Abe, Makoto; Boztuæg, Yasemin; Hildebrandt, Lutz

Working Paper

Investigation of the stochastic utility maximization process of consumer brand choice by semiparametric modeling

SFB 373 Discussion Paper, No. 2000,84

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Abe, Makoto; Boztuæg, Yasemin; Hildebrandt, Lutz (2000) : Investigation of the stochastic utility maximization process of consumer brand choice by semiparametric modeling, SFB 373 Discussion Paper, No. 2000,84, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10048100

This Version is available at:
https://hdl.handle.net/10419/62233

Investigation of the Stochastic Utility Maximization Process of Consumer Brand Choice by Semiparametric Modeling

Makoto Abe

University of Tokyo
Faculty of Economics
Hongo, Bunkyo-ku
Tokyo 113-0033 JAPAN
TEL +81-3-3812-2111
FAX +81-3-3818-7082
E-mail: abe@e.u-tokyo.ac.jp

Yasemin Boztug

Humboldt University of Berlin
Institute of Marketing
Spandauer Str. 1
10178 Berlin, GERMANY
TEL +49-30-2093-5750
FAX +49-30-2093-5675
E-mail: boztug@wiwi.hu-berlin.de

Lutz Hildebrandt

Humboldt University of Berlin
Institute of Marketing
Spandauer Str. 1
10178 Berlin, GERMANY
TEL +49-30-2093-5691
FAX +49-30-2093-5675
E-mail: hildebr@wiwi.hu-berlin.de

October 2000

ABSTRACT

The use of nonparametric methods, which posit fewer assumptions and greater model flexibility than parametric methods, could provide useful insights when studying brand choice. It was found, however, that the data requirement for a fully nonparametric brand choice model is so great that obtaining such large data sets is difficult even in marketing. Semiparametric methods balance model flexibility and data requirement by imposing some parametric structure on components that are not sensitive to such assumptions while leaving the essential component nonparametric.

In this paper, the authors compare two semiparametric brand choice models that are based on the generalized additive models (GAM). One model is specified as a nonparametric logistic regression of GAM (Hastie and Tibshirani 1986) with one equation for each brand. The other model is a multinomial logit (MNL) formulation with a nonparametric utility function, which is derived by extending the GAM framework (Abe 1999). Both models assume a 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 the competitive brands, whereas that of the MNL formulation is guided by the choice theory of stochastic utility maximization (SUM).

Simulation study and application to actual scanner panel data seem to support the behavioral assumption of SUM. In addition, if we relax the SUM assumption by letting data specify the competitive structure, a substantially larger amount of data, perhaps an order of magnitude more, would be required. Therefore, if alternative brands are chosen carefully, nonparametric relaxation to capture cross effect (i.e., nonparametrization of the MNL structure) may not be warranted unless the size of database becomes substantially larger than the one currently used.

Keywords: nonparametric method, semiparametric model, generalized additive models, brand choice, stochastic utility maximization, scanner panel data

1. INTRODUCTION

In light of the availability of large panel purchase records, nonparametric methods are viable alternatives to the traditional parametric modeling (Rust 1988, Abe 1991, 1995). The methods posit fewer assumptions, thereby reducing the effect of model misspecification. Investigating nonlinearity of response functions with a nonparametric method could provide important managerial insights into consumer brand choice behavior, which may not be obtained otherwise. Previous studies of nonparametric methods, however, have pointed out their limitations in real applications, namely large computation and data requirements. The data requirement, often called the "curse of dimensionality" (Silverman 1986), 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. Although the computational issue could be solved with the advance of technology, the data problem will persist. Introducing some prior structure into a nonparametric model can often alleviate the data problem by restricting the degrees of freedom. Such models are often called semiparametric models.

Semiparametric methods balance model flexibility and data requirement by imposing parametric structure on components that are not sensitive to such assumptions while leaving the essential component nonparametric. It is important to know what kind of structure and assumptions can be applied to which component, so that the resulting semiparametric model can provide us with useful insights while minimizing the chance of misspecification. One structure that is commonly imposed is additive separability in covariates. Because the additivity permits estimation of one-dimensional nonparametric functions of a single covariate, the curse of dimensionality does not pose much threat in such nonparametric models.

Previous study in brand choice found 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 (Abe 1999, Briesch, Chintagunta, and Matzkin 1997). These findings implicitly supported a competitive structure of a multinomial logit (MNL) model that was augmented by a nonparametric utility specification.

The objective of this paper is to investigate whether the MNL structure of inter-brand competition is a reasonable assumption for semiparametric models of brand choice. If so, it would justify that the consumer choice process follows stochastic utility maximization (SUM). For this purpose, we compare the performance of two comparable additive-in-covariates semiparametric choice models that differ in one aspect. One that infers brands' competition nonparametrically from data and the other that assumes a MNL competitive structure on the basis of SUM from choice theory.

For the former model, we chose a nonparametric logistic regression proposed by Hastie and Tibshirani (1986, 1987). It is based on the generalized additive models (GAM) --- semiparametric models that relate a response variable to an additive-in-nonparametric-covariates predictor via a parametric 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.

The latter semiparametric model imposes a competitive structure that follows SUM of consumer choice theory. SUM postulates that attractiveness of each alternative among a set of available alternatives is characterized by the utility and an alternative with the highest utility to the decision maker is chosen. Utility is stochastic and expressed as the sum of a deterministic component and a random disturbance term. We assume that the deterministic component of utility is additive in a one-dimensional nonparametric function of each covariate and the random term has an i.i.d. extreme-value distribution. Its parametric counterpart is an ubiquitous multinomial logit model with a linear-in-parameters deterministic utility function.

Our finding supports the behavioral assumption of SUM. In addition, if we relaxed the SUM assumption by letting data specify the competitive structure, a substantially larger amount of data, perhaps an order of magnitude more, would be required. Therefore, at least in brand choice modeling, nonparametrization of the MNL structure may not be warranted unless the size of database becomes substantially larger than the one typically used by academic researchers.

In Section 2, the two semiparametric 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 semiparametric models are applied to German scanner panel data of brand choice in a health care product category, followed by a discussion in Section 5.

2. MODELS

Let us describe the two semiparametric models: one that estimates the competitive structure nonparametrically and the other that assumes a well-known competitive structure.

2.1. Nonparametric Logistic Regression -- Estimating Competitive Effect

Because this model is based on GAM whose idea was originated from their parametric version, generalized linear models (GLM), let use 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 h, which is linear in parameters of explanatory variables x_p ($p=1,2,..,P$) such that $h(x) = \sum_p b_p x_p$. 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 so that the predictor index takes a form $h(x) = \sum_p f_p(x_p)$. For example, the GAM for logistic regression of a binary response variable *y* is expressed as

$$Prob(x) \equiv m(x) = E(y \mid x) = G\left(\sum_{p=1}^{P} f_p(x_p) \right) = G\big(h(x)\big) \qquad (1)$$

where f_p is a nonparametric function of the *p*-th explanatory variable x_p and *G(.)* is a logistic link function of a form:

$$G(h(x)) = \frac{1}{1+e^{-h(x)}}. \qquad (2)$$

In modeling a choice of a particular brand, covariates can include marketing mix variables of that brand as well as that 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 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 (Boztug and Hildebrandt 1998), 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. Nonparametric Multinomial Logit Model -- Imposing Competitive Structure

Among many classes of SUM models for discrete choice, our nonparametric utility specification is built on a multinomial logit (MNL) model. This is because MNL has been used extensively in studying brand choice using scanner panel data (Guadagni and Little 1983). Use of MNL models to analyze scanner data is part of everyday operation in some commercial firms.

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

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

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

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

and $f_p(.)$ is a nonparametric function of the p-th explanatory variable.

Although similar in form to equation (2) 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}:

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

Notice that the predictor $h(x)$ 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 that, to be consistent with the SUM assumption, the utility function of a brand cannot include covariates of other alternative brands (Manski and McFadden 1981). The cross effect is driven by this assumption, and hence MNL models exhibits a well-known proportional draw (IIA) competitive structure.

