal, and a fully general form of competitive structure cannot be realized (Rossi, McCulloch and Allenby 1996).

For the latter model, we chose a nonparametric logistic regression proposed by Hastie and Tibshirani (1986, 1987). It is based on the generalized additive model (GAM) that relates a response variable to an additive–in–nonparametric–covariates predictor via a logistic link function. By regressing a binary choice variable (indicating whether a brand is chosen or not) on marketing mix variables for that brand as well as for alternative brands, competitive marketing effect can be estimated nonparametrically. A single regression equation is estimated for each brand. Its parametric counterpart is the usual linear–in–parameters logistic regression. Aside from the fact that the stochastic component has the same logistic distribution as a MNL model, the reason for choosing this model is as follows.

Generalized extreme–value (GEV) models, a general class of parametric random utility choice models proposed by McFadden (1978), are not restricted by the IIA assumption. Their expression resembles that of a MNL model with the expception that a brand's utility depends not only on its own attributes but also on utilities of alternative brands in a complicated nonlinear fashion as shown below.

$$P_n(i) = \frac{e^{V_{in} + \ln G_i(\cdots)}}{\sum_{j=1}^{J_n} e^{V_{jn} + \ln G_j(\cdots)}} \tag{1}$$

where $G(\cdots)$ is a function of utilities of alternative brands possessing certain properties, and $G_j(\cdots)$ is the first derivative with respect to its j–th argument. Further details can be found in Ben–Akiva and Lerman (1985), equation (5.47). Hence, by introducing a nonparametric function of attributes for alternative brands, one can mimic the parametric GEV model that is not constrained by IIA or MNL competitive structure.

Our finding indicates that even though the estimation of a nonparametric MNL is biased by non–IIA data in a simulation setting, its result is quite robust in actual scanner data. In addition, if we relaxed the MNL assumption by letting data specify the competitive structure, a substantially larger amount of data, perhaps an increased order of magnitude, would be required. At least in brand choice modeling, therefore, nonparametric relaxation is useful only for utility specification but not the MNL structure itself unless the size of database becomes substantially larger than the one typically used by academic researchers.

In Section 2 the two nonparametric models and their estimation methods are described. Section 3 describes the result of a simulation study to compare the two models under a known competitive structure. In Section 4 these two nonparametric models are applied to German scanner panel data of brand choice in a health care product category and is followed by a discussion in Section 5.

2 Models

Let us describe the two nonparametric models whose error component is distributed logistically: one estimates the competitive structure nonparametrically and the other assumes a MNL competitive structure.

2.1 Estimating Competitive Effect — Nonparametric Logistic Regression

Since this model is based on GAM, whose idea originated from its parametric version, generalized linear models (GLM), let us describe GLM first. GLM (Nelder and Wedderburn 1972) generalize the standard linear methodology to accommodate diverse types of a response variable. GLM allow for a flexible relationship between a response variable y and a predictor index η, which is linear regarding parameters of explanatory variables $x_p (p = 1, 2, .., P)$ such that $\eta(x) = \sum_p \beta_p x_p$. The appropriate specification of the random component and the link function in GLM leads to various regression models such as usual multiple regression, logistic regression, a binary probit model, and log–linear models.

Generalized additive models (GAM) (Hastie and Tibshirani 1990) are nonlinear extensions of GLM, and relax the linear–in–parameters assumption to a sum of one–dimensional nonparametric functions of the explanatory variables. Thus, the predictor index takes a form $\eta(x) = \sum_p f_p(x_p)$. For example, the GAM for logistic regression of a binary response variable y is expressed as

$$E(y|x) \equiv \mu(x) = \Pr(x) = G\left(\sum_{p=1}^{P} f_p(x_p)\right) = G\left(\eta(x)\right) \tag{2}$$

where f_p is a nonparametric function of the p–th explanatory variable x_p and $G(\cdot)$ is a logistic link function of a form:

$$G(\eta(x)) = \frac{1}{1 + e^{-\eta(x)}}. \tag{3}$$

In modeling a choice of a particular brand, covariates can include marketing mix variables of that brand as well as those of alternative brands. Estimated nonparametric functions, $f_p(\cdot)$, for the brand's own covariates suggest how its pricing and promotion influence its choice, whereas estimated functions of covariates for alternative brands provide insights into the impact of competitive marketing activity on that brand. Hence, the model is not restricted by IIA but instead captures the competitive effect nonparametrically.

One drawback of the regression formulation is that, because a separate binary regression model is estimated for each brand, the sum of choice probabilities

over available brands does not become one. While this may not be problematic when interpreting the estimated nonparametric functions (Boztuğ and Hildebrandt 2001), it poses a logical inconsistency when predicting brand choice probabilities. A typical solution is to normalize probabilities so that they sum up to one for each purchase incident.

