ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Econometrics & Operations Research

A comparison of Thin Plate Splines and P-splines in the Generalized Additive Model for Discrete-Choice Data

Sanne Bakker - 450842

Supervisor: Gruber, K.

Second corrector: Dijkstra, N.F.S.

July 7, 2019

Abstract

The utility functions used in the Multinomial Logit model for to describe consumer brand choice data are usually specified as linear-in-parameter. Due to nonlinear price functions, it can be more realistic to implement nonparametric functions instead, which can be done by spline smoothing techniques. This research compares a knot-based approximation of Thin Plate Splines to P-splines with second order difference penalties, as the latter could be more suitable in brand choice related context. The results, however, show that one technique is not significantly better than the other. Furthermore, the influence of the number of knots on the estimation accuracy is investigated. The results show that when including more knots, the performance of the fit increases. However, after a certain point, the computation time increases significantly, while the performance does hardly improve anymore.

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

The Multinomial Logit (MNL) model is often used to describe consumer brand choice data. For example, the brand of catsup a consumer chooses in the supermarket. MNL is popular because the coefficients of the parameters, usually characteristics like price and promotions, have a clear interpretation. In general, the discrete choice probabilities are estimated in terms of utility functions, which means that the alternative with the highest utility is chosen. These functions are usually specified as linear in the parameters. However, this can be an unrealistic assumption due to the fact that individuals have a subjective perception of price. First of all, consumers are often insensitive to price around a certain reference price (Kalyanaram and Little (1994)). This is called latitude of acceptance: when deciding whether or not to buy the product, the consumer is not influenced by the price. Secondly, there can be a so-called saturation effect, meaning that consumers discount the discount when it concerns a higher price reduction (Gupta and Cooper (1992)). As a result, the function of price becomes nonlinear.

Considering this, it is appealing to make the model less restrictive by relaxing the linear-in-parameters structure of the utility function to allow for nonlinear influences. As a consequence, it becomes harder to find a suitable fit by means of parametric functions. Here, parametric means it is known exactly in advance how many parameters will be fitted. The counterpart of this method is nonparametric estimation, where the set of parameters is not fixed and do not have to be linear. Another advantage of this method is that no assumptions about the relationship of the data points is required in advance.

Nonparametric models can be estimated using smooth functions. These are often based on splines and can be seen as piece-wise linear functions, which this paper will focus on. First of all, two widely used smoothing techniques will be compared. These are Thin Plate Splines and P-splines. After that, the influence of the number of knots on the estimation accuracy is investigated. This is summarized in the following two questions: "*How do P-splines compare in performance to Thin Plate Splines?*" and "*How does the number of knots influence the estimation accuracy?*". The research is in the framework of the MNL model for a multivariate response variable.

The remainder of the paper is structured as follows. Section 2 provides an overview about the relevant literature on spline smoothing and the associated model estimation. The data set analysed in this research is described in Section 3. Consequently, the methods used to construct the splines as well as to evaluate the models are presented in Section 4. After that, the results are summarized in Section 5 and Section 6 provides the conclusion of the research.

2 Literature

This section reviews the literature on different spline smoothing techniques, as well as the context in which they are estimated.

2.1 Generalized Linear Models

Generalized Linear Models (GLM; Nelder and Wedderburn (1972)) extends the class of linear models, as it allows the error term to follow another distribution than a normal distribution. GLM is a class of likelihood based regression models based on parametric functions. The model exists of a response variable y and linear predictor η which consists of a set of covariates $x_1, ..., x_p$. In general, it has a linear-in-parameter structure, such that $\eta = \sum \beta_p x_p$. Depending on η is a link function $g(\cdot)$, which links the expected value of the outcome to the predictor function.

When the dependent variable involves a discrete choice, the model is usually expressed in terms of maximizing random utility. The random utility of alternative j $(j = 1, 2, ..., J)$ for individual i $(i = 1, 2, ..., N)$ is defined as $u_{ij} = v_{ij} + e_{ij}$, where e_{ij} is the error term. In the MNL setting, the link function equals the probability that i chooses j:

$$Pr_i(j) = g(\eta) = \frac{exp(v_{ij})}{\sum_{k \epsilon C_i} exp(v_{ik})}, \quad where \quad \eta = v_{ij} = \sum_p \beta_p x_p. \tag{1}$$

2.2 Generalized Additive Models

Hastie and Tibshirani (1986, 1987) introduced the class of Generalized Additive Models (GAM), which is an extension of GLM. In a GAM, the assumption of linearity of the predictor variables is relaxed by specifying unknown smooth functions $s(\cdot)$. Often, the assumption is made that the explanatory variables are additive separable, such that the predictor $\eta = \sum s_p(x_{ijp})$.

When the model is again expressed in utility terms, a classification can be made according to the (non)parametric specification of the two components. This is summarized in Table 1. Here, semiparametric means that the model consists of both a parametric part as well as a nonparametric part.

Table 1: Three categories on a qualitative response variable according to the parametric specification.