2.3. Comparison of the Two Models

At this point, it is worthwhile to compare the two semiparametric 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 SUM, 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

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

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 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.

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.

Table 1. Comparison of the Two Semiparametric Choice Models

	Multinomial Logit	**Logistic Regression**
Behavioral Theory assumed	Stochastic utility maximization	None.
Competitive Structure implied	Proportional Draw (IIA)	None. Specified by data by including attributes for other alternatives.
Parametric Assumptions of the Random Component	Stochastic utility has an extreme-value distribution	Logistic link function

3. SIMULATION STUDY

The purpose of this simulation is to investigate how well the nonparametric MNL and logistic regression models that does and does not assume SUM, respectively, fit to data sets that does and does not follow the SUM process. For data that follow the SUM process, we expect both models to fit the data well. MNL fits well because the model assumption is consistent with the data. Logistic regression should perform well because it does not presume a particular competitive structure, and nonparametric function h 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 SUM process in the data. For data that do not follow the SUM process, we expect that MNL -- whose competitive assumption is incompatible -- performs poorly, whereas logistic regression can still fit the data well. Please refer to Table 2.

Table 2: Data and Model in Simulation Study

		Semiparametric Model	
		MNL	Logistic Regression
Data process and Competitive structure	Stochastic Utility Maximization	°	°
	No cross effect	×	°

Simulated brand choice data for two alternatives consisting of 1000 choice incidents were generated according to two processes: one that is based on SUM and the other that is not. We used two continuous variables for alternative j (where j = 1 or 2) as X_{j1} (e.g., loyalty) and X_{j2} (e.g., price), whose nonlinear response must be estimated by the semiparametric models. To make the simulation more challenging and realistic, we also intruduced two binary indicator variables for alternative j, Z_{j1} (e.g., feature) and Z_{j2} (e.g., display).

The choice data with SUM were generated according to a multinomial logit process of equation 5 with the following utility function for brand j.

$$v_j = 0.317 \times asc2_j - 10(X_{j1} - 0.5)^2 + 30(X_{j2} - 0.75)^2 + 0.567 \times Z_{j1} + 0.700 \times Z_{j2} \qquad (8)$$

$asc2_j$ is a brand dummy for brand 2. X_{j1} and X_{j2} were generated randomly from uniform distributions of [0,1] and [0.5,1], respectively. The values of Z_{j1} and Z_{j2} 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.

The choice data that do not assume SUM were generated according to a logistic regression of equation 2 for brand 1 where

$$\begin{aligned} h = {} & 0.317 - 10(X_{11} - 0.5)^2 + 0 \times (X_{21} - 0.5)^2 + 30(X_{12} - 0.75)^2 + 0(X_{22} - 0.75)^2 \\ & + 0.567 \times Z_{11} + 0 \times Z_{21} + 0.700 \times Z_{12} + 0 \times Z_{22} \end{aligned} . \qquad (9)$$

In this data set, h depends on the attribute values of only brand 1 (X_{11}, X_{12}, Z_{11} and Z_{12}) but not of brand 2 (X_{21}, X_{22}, Z_{21} and Z_{22}). 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 MNL model that implicitly assumes the competitive effect would have difficulty recovering the quadratic response of X_{11} and X_{12}, whereas the logistic regression should be able to recover

the quadratic response from brand 1 and a flat response from brand 2 through separate nonparametric functions.

Let us first discuss the result for the data that are generated according to SUM. Figure 1 shows the estimation result of a nonparametric MNL model. As expected, the model correctly recovered the quadratic shapes of minimum at 0.5 and maximum at 0.75 for the first and second covariate, respectively.

An estimation result by the nonparametric regression is shown in Figure 2. The model now contains four continuous variables, two for each alternative as X_{11}, X_{12}, X_{21}, X_{22}. Since the data generating process (i.e., SUM) of equation (5) can be rewritten as equation (2) with $h = v_1 - v_2$, the correctly recovered response for brand 2's covariates should be opposite (i.e., mirror image about the x-axis) of that for brand 1's covariates. This was indeed the case where the minimum and maximum of brand 2's covariates are switched from those of brand 1's covariates.

Let us now turn to discuss the result for the data that do not follow the SUM process as in (9). The estimation result 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 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 y-axis for the two covariates of brand 2, suggesting that the effect from covariates for brand 2 (i.e., cross effect) is quite small. This was indeed the underlying data assumption.

To summarize, the nonparametric MNL model could recover the underlying nonlinear response correctly only when the data follows the SUM process. In contrast, the nonparametric logistic regression was flexible enough to recover arbitrary competitive structure in data, whether they exhibit SUM or not. This simulation confirmed Table 2.

4. APPLICATION TO PANEL DATA OF CONSUMER 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. *PROMOTION* was defined to be 1 if both feature and display occurred simultaneously and 0 otherwise. This was done due to the high correlation between these two promotional activities.

The objective here is to discover the shape of the price response on brand choice, controlling for other impact variables *LOYALTY* and *PROMOTION*. Figure 5 presents the estimated functions of the logistic regression for brand 1. Because *LOYALTY* variables sum up to one across the brands, *LOYALTY* of only the first two brands are included to avoid the multicollinearity problem. The degrees of freedom is about 4 for all explanatory variables. The predictor index, h, for brand 1 increases almost linearly with *LOYALTY* of brand 1. However, h 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 seems to be monotonically decreasing with price of brand 1, but not monotonically increasing with price of the competitive brands 2 and 3.

To provide support for this result, a corresponding parametric model, a linear-in-parameters logistic regression for brand 1 choice, is estimated. The result is reported in Table 2. *LOYALTY* of brand 1 has a significant positive slope, whereas the negative slope of brand 2 is not significant. *PRICE* of brands 1 and 2 has an expected sign at the 5% significance level but not that of brand 3. *PROMOTION* of each brand has a significant expected sign. Note the comparable magnitudes for *PRICE2* and *PRICE3* as well as those of *PROMOTION2* and *PROMOTION3*. This implies that the competitive effects from brand 2 and brand 3 on the choice of brand 1 are similar. The loglikelihood value increases from -3333.02 for the linear specification to -3143.56 for the semiparametric one, a rather large improvement.

Table 2. Parametric Estimation Result for Logistic Regression of Brand 1 Choice

Variable	Estimation for *b*	*t*-value
LOYALTY 1	7.38	15.3
LOYALTY 2	-0.59	-1.1
PRICE 1	-6.99	-13.3
PRICE 2	2.67	2.1
PRICE 3	2.50	1.8
PROMOTION 1	0.65	4.9
PROMOTION 2	-0.64	-4.5
PROMOTION 3	-0.62	-4.5

We now turn to the result for the semiparametric model of the MNL formulation. As shown in Figure 6, utility increases with *LOYALTY* in a slightly nonlinear fashion and decreased linearly with *PRICE*. To be comparable to the logistic regression model, the degrees of freedom is chosen to be the same 3.9 for both functions. 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. In addition, all explanatory variables are highly significant in the expected direction, as it can be seen in Table 3.

Table 3. Parametric Estimation Result for MNL Model

Variable	Estimation	*t*-value
LOYALTY	3.96	24.2
PRICE	-7.18	-16.8
PROMOTION	1.10	16.3
Brand 2	-1.36	-13.2
Brand 3	-2.27	-15.6

The comparison of the results by the two models shows that the MNL model produces the robust estimate that have face validity, whereas the regression model fails to unveil the competitive structure from the data. One reason for the poor estimation by the logistic regression formulation can be attributed to the curse of dimensionality. To infer the competitive structure from the data, 18 nonparametric additive functions --- six functions (three for *LOYALTY* and three for *PRICE*) for each of the three regressions --- must be estimated from 8892 (2964x3) choices. From the same amount of data, only two nonparametric functions are estimated in the multinomial logit formulation. Difference in the amount of data in constructing these nonparametric functions can be seen clearly from Figures 5 and 6 as difference in the densities of observation points.

Furthermore, the parametric estimate, shown in Table 2, 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 semiparametric logistic regression.