2.2 Imposing Competitive Structure — Nonparametric MNL Model

The choice probability of alternative j as expressed in a usual linear–in–parameters MNL model is

$$\Pr(j) = \frac{e^{v_j}}{\sum_k e^{v_k}} \quad \text{where} \quad v_j = \sum_p \beta_p x_{jp} \tag{4}$$

and x_{jp} denotes the p–th explanatory variable for alternative j. Our objective here is to obtain a MNL model with a flexible utility structure such that

$$\Pr(j) = \frac{e^{v_j}}{\sum_k e^{v_k}} \quad \text{where} \quad v_j = \sum_p f_p(x_{jp}) \tag{5}$$

Although similar in form to equation (3) for a binary case, its extension to a multinomial setting is not trivial. This can be seen by dividing the numerator and denominator of (5) by e^{v_j}:

$$\Pr(j) = \frac{1}{1+e^{-\eta(x)}} \quad \text{where} \quad \eta(x) = \sum_p f_p(x_{jp}) - \log\left\{\sum_{k\neq j} e^{\left(\sum_p f_p(x_{kp})\right)}\right\} \tag{6}$$

Notice that the predictor $\eta(x)$ is no longer additive in a function of each covariate, f_p, and does not conform to the logistic regression of GAM. Abe (1999) derived the nonparametric additive utility specification for MNL, shown in (5), from a generic formulation of GAM using a penalized likelihood function.

Note in order to be consistent with the random utility maximization assumption, the utility function of a brand cannot include covariates of other alternative brands (McFadden 1981). The cross–effect is driven by this assumption, and hence MNL models exhibit the IIA competitive structure.

2.3 Comparison of the Two Models

At this point, it is worthwhile to compare the two nonparametric models: one that is based on logistic regression and the other that is based on MNL.

The MNL formulation is built on the behavioral theory of IIA, which, in turn, specifies its competitive structure. For example, the effect of the price change of brand 2 on the choice of brand 1 is determined by the difference in utilities for the two brands through the MNL form expressed as

$$\Pr(j) = \frac{e^{v_j}}{\sum_k e^{v_k}}. \tag{7}$$

In the logistic regression formulation, on the other hand, there is no theory specifying the competitive structure. This leads to a more flexible model. In turn, the competitive effect must be captured from the data by introducing covariates of alternative brands. For example, to account for the cross–effect of the price change of brand 2 on the choice of brand 1, logistic regression for brand 1's choice must include a price variable for both brands 1 and 2. Therefore, the logistic regression formulation is more data–driven and nonparametric–oriented than the MNL formulation.

And for this very reason, the model is more prone to "the curse of dimensionality" problem in actual setting, which refers to an exponential increase in sample size to maintain the accuracy of an estimator as the complexity of the problem (e.g., the number of alternatives and covariates) increases (Silverman 1986). It remains to be seen in our empirical study how the tradeoff between model flexibility and data requirement turns out, relative to nonparametric MNL.

The parametric assumption of the random component is the same in both models. We assume a logistic link function, which results from an extreme value distribution of the error terms in the MNL model. The comparison of the two models is summarized in Table 1.

	MNL	**Logistic Regression**
Competitive Structure assumed[a]	IIA	None. Specified by data by including attributes for other alternatives.
Parametric Assumptions of the Random Component	Logistic link function	Logistic link function

[a] The two models have the same random component but differ in competitive structure.

Table 1: Comparison of the Two Nonparametric Choice Models

3 Simulation Study

3.1 Procedure

The purpose of this simulation is to investigate how well the nonparametric MNL and logistic regression models that respectively do and do not assume IIA, fit data sets that are and are not restricted by IIA. Please refer to Table 2.

For data that follow the IIA restriction (row 1), we expect both models to fit the data well (indicated by "$\circ$"). MNL fits well because the model assumption is consistent with the data. Logistic regression should also perform well because it does not presume a particular competitive structure, and nonparametric function η should be sufficiently flexible to fit to a variety of competitive structure. We are particularly interested in how well nonparametric logistic regression can recover the underlying IIA restriction in the data. For data that do not follow the IIA restriction (row 2), we expect that MNL – whose competitive assumption is incompatible – performs poorly ("$\times$"), whereas logistic regression can still fit the data well.

		Model (nonparametric)[a]	
		MNL	Logistic Regression
Data	IIA	$\circ$	$\circ$
	Non-IIA	$\times$	$\circ$

[a] Nonparametric MNL is expected to perform poorly on non–IIA data, where nonparametric logistic regression can fit both IIA and non–IIA data well.

Table 2: Data and Model in Simulation Study

3.2 Simulated Choice Data

Simulated brand choice data for two alternatives consisting of 1000 choice incidents were generated according to two processes: one that is restricted by IIA and the other that is not. From the two sets of generated data, one was used to calibrate the models and the other was used to test the predictive validation. We used two continuous variables for alternative j (where $j = 1$ or 2) as X_{lj} (e.g., loyalty) and X_{pj} (e.g., price), whose nonlinear response must be estimated by the nonparametric models. To make the simulation more challenging and realistic, we also introduced two binary indicator variables for alternative j, Z_{fj} (e.g., feature) and Z_{dj} (e.g., display).

IIA Data

The choice data with the IIA restriction were generated according to a MNL process of equation (5) with the following utility function for brand j.

$$v_j = 0.317 \times asc2_j - 10(X_{lj} - 0.5)^2 + 30(X_{pj} - 0.75)^2 + 0.567 \times Z_{fj} + 0.700 \times Z_{dj} \tag{8}$$

$asc2_j$ is a brand dummy for brand 2 such that $asc2_1 = 0$ and $asc2_2 = 1$. X_{lj} and X_{pj} were generated randomly from uniform distributions of [0,1] and [0.5, 1] respectively. The values for Z_{fj} and Z_{dj} were taken from those of actual promotional indicator variables, feature and display, in scanner panel data. The magnitude of the coefficients was chosen to be comparable to that of real choice data.