Type	Deterministic component: v_{ij}	Random component: e_{ij}
1. Nonparametric	Nonparametric	Distribution-free
2. Semiparametric I	Parametric (usually linear-in-parameter)	Distribution-free
3. Semiparametric II	Nonparametric	Parametric distribution

In the second class of semiparametric specification, the deterministic component v_{ij} takes a nonparametric form, while the random component e_{ij} assumes a certain parametric distribution. Hastie and Tibshirani (1986) used a logistically distributed random component in a GAM for binary response.

For modeling brand choice behavior with multinomial response, Briesch et al. (2002) compared nonparametric with semiparametric estimation while assuming an extreme-value distribution for the stochastic term. Abe (1999) followed this approach in the MNL setting, which results in the following model:

$$Pr_i(j) = g(\eta) = \frac{exp(v_{ij})}{\sum_{k \epsilon C_i} exp(v_{ik})}, \quad where \quad \eta = v_{ij} = \sum_p s_p(x_{ijp}). \tag{2}$$

Here, e_{ij} follows an iid Type I extreme-value distribution and the deterministic component contains additive one-dimensional smooth functions $s_p(\cdot)$, instead of the usual linear-in-parameter covariates.

2.3 Spline smoothing

Estimation of a nonparametric function can be done by means of spline smoothing. In general, this can be expressed as:

$$s(x) = \sum_{q=1}^{Q} \alpha_q B_q(x), \tag{3}$$

where $B_q(x)$ are some basis functions and Q is the basis dimension. To obtain intervals for the spline, knots are placed within the domain of the covariate. When there are less knots than number of observations, this is called a regression spline. In other words, the spline bases to construct a model for the original data set are found by using a much smaller data set. Therefore, it can be seen as computationally efficient. However, the main issue here is deciding about the number and the location of the knots, since this can have a remarkable influence on the fit of the model to the data.

Beside the fact that the number of knots controls the flexibility of the model (more knots means more fluctuations), the smoothness can also be controlled by penalties. This means that each basis has a related "wiggliness" penalty, where the smoothing parameter λ controls the trade-off between the roughness of the function estimate and the fitting of the data. In terms of extremes: if λ equals zero this will result in a linear curve. Conversely, if λ goes to infinity, the curve becomes very wiggly, since there will be a curve between every data point. In that case, we speak of an interpolating spline. Depending on the context either the data will be overfitted, or the estimate will still be valid but not very meaningful. As a result, it will be

hard to extrapolate a “trend” into the future.
Note that there is a trade-off between objectivity and efficiency of the model estimation. On the one hand, when there are as many knots as data points, we do not need to decide about the number of knots, which can be seen as objective. However, at the same time this means that there are n parameters to estimate in case of n data points, which can be computationally costly.

In this research we focus on two types of smoothing splines: a knot-based approximation of Thin Plate Splines and P-splines with difference based penalties. Both techniques are described as optimal in literature, but each in their own way: see for example Wood (2017) and Eilers and Marx (1996), respectively. Therefore, both techniques are widely used and valued. In other words, enough reason to compare the performance of these two techniques.

2.3.1 Thin Plate Splines

A very general technique is the Thin Plate Spline (TPS), which has mainly been investigated by Wood (2003). In this smoothing technique, no basis functions need to be selected, because they are found during the procedure itself. As a result, it can be seen as an ideal smoother in the sense that no other smooth function will better match the data. In case one has no information about the specifications of the model it can be a favorable choice, despite the problem of the n free parameters. In addition to the *full* TPS method, there exist a knot-based approximation, which is described in Section 4.1. Note in this case, one has to accept an increase in subjectivity of the fit since we have to choose the number knots.

2.3.2 P-splines

With regard to the penalized regression splines, P-splines (PS) are widely used. Eilers and Marx (2010); Eilers et al. (2015) did a lot of research into this area. The related formulas are described in Section 4.2. PS exist of B-spline bases with discrete penalties based on first or second order differences of the spline coefficients (Eilers and Marx (1996)). In general, d-th order differences can be expressed as: $D_d\alpha = \Delta^d\alpha$, where D_d a matrix and α the vector with spline coefficients. It makes use of $\Delta^d\alpha_j = \Delta(\Delta^{j-1}\alpha_j)$ and $\Delta\alpha_j = \alpha_j - \alpha_{j-1}$. This implicates the assumption that the neighbouring coefficients are more similar. In other words, more information about the specification of the model is used and simultaneously some structure is brought into the penalization. For example in brand choice related context, if it is reasonable to assume that the actual purchase is more similar to the last purchase, this could be a reason to choose PS instead of TPS. At the same time, one could suggest that when there is some structure in the penalization, the model can be estimated using less knots. These are issues which will be addressed in this paper, as there has not been a lot of research in this area.