For these two reasons, at least for this dataset, the SUM 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). Semiparametric methods balance model flexibility and data requirement by imposing some parametric structure on components that are not sensitive to such assumptions while leaving the essential component nonparametric. 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 semiparametric models that were based on GAM. 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 a 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 the competitive brands, whereas that of the MNL formulation is guided by the choice theory of SUM. 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. 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. Even for a 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), price of a brand 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 sum to one. This 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 semiparametric model, which is based on the MNL formulation, produced intuitive and stable estimates. Its competitive structure is supported by SUM theory, 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 behavioral theory of SUM, which can be accommodated by this semiparametric MNL model to reduce the curse of dimensionality problem.

The computation times for both semiparametric 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 semiparametric MNL model. However, the code is fairly simple to write and was implemented in MATLAB.

We compared two semiparametric choice models in this paper. Yet, there exists a continuum of models from a parsimonious parametric model to a fully nonparametric model. Our study using a typical academic scanner database suggested that, if alternative brands are carefully chosen, SUM is a fairly safe assumption to impose. 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 size of database 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). Another direction is to use a model that relaxes the additivity-in-covariates of the semiparametric logistic regression to accommodate the interaction effect. A new method, called marginal integration, can estimate a marginal influence of each explanatory variable under presence of such interaction (Nielsen and Linton 1998). A flexible software for this approach is now available under the library of XploRe (Härdle et al. 2000)

REFERENCES

Abe, M. (1991), "A Moving Ellipsoid Method for Nonparametric Regression and its Application to Logit Diagnostics using Scanner Data," *Journal of Marketing Research*, 28, 339-346.

Abe, M. (1995) "A Nonparametric Density Estimation Method for Brand Choice using Scanner Data," *Marketing Science*, 14(3), 300-325.

Abe, M. (1998), "Measuring Consumer, Nonlinear Brand Choice Response to Price," *Journal of Retailing*, 74 (4), 541-568.

Abe, M. (1999) "A Generalized Additive Model for Discrete Choice Data," *Journal of Business and Economic Statistics*, 17 (3), 271-284.

Boztug, Y., and Hildebrandt, L. (1998). Nicht- und semiparametrische Markenwahlmodelle im Marketing. Discussion Paper, SFB 373, Humboldt University of Berlin.

Briesch, R. A., Chintagunta, P. K. and Matzkin, R. L. (1997), "Nonparametric and Semiparametric Models of Brand Choice Behavior," Working Paper, The University of Texas at Austin.

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.

Härdle, W., Klinke, S. and Müller, M. (2000), *XploRe Learning Guide*, Springer-Verlag.

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," *Statistical Science*, 1(3), 297-318.

Hastie, T. and Tibshirani, R. (1987), "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, 82(398), 371-386.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall.

Manski, C. F. and McFadden, D. (Editors) (1981), *Structural Analysis of Discrete Data with Econometric Applications*, The MIT Press.

Nelder J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Series A, 135, 1134-1143.

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.

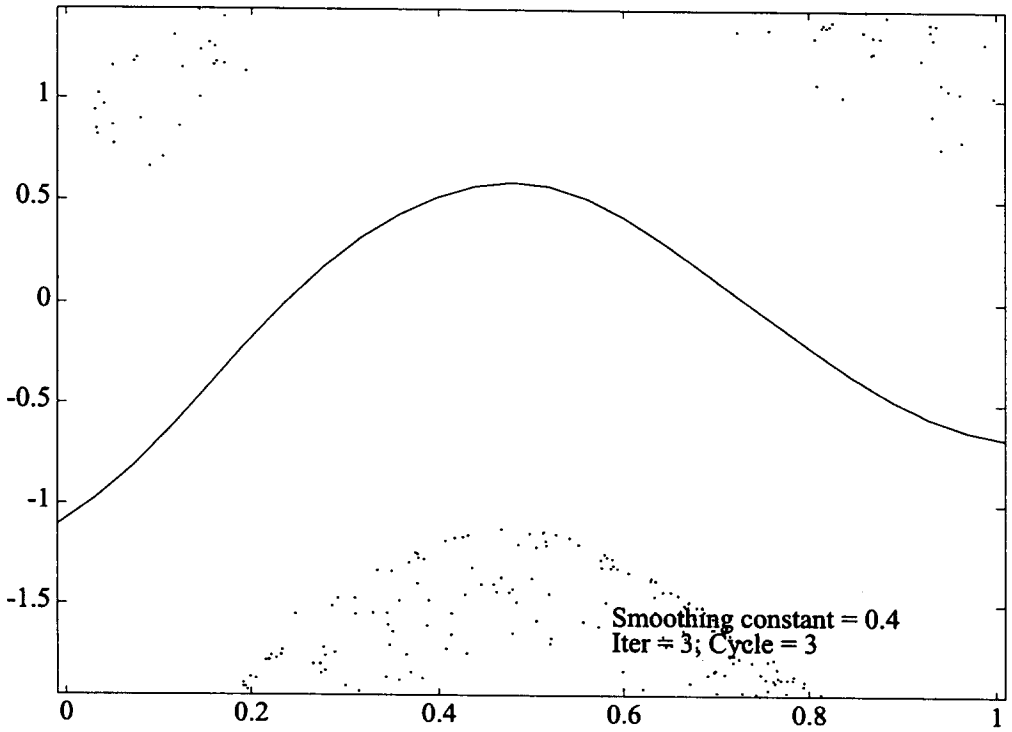
Rust, R. T. (1988), "Flexible Regression," *Journal of Marketing Research*, 25, 10-24.

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

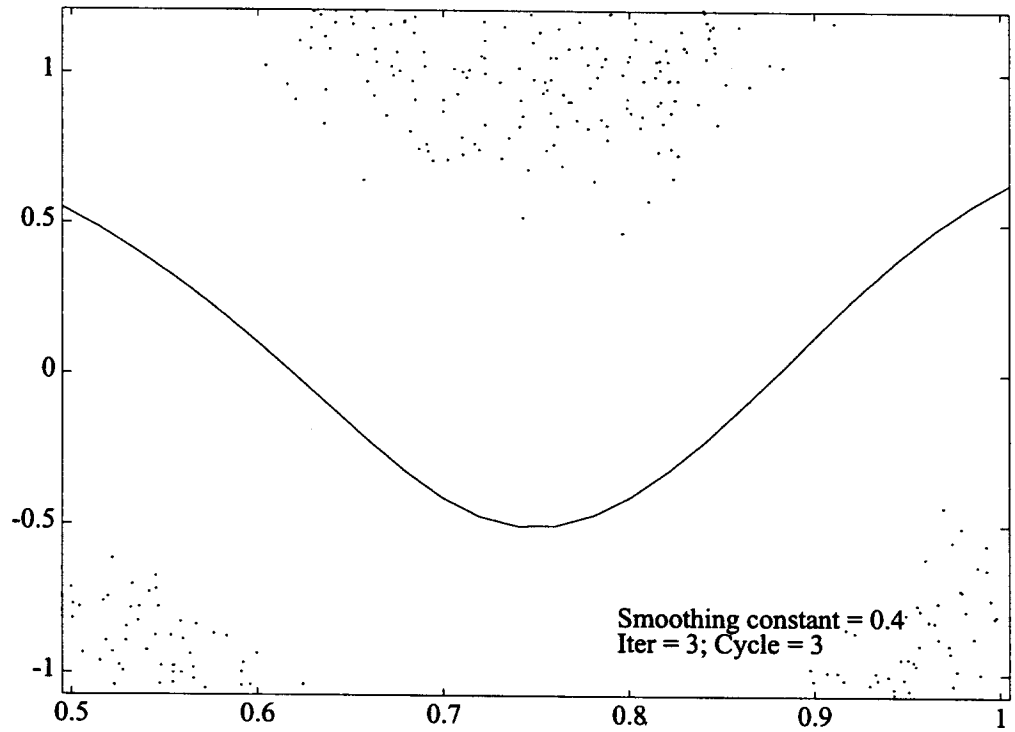
Venables, W. N. and Ripley, B. D. (1994), *Modern Applied Statistics with S-Plus*, Springer-Verlag.