Non-IIA Data

The choice data that are not restricted by IIA were generated according to a logistic regression of equation (3) for brand 1, where

$$\begin{aligned} \eta = {} & 0.317 - 10(X_{l1} - 0.5)^2 + 0(X_{l2} - 0.5)^2 + 30(X_{p1} - 0.75)^2 \\ & + 0(X_{p2} - 0.75)^2 + 0.567 \times Z_{f1} + 0 \times Z_{f2} + 0.700 \times Z_{d1} + 0 \times Z_{d2} \end{aligned} . \tag{9}$$

In this data set, η depends on the attribute values of only brand 1 (X_{l1}, X_{p1}, Z_{f1} and Z_{d1}) but not those of brand 2 (X_{l2}, X_{p2}, Z_{f2} and Z_{d2}). In other words, the choice probability of brand 1 is unaffected by the change in the values of brand 2's attributes. It is expected that the nonparametric MNL model that implicitly assumes the IIA competitive structure would have difficulty recovering the quadratic response of X_{l1} and X_{p1}, whereas the nonparametric logistic regression should be able to recover the quadratic response from brand 1 and a flat response from brand 2 through separate nonparametric functions.

3.3 Results

Let us discuss the results for the IIA data first. Figure 1 shows the estimation results of a nonparametric MNL model. According to equation (5), the two nonparametric functions, one for loyalty and the other for price, are estimated here. The model correctly recovered quadratic shapes with a minimum at 0.5 and a maximum at 0.75 for the first and second covariate, respectively, as expected.

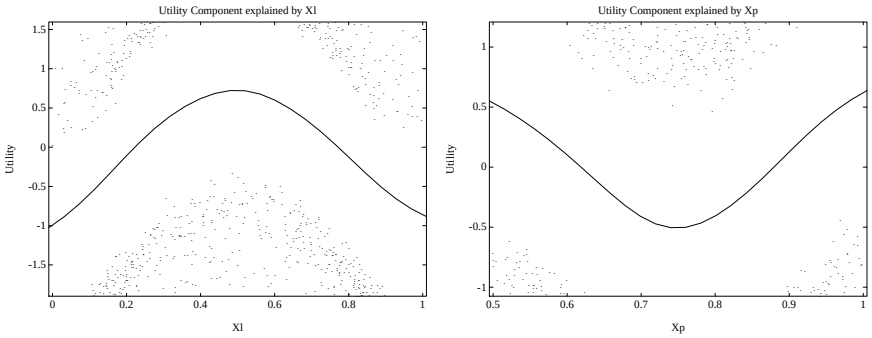

Figure 1: Result of the Nonparametric MNL for IIA Data [2]

Estimation results from the nonparametric regression are shown in Figure 2.

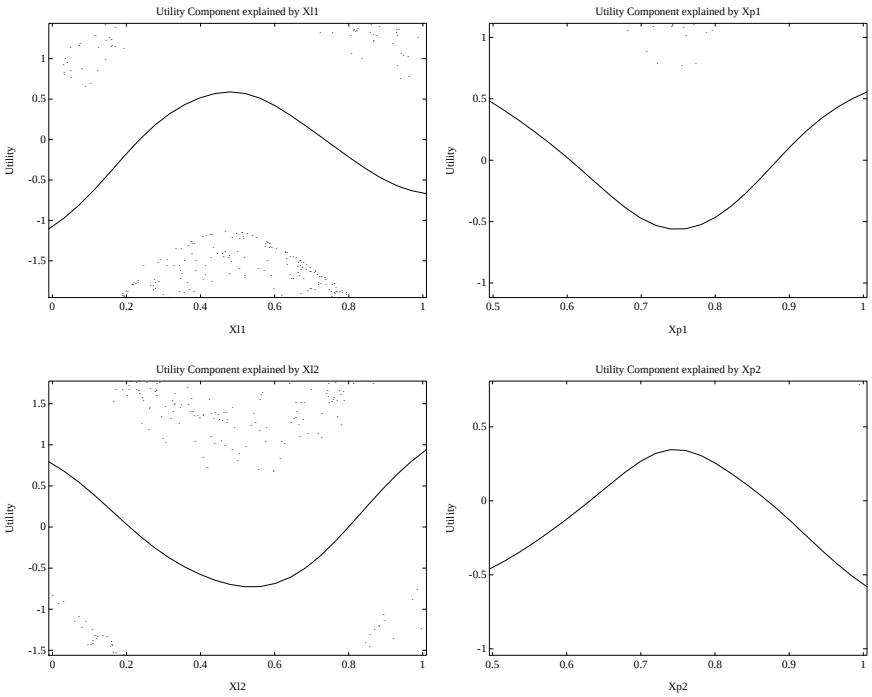

Figure 2: Result of the Nonparametric Logistic Regression for IIA Data[3]

[2]Nonparametric MNL recovers the maximum at 0.5 and minimum at 0.75 for the quadratic functions of X_l and X_p, respectively, in v_j quite well. Each small dot on the graph corresponds to a single choice occasion.

[3]In nonparametric logistic regression, because $\eta = v_1 - v_2$, the effect of X_{l1} and X_{p1} and that of X_{l2} and X_{p2} have opposite signs. The maximum at 0.5 and minimum at 0.75 for X_{l1} and X_{p1} and the minimum at 0.5 and maximum at 0.75 for X_{l2} and X_{p2} in η are recovered quite well.

As explained in equation (3), the model now contains four continuous variables, two for each alternative as X_{l1}, X_{l2},X_{p1}, X_{p2}. Since the data generating process (i.e., IIA) of equation (5) can be rewritten as equation (3) with $\eta = v_1 - v_2$ by substituting equation (8) for $j = 1$ and 2, the correctly recovered response for brand 2's covariates should be the opposite (i.e., mirror image about the x–axis) of that for brand 1's covariates. As can be seen from Figure 2, this is what we obtained.

Let us now turn to discuss the result for the non–IIA as in (9). The estimation results of a nonparametric MNL model are shown in Figure 3. Since this model is not consistent with the data assumption, the estimated nonparametric functions are extremely poor.

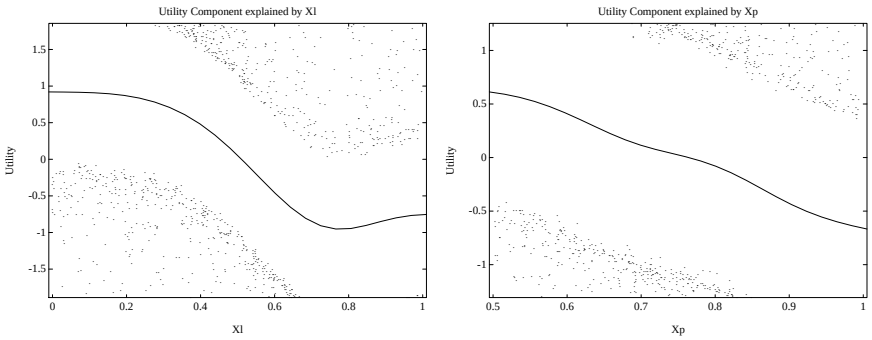

Figure 3: Result of the Nonparametric MNL Model for non-IIA Data[4]

Figure 4 shows the estimation result of a nonparametric regression model, which recovers the underlying nonparametric response of the four covariates quite well. Note the small scale of the y–axis for the two covariates of brand 2, X_{l2} and X_{p2}, suggesting that the effect from covariates for brand 2 (i.e., cross effect) is quite small. This was indeed the underlying data assumption of equation (9).

Our visual finding was confirmed by model's fit to the data, which is shown in Table 3. Each cell contains three fit statistics, loglikelihood, mean probability of correct choices, and hit rate, for the calibration and holdout samples based on the average of 10 simulation runs. Note the lower–left cell in which the nonparametric MNL model poorly fits to non–IIA data. This is because MNL assumes IIA.

[4]Because nonparametric MNL presumes IIA, the estimation for the quadratic functions of X_l and X_p for v_j is quite poor.

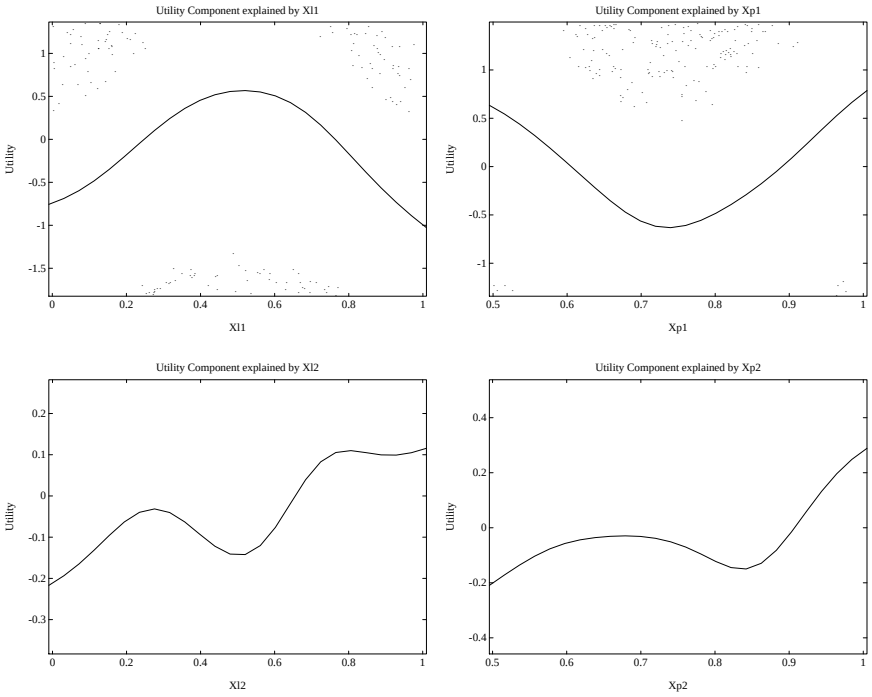

Figure 4: Result of the Nonparametric Logistic Regression for non-IIA Data[5]