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.

Figure 1. Nonparametric MNL with SUM Data.

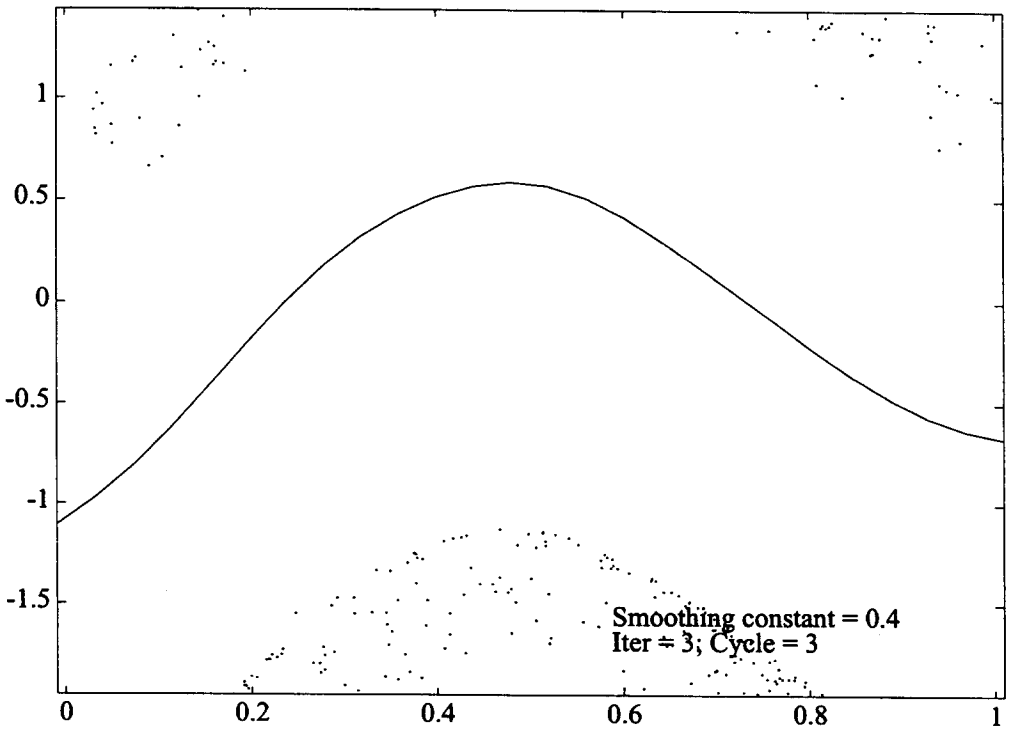


X_1

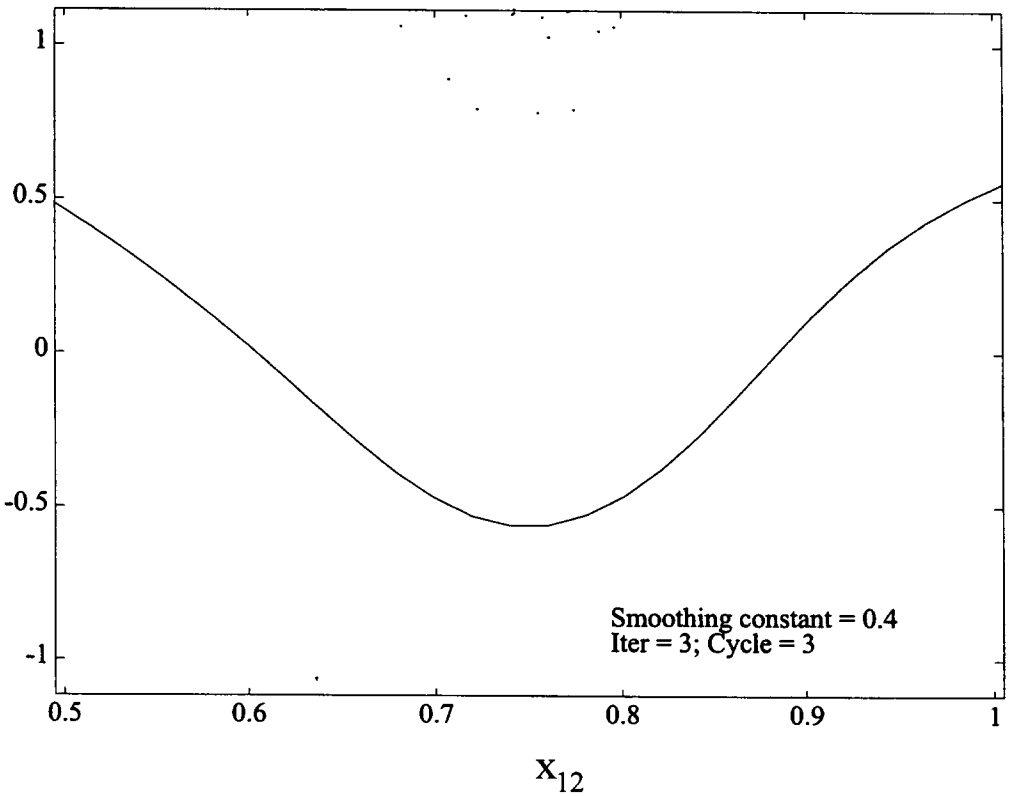


X_2

Figure2a. Nonparametric Regression for Brand 1 with SUM Data