			Model[a]	
			MNL	Logistic Regression
Data	IIA	LL	-530.0 (-538.7)[b]	-522.1 (-525.4)
		mean prob.[c]	0.646 (0.622)	0.652 (0.635)
		hit rate[d]	74.2% (72.6%)	75.2% (73.4%)
	Non-IIA	LL	-618.3 (-634.2)	-560.8 (-562.3)
		mean prob.	0.571 (0.549)	0.620 (0.611)
		hit rate	66.1%(64.2%)	70.9% (70.4%)

[a] The nonparametric MNL model could recover the underlying nonlinear response correctly only when the data follows the IIA restriction. In contrast, the nonparametric logistic regression was flexible enough to recover arbitrary competitive structure in data, whether they exhibit IIA or not.

[b] Figures in parentheses are for holdout.

[c] Average of the predicted probabilities of a chosen brand for all choice occasion.

[d] Fraction of choices predicted correctly when the predicted brand is defined as the brand with the highest predicted probability among alternatives. Perfect prediction results in 1.0 for the mean probability of correct choices and 100% for the hit rate.

Table 3: Fit to Calibration and Holdout Data in Simulation Study

[5] Nonparametric logistic regression recovers the maximum at 0.5 and minimum at 0.75 for X_{l1} and X_{p1} and the flat response for X_{l2} and X_{p2} in η quite well.

To summarize, the nonparametric MNL model could recover the underlying nonlinear response correctly only when the data follows the IIA restriction. In contrast, the nonparametric logistic regression was flexible enough to recover arbitrary competitive structure in data, whether they exhibit IIA or not.

4 Application to Panel Data of Consumer Brand Choice

From the simulation study, we learned that the nonparametric logistic regression can fit both IIA and non–IIA data. We also found that the fit of nonparametric MNL to non–IIA data could be quite poor. Is this the case for real data of a typical size? Are there other issues that did not surface in the simulation study? To answer these questions, we now apply the two nonparametric models, MNL and logistic regression, to the actual scanner panel data of brand choice.

The data were provided by the GfK Instrument BehaviorScan of Germany. They contained panel purchase records at one store of a healthcare product category over a period of 104 weeks. Also included were price and binary promotion indicator variables, feature and display, for each brand. We created a subset of the data by extracting purchases of panelists who had bought only three leading brands. This has resulted in a database with 2651 purchases made by 964 households.

We used two continuous explanatory variables, *PRICE* and *LOYALTY*, and one binary explanatory variable, *PROMOTION*, for our models. *LOYALTY*, whose definition was adopted from Guadagni and Little (1983), captured household heterogeneity through purchase history.[6] To minimize unwanted effect arising from heterogeneity, it is important to address the differences across households. Existing approaches to heterogeneity in marketing include (1) incorporating heterogeneity covariates, (2) latent class, (3) random coefficient model with hierarchical Bayes structure. The latter two apply to parametric models and require many degrees of freedom, which is not suitable for nonparametric models like ours. Thus, though not perfect, we adopted the first approach by including the household–specific loyalty variable. *PROMOTION* was defined to be 1 if both feature and display occurred simultaneously and 0 otherwise. This was done due to high correlation between these two promotional activities.

Let us first describe the estimation result of the nonparametric logistic regression for brand 1 choice shown in Figure 5. Because *LOYALTY* variables sum up to one across the brands, *LOYALTY* of only the first two brands are

[6]The loyalty variable must be initialized (Guadagni and Little 1983). This was done using the sample prior to the dataset we used for model calibration.

included to avoid the multicollinearity problem. The degrees of freedom is 3.9 for all explanatory variables.[7]

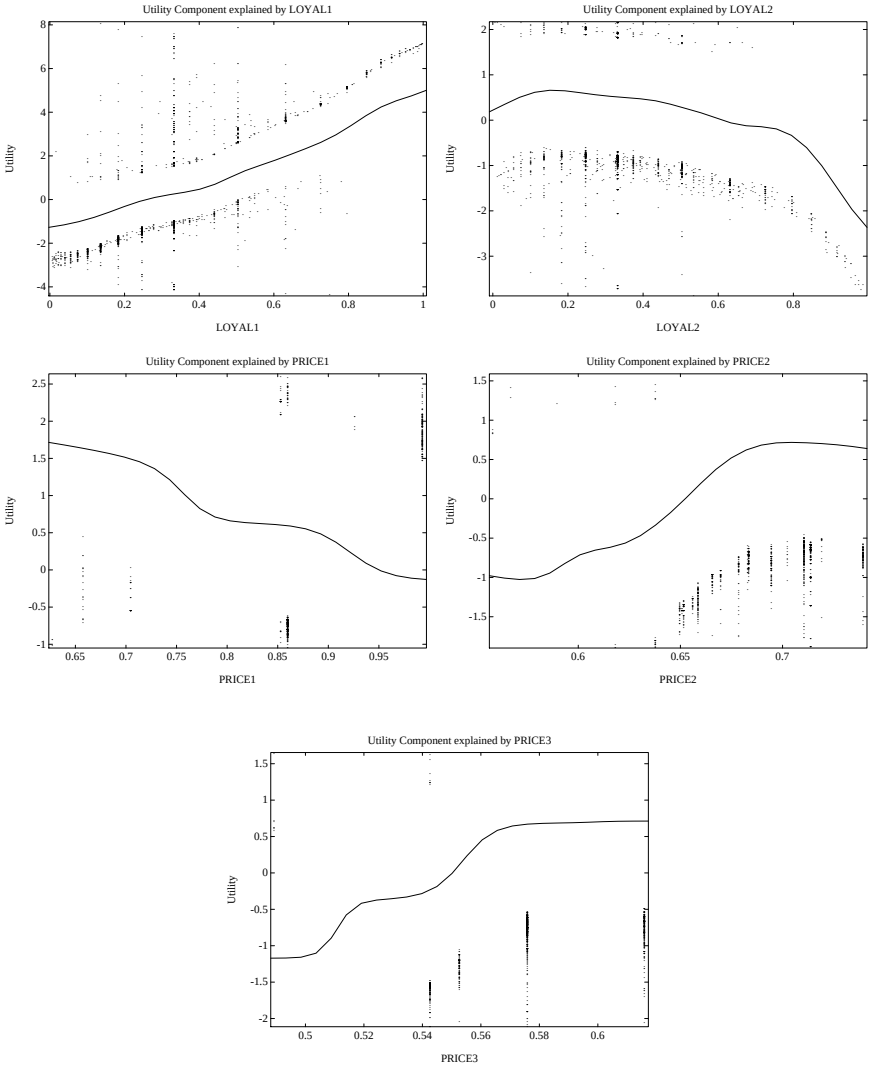

Figure 5: Result of the Nonparametric Logistic Regression for Scanner Panel Data[8]

[7]Regarding calculation of degrees of freedom for nonparametric models, please see Hastie and Tibshirani 1990, ch.4.

[8]η for brand 1 choice increases with *LOYALTY1* and decreases but not monotonically with *LOYALTY2*. η for brand 1 choice decreases with *PRICE1* and increases with *PRICE2* and *PRICE3*, which is intuitive. However, the data points in figures for *PRICE2* and *PRICE3* are rather sparse to allow their accurate nonparametric estimation.

Predictor index η for brand 1 choice increases almost linearly with $LOYALTY$ of brand 1. However, η does not decrease monotonically with $LOYALTY$ of brand 2, which is somewhat counter–intuitive. As for the nonparametric estimates of price, sparseness of the observed price levels makes the interpretation difficult, especially for brands 1 and 3. The predictor index for brand 1 choice seems to be monotonically decreasing with brand 1's price, but not monotonically increasing with the prices of competitive brands 2 and 3. In sum, it appears that the amount of data is insufficient to allow for reliable estimation.

We now turn to the result for the nonparametric model of the MNL formulation. To be comparable to the logistic regression model, the degrees of freedom value is chosen to be the same 3.9 for both functions. As shown in Figure 6, utility increases with $LOYALTY$ in a slightly nonlinear fashion and decreases linearly with *PRICE*.[9] Datapoints for the PRICE variable appear sufficiently dense to provide reliable nonparametric estimation.

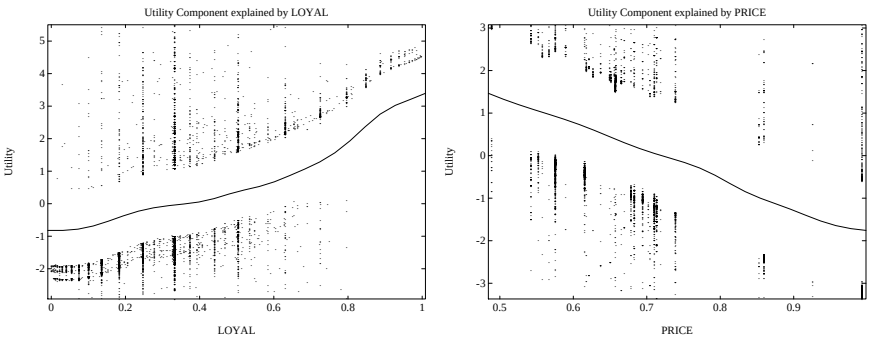

Figure 6: Result of the Nonparametric MNL Model for Scanner Panel Data[10]

The comparison of results from the two models shows that the MNL model produces robust estimates that have face validity, whereas the regression model fails to unveil the competitive structure from the data. This was indeed confirmed by the fit statistics shown in Table 4. Contrary to our result from the simulation study, in which the nonparametric logistic regression fitted the data well regardless of following IIA and the nonparametric MNL had a bad fit to non–IIA data, for real data here, the nonparametric logistic regression performed poorly.

[9] Support for the near linearity in covariates is obtained by estimating a parametric counterpart, a standard linear–in–parameters MNL model. The loglikelihood value decreases by a small amount from -1910.98 for the semiparametric specification to -1917.38 for the linear parametric one. The linearity in covariates cannot be rejected by the likelihood ratio test ($\chi^2 = 12.8$ for 5.8 degrees of freedom).

[10] Utility increases with $LOYALTY$ in a slightly nonlinear fashion and decreased linearly with *PRICE*. The observation points seem dense enough to warrant a sufficiently accurate nonparametric estimation.