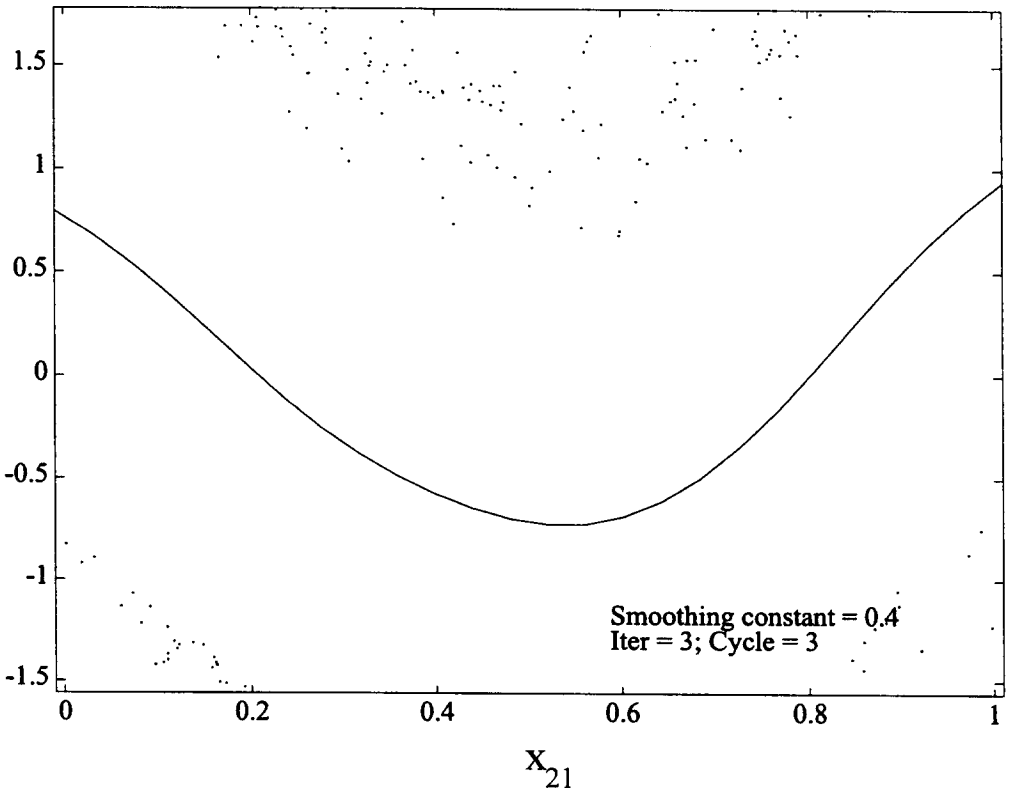


X_{11}



X_{12}

Figure2b. Nonparametric Regression for Brand 2 with SUM Data



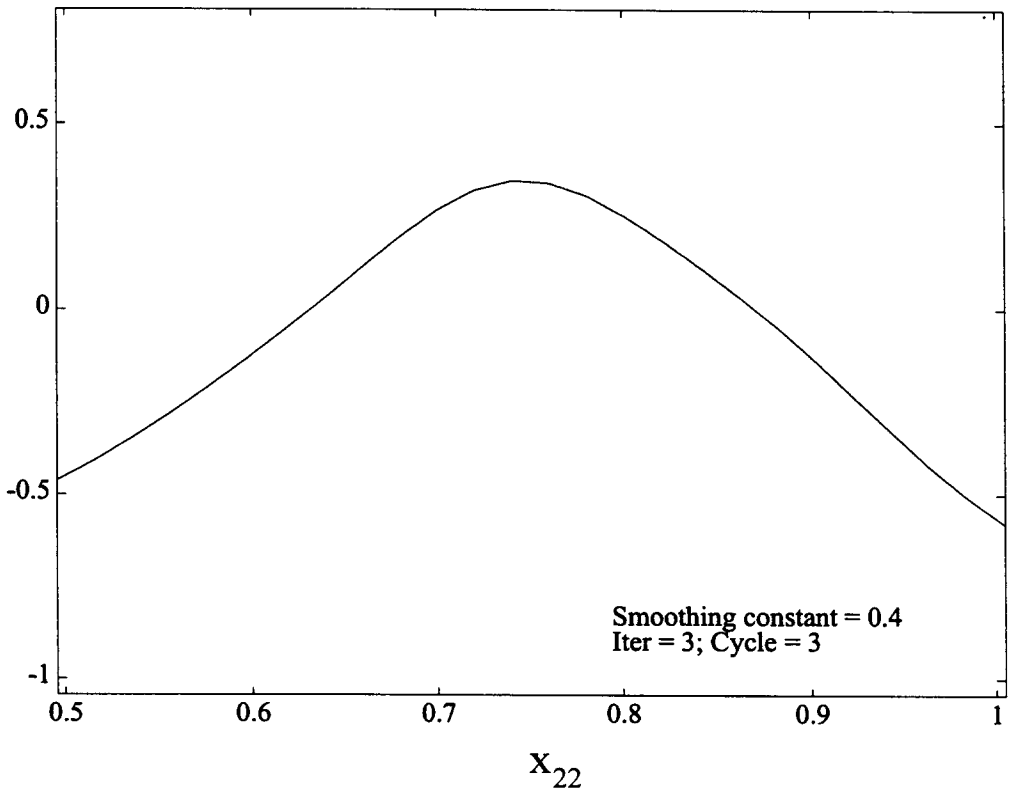
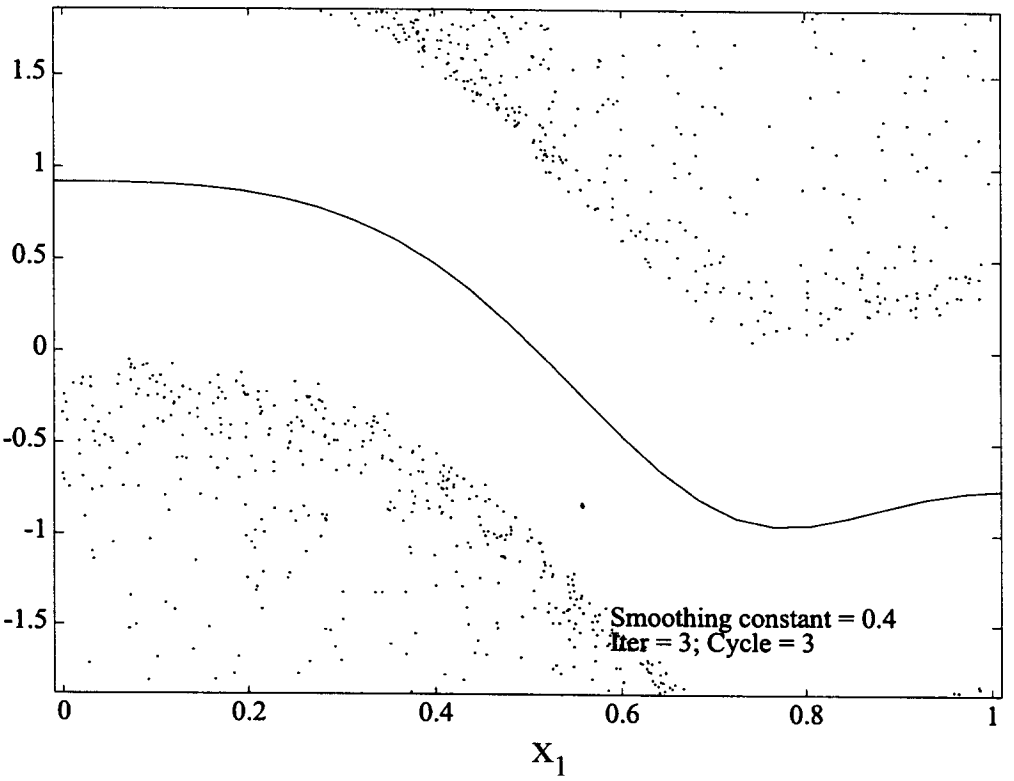