Model (nonparametric)[a]	
Logistic Regression	MNL[b]
-3333.02	-1910.98
0.446	0.585
54.2%	70.4%

[a] Figures in each cell are loglikelihood, mean probability of correct choices, and hit rate.
[b] Nonparametric MNL had a much superior fit than nonparametric logistic regression in actual scanner panel data.

Table 4: Model Fit to Actual Scanner Panel Data

One reason for the poor estimation by the logistic regression formulation can be attributed to the curse of dimensionality. To infer competitive structure from data, 15 nonparametric additive functions — five functions (two for *LOYALTY* and three for *PRICE*) for each of the three regressions — must be estimated from 8892 (2964×3) choices. From the same amount of data, only two nonparametric functions are estimated in the MNL formulation. Difference in the amount of data in constructing these nonparametric functions can be seen clearly from Figures 5 and 6 as a difference in the densities of observation points.

Furthermore, the estimation result of the usual linear–in–parameters logistic regression for brand 1 choice, shown in Table 5, indicates that the magnitudes of the cross–effect are similar for brands 2 and 3. This implies that it is not necessary to capture the competitive structure through separate covariates for brands 2 and 3, as is the case for the nonparametric logistic regression.

Variable	β^a	***t*-value**
LOYALTY1	7.38	15.3
LOYALTY2	-0.59	-1.1
PRICE1	-6.99	-13.3
PRICE2	2.67	2.1
PRICE3	2.50	1.8
PROMOTION1	0.65	4.9
PROMOTION2	-0.64	-4.5
PROMOTION3	-0.62	-4.5

[a] Magnitudes of the cross–effect on brand 1 are similar, confirming that the data follow the IIA restriction reasonably well.

Table 5: Estimate for Linear–in–Parameters Logistic Regression of Brand 1 Choice

For these two reasons, at least for this dataset, the IIA assumption seems to be reasonable to impose on a nonparametric model, providing a more robust estimate.

5 Conclusions

For studying brand choice in marketing, use of nonparametric methods, which posit fewer assumptions and greater model flexibility than parametric methods, is an appealing alternative. It was found, however, that the data requirement for a fully nonparametric brand choice model is so great that obtaining such large data in marketing may not be practical (Abe 1995). By imposing an appropriate structure on components that are not sensitive to such restriction while leaving the essential component nonparametric, one can make best use of nonparametric modeling. Previous studies in brand choice indicated that, even if a parametric distributional assumption is imposed on the random component (noise/uncertainty), much of the benefit of a fully nonparametric method can be realized as long as the response function of covariates is kept nonparametric.

In this paper, we compared two such nonparametric models that were both based on GAM but differ in the degree of nonparametrization. One is a standard logistic regression of GAM, in which a choice of each brand is modeled separately in a binary fashion. The other is a MNL formulation with a nonlinear utility function, which is derived by extending the GAM framework. Both models assume the same parametric distribution for the random component but capture the response of covariates nonparametrically. The competitive structure of the logistic regression formulation is specified by data through nonparametric response functions of the attributes for competitive brands, whereas that of the MNL formulation is guided by choice theory with an i.i.d. error term. Hence, the former model can be considered to be more nonparametric and data–driven than the latter model.

The simulation study and application to actual scanner panel data of consumer brand choice provided useful insights. Because the logistic regression formulation involves fewer assumptions than the MNL formulation, the former model shares similar advantages and limitations of a fully nonparametric method. In other words, it is more flexible in modeling competitive structure, but also more prone to the curse of dimensionality problem. The regression formulation estimated more one–dimensional nonparametric functions than the MNL formulation did from the same amount of data — four times more for the simulated data and 7.5 times more for the real scanner data.[11] In general, it must estimate J^2 times more functions to capture the effect of inter–brand competition, where J is the number of brands.[12] Even for a

[11] In the simulation, the logistic regression has four nonparametric functions corresponding to X_{l1}, X_{l2}, X_{p1}, X_{p2} for each brand choice, whereas MNL had X_l and X_p for both brands. In the actual data, the numbers were 15 for logistic regression versus 2 for MNL.

[12] If MNL has K nonparametric functions, logistic regression has $J \times K$ functions for each of the J brands, thereby the factor of J^2. In the actual data of $J = 3$, the factor was 7.5 rather than 9 because only $J - 1$ instead of J loyalty variables existed due to dependency across brands.

modest value of J, the number of nonparametric functions to be estimated in the regression formulation can be quite large, thereby posing the curse of dimensionality problem.

Insufficiency of data in the logistic regression model was evidenced by the sparse and counter–intuitive shape of the response estimates for the actual scanner data. The problem was aggravated by the fact that in real data of even a moderate size (2651 purchases), brand prices tended to occur at a few discrete levels. For instance, only seven and five levels existed for brands 1 and 3, respectively. Another limitation of the logistic formulation is a logical inconsistency in which choice probabilities across brands did not result in a sum of 1. We overcome the problem by normalizing the probabilities so that they result in a sum of 1, it may be one reason for the poor fit characterized by the loglikelihood value. Future research must address the issues of data requirement and logical inconsistency and provide more applications to real data.