Figure3. Nonparametric MNL with Non-SUM Data



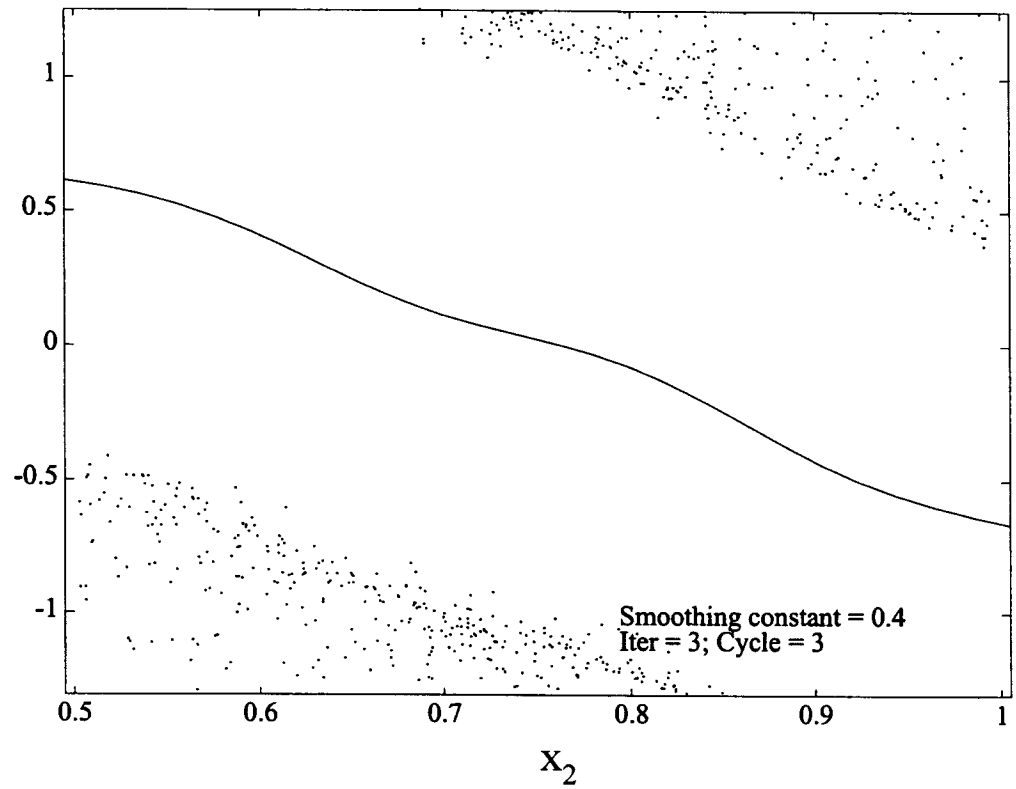
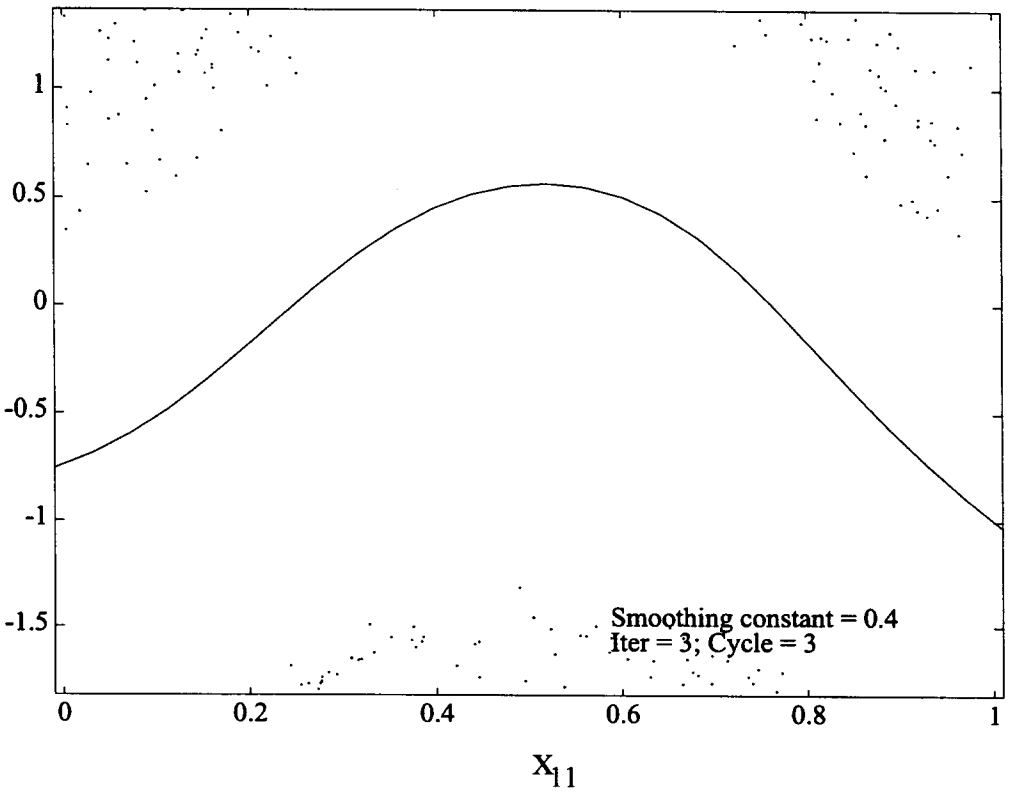


Figure4a. Nonparametric Regression for Brand 1 with Non-SUM Data



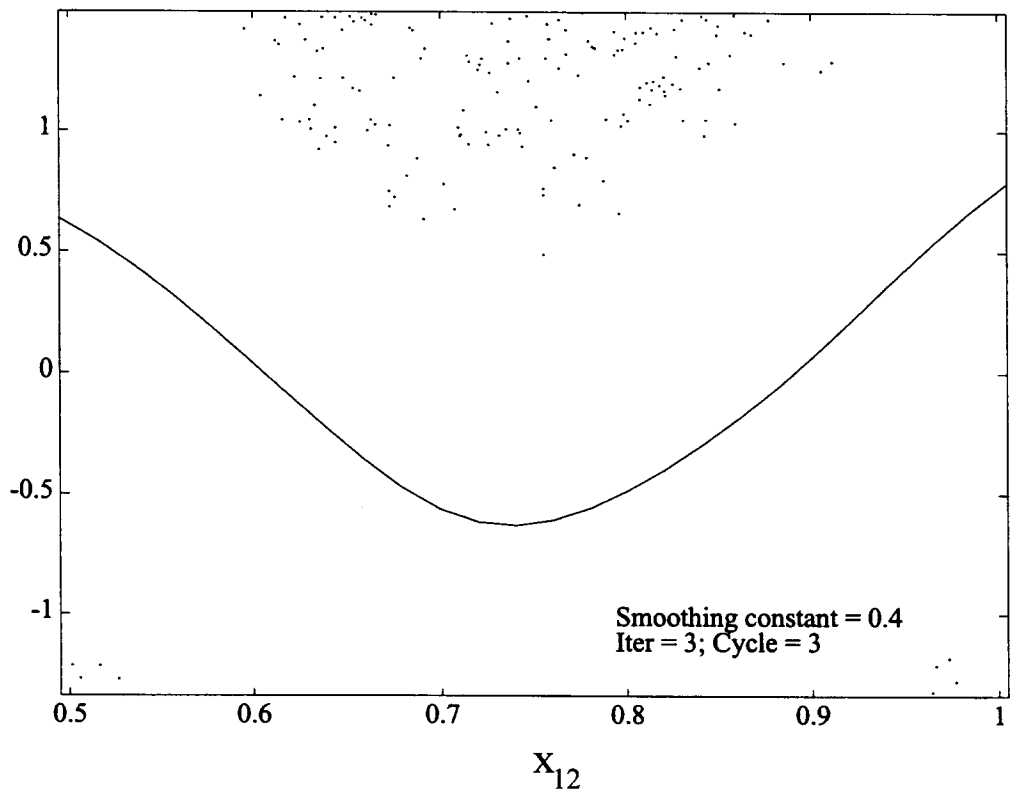
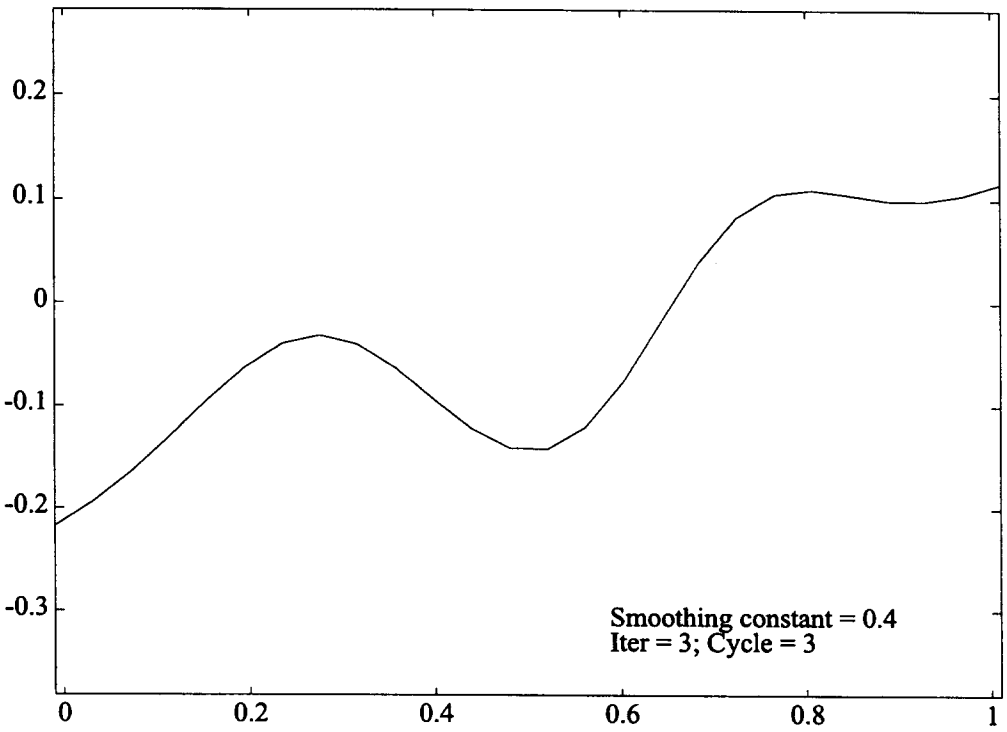
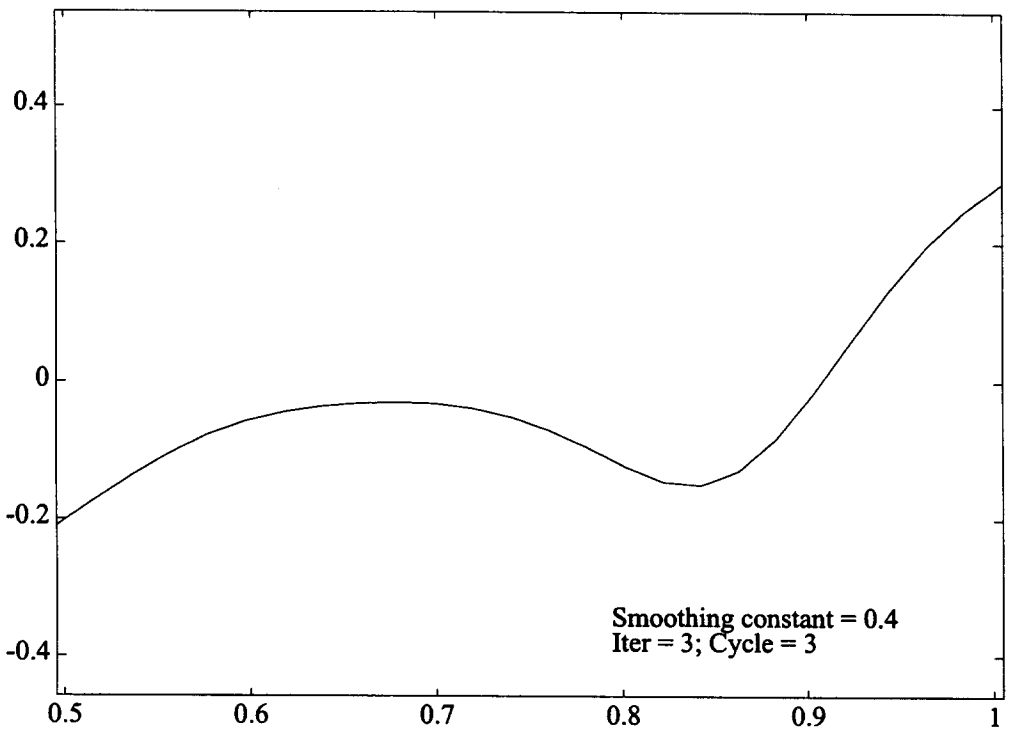


Figure4b. Nonparametric Regression for Brand 2 with Non-SUM Data

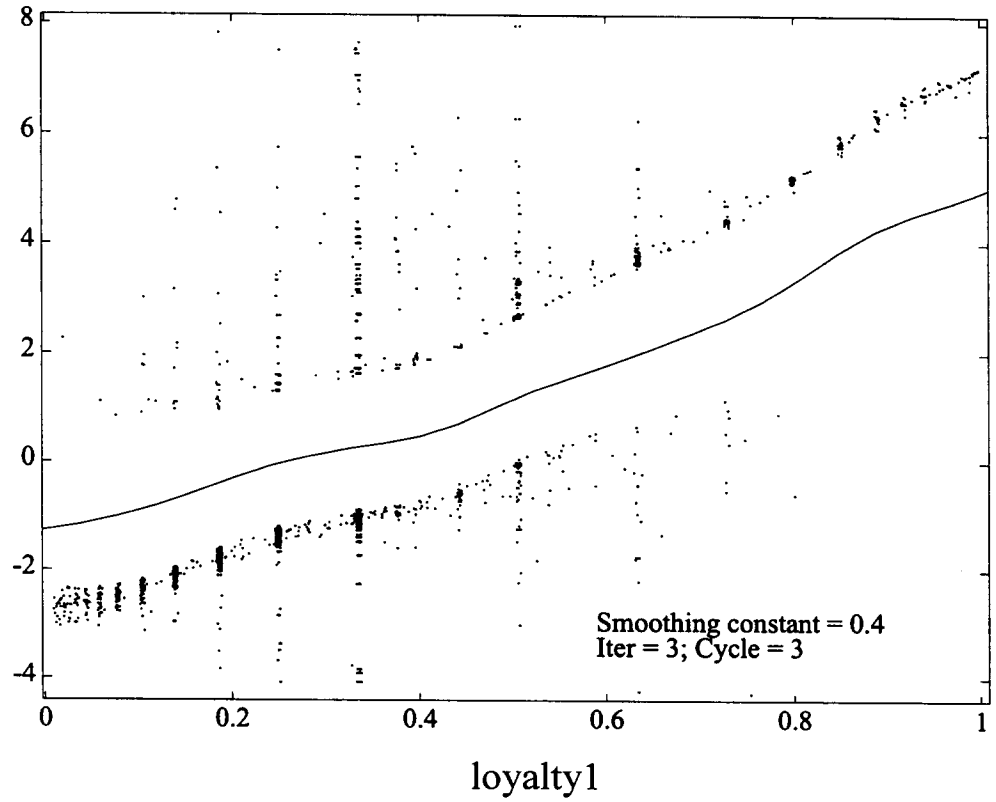


X_{21}



X_{22}

Figure5a: Nonparametric Regression for Brand 1 with Scanner Data



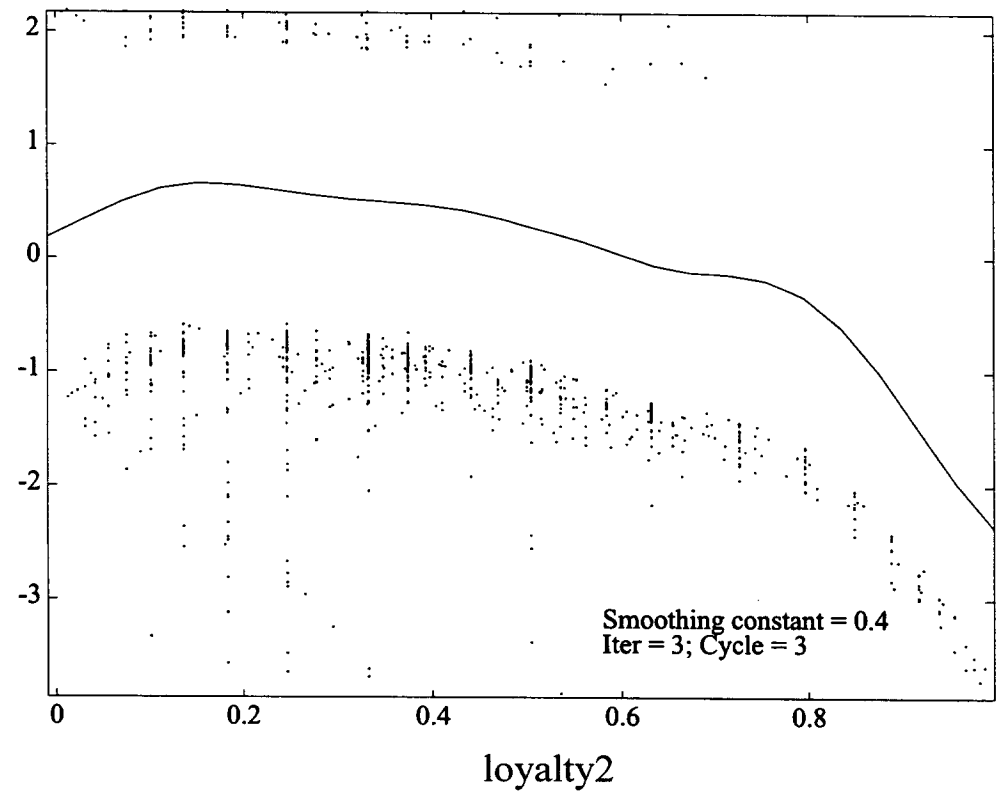
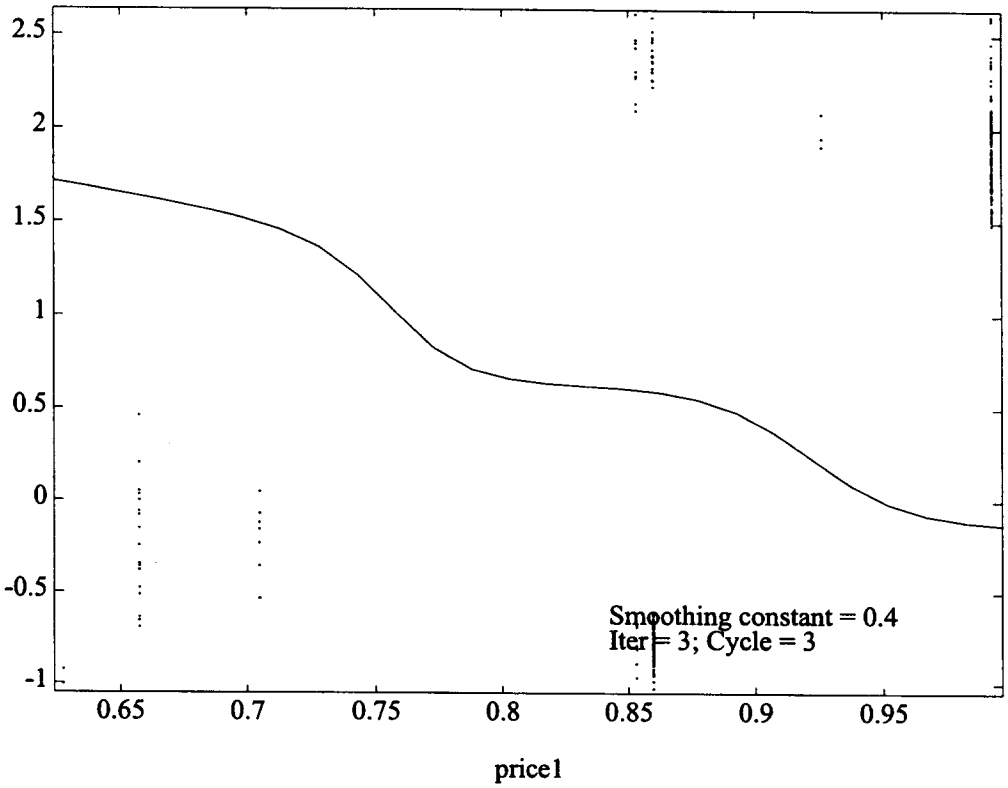