The other nonparametric model, which is based on the MNL formulation, produced intuitive and stable estimates. Its competitive structure presumes IIA, resulting in estimation of a fewer response functions and producing more robust results. Abe (1998, 1999) applied the model successfully to American scanner panel data in four product categories. All of these results seem to justify the IIA presumption, which can be accommodated by this nonparametric MNL model to reduce the curse of dimensionality problem.

The computation times for both nonparametric models are comparable and within a practical range. In our study, they were under one minute on a desktop computer. One advantage of the logistic regression formulation is that the popular software for GAM, called S–Plus (Venables and Ripley 1994), can be adopted without modification. At the moment, no commercial software is available to estimate the nonparametric MNL model. However, the code is fairly simple to write and was implemented in MATLAB.

We compared two nonparametric choice models in this paper. Yet, there exists a continuum of nonparametric models from a parsimonious one to a fully nonparametric model that assumes almost no structure. Our study using a typical academic scanner database suggested that if alternative brands are carefully chosen, IIA is a fairly safe assumption to impose upon. Nonparametric relaxation to capture cross effect seemed to result in the curse of dimensionality and may not be a fruitful direction to pursue unless the database size becomes substantially larger than the one currently used.

One interesting future direction is to compare non–statistical nonparametric modeling such as artificial neural network and data mining techniques. There is a striking similarity between the logistic regression of GAM and a neural network with a hidden layer and a logistic sigmoid function (West, Brockett and Golden 1997, Hruschka, Probst and Fettes 2001). Another

direction is to use a model that relaxes the additivity–in–covariates of the nonparametric logistic regression to accommodate the interaction effect. A new method, called marginal integration, can estimate a marginal influence of each explanatory variable under the presence of such interaction (Nielsen and Linton 1998). A flexible software for this approach is now available in the library of XploRe (Härdle et al. 2000)

References

[1] Abe, M. (1995), 'A nonparametric density estimation method for brand choice using scanner data', *Marketing Science* **14**(3), 300–325.

[2] Abe, M. (1998), 'Measuring consumer, nonlinear brand choice response to price', *Journal of Retailing* **74**(4), 541–568.

[3] Abe, M. (1999), 'A generalized additive model for discrete–choice data', *Journal of Business & Economic Statistics* **17**(3), 271–284.

[4] Ben–Akiva, M., and Lerman, S. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.

[5] Boztuğ, Y. and Hildebrandt, L. (2001), 'Nichtparametrische Methoden zur Schätzung von Responsefunktionen', *in* H. Hippner, U. Küsters, M. Meyer and K. Wilde, eds, 'Handbuch Data Mining im Marketing', Vieweg, pp. 241–251.

[6] Briesch, R. A., Chintagunta, P. K. and Matzkin, R. L. (1997), Nonparametric and semiparametric models of brand choice behavior, Technical report, University of Texas at Austin.

[7] Gonul, F. and Srinivasan, K. (1993), 'Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models: Methodological and Managerial Issues', *Marketing Science*, **12**(3), 213–229.

[8] Guadagni, P. M. and Little, J. D. C. (1983), 'A logit model of brand choice calibrated on scanner data', *Marketing Science* **2**(3), 203–238.

[9] Härdle, W., Klinke, S. and Müller, M. (2000), *XploRe Learning Guide*, Springer–Verlag, Berlin.

[10] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman & Hall, London.

[11] Hastie, T. J. and Tibshirani, R. J. (1986), 'Generalized additive models', *Statistical Science* **1**(3), 297–318.

[12] Hastie, T. J. and Tibshirani, R. J. (1987), 'Generalized additive models: Some applications', *Journal of the American Statistical Association* **82**(398), 371–386.

[13] Hruschka, H., Probst, M. and Fettes, W. (2001), Homogeneous and Latent Class Versions of the Neural Net–Multinomial Logit Model (NN–MNL): A Semiparametric Approach to Analyze Brand Choice, Discussion Paper 363, Faculty of Economics, University of Regensburg.

[14] Kamakura, W. A. and Russell, G. J. (1989), 'A Probabilistic Choice Model for Market Segmentation and Elasticity Structure', *Journal of Marketing Research*, **26**(4), 379–390.

[15] Manski, C. F. and McFadden, D. (1981), Alternative estimators and sample designs for discrete choice analysis, *in* C. F. Manski and D. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', The MIT Press, pp. 2–50.

[16] McCulloch, R. E. and Rossi, P. E. (1994), 'An Exact Likelihood Analysis of the Multinomial Probit Model', *Journal of Econometrics*, **64**, 207–240.

[17] Nelder, J. A. and Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society, Series A* **135**(3), 370–384.

[18] Nielsen, J. P. and Linton, O. B. (1998), 'An optimization interpretation of integration and back–fitting estimators for separable nonparametric models', *Journal of the Royal Statistical Society, Series B* **60**(1), 217–222.

[19] Rust, R. T. (1988), 'Flexible regression', *Journal of Marketing Research* **25**, 10–24.

[20] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

[21] Venables, W. and Ripley, B. D. (1994), *Modern Applied Statistics with S–Plus*, Springer.

[22] West, P. M., Brockett, P. L. and Golden, L. L. (1997), 'A comparative analysis of neural networks and statistical methods for predicting consumer choice', *Marketing Science* **16**(4), 370–391.