Figure5b: Nonparametric Regression for Brand 1 with Scanner Data



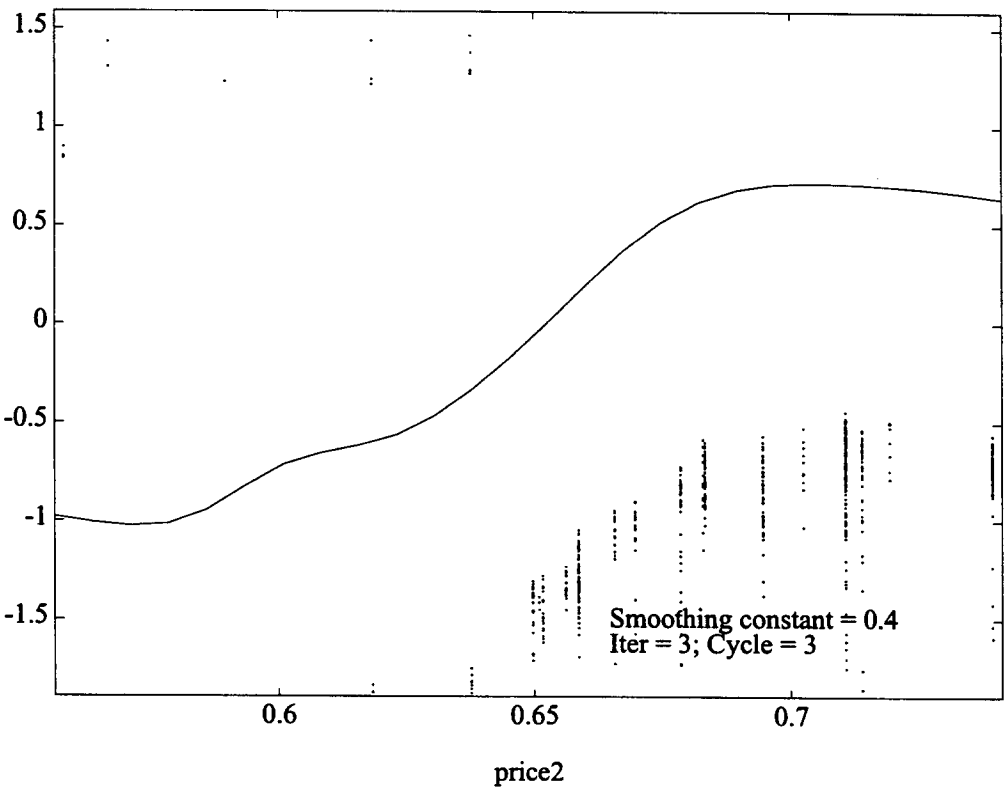


Figure5c: Nonparametric Regression for Brand 1 with Scanner Data

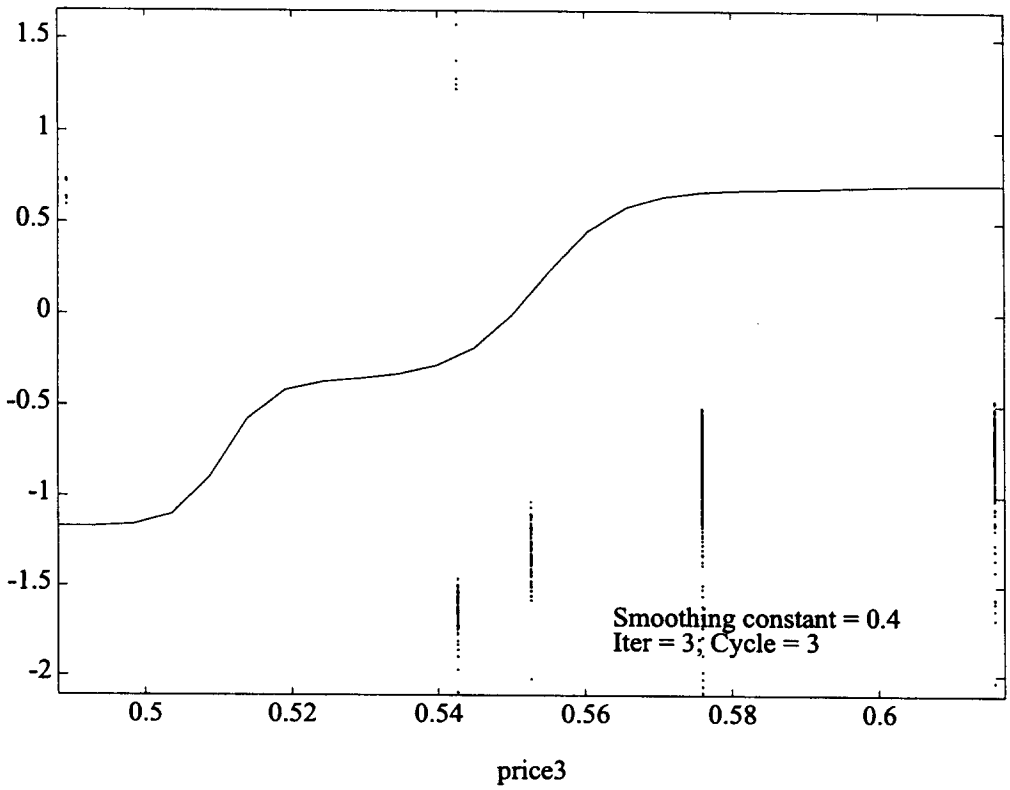


Figure 6. Nonparametric MNL with Scanner Data