2.4 Model estimation

GAM's based on one-dimensional smoothing functions are fitted by means of a penalized likelihood approach. In practice, this amounts to solving the penalized iteratively re-weighted least squares problem (PIRLS). Hastie and Tibshirani (1986) developed an algorithm which is analogous to this formulation, as it is shown that it converges to the maximum expected log-likelihood. This so called local scoring algorithm was designed for a binary dependent variable. Abe (1999) further extended their method for multivariate qualitative response in the MNL model. By simulation studies he showed that his proposed model can fit several nonlinear structures, even when the distributional assumption of the random term is violated. The local scoring procedure includes a nonparametric regression, which is conducted by a weighted version of the backfitting algorithm (Buja et al. (1989)). Here, partial residuals of the corresponding covariates are iteratively smoothed. The advantage of this method is that any type of smooth functions can be used. Despite the critique that diagnostics can be hard to obtain (for further explanation see Section 2.5), this approach is still widely used due to its computational efficiency.

2.5 Smoothing parameters

As mentioned in Section 2.3, the smoothing parameter λ controls the influence of the penalty on the basis function. Ideally, the value of λ is optimal when the estimated function is close to the true function.

Different techniques can be used to determine λ. One way of estimating λ is by General Cross Validation (Gu and Wahba (1991)). However, when the backfitting algorithm is used, the selection of the parameter at each iterate becomes computationally costly when there are more than two or three covariates. Therefore, Wood (2004) came up with a numerical optimization of the cross-validation. Another way to estimate λ is by the Restricted Maximum Likelihood approach (REML). The advantage of this method is that it converges faster than GCV, despite the fact that it is also less robust than GCV. In case of a GAM model usually the REML method is preferred, because it is most effective.

3 Data

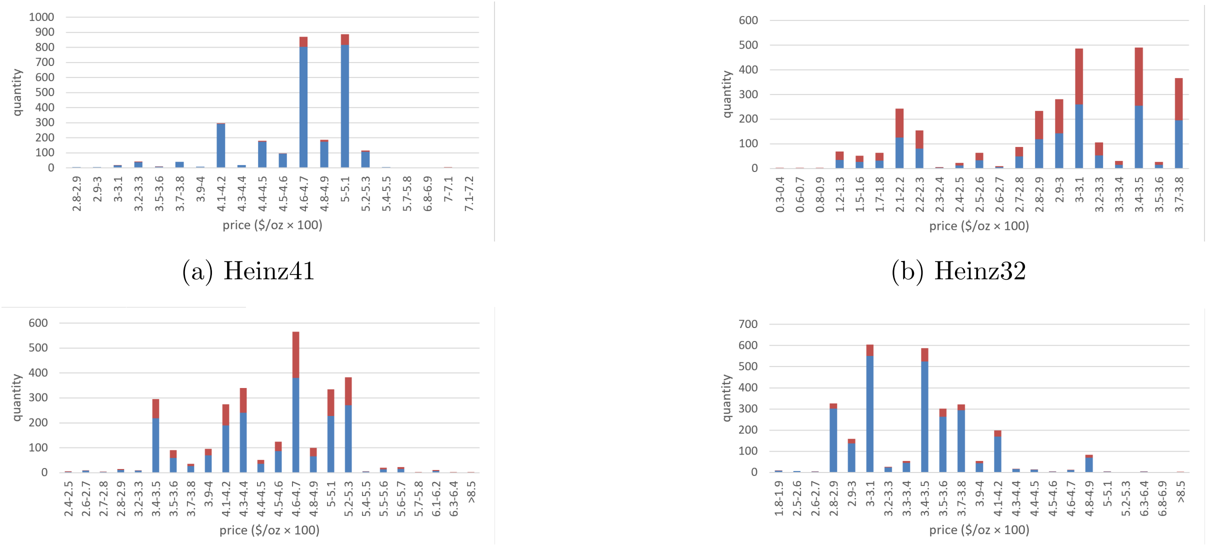
The data set which is used in this research, was first used by Jain et al. (1994). It contains cross-section data of 2798 individuals in Springfield, Missouri with regard to their choice of brand for catsup. They could choose from four different alternatives: Heinz41, Heinz32, Heinz28 or Hunts32. For each brand z, three variables are known: $disp.z$, $feat.z$ and $price.z$. The first two are dummy variables indicating whether there is a display or feature advertisement for brand z respectively. The last variable represents the price of brand z. Table 2 gives some descriptive statistics on the data set. Note that the price variable is the average sample price across all purchases.

Table 2: Descriptive statistics of the catsup data set.

Brand	Share (%)	Average price ($\$/oz \times 100$)	Fraction feature	Fraction display
Heinz41	0.07	4.63	0.08	0,13
Heinz32	0.52	3.14	0.10	0.15
Heinz28	0.30	4.32	0.13	0.19
Hunts32	0.11	3.36	0.09	0.19

From Table 2 one can see that Heinz32 is the most popular choice and has the lowest average price. Note that Heinz28 has the second largest share, while it is also second most expensive option. With regard to advertisement, Heinz28 is on promotion most often.

Figure 1 shows how the prices per brand are distributed. The price occurrence as well as the purchase frequency per price are presented. Especially for Heinz32 and Hunts32 one can clearly see when there is a price discount.



(a) Heinz41

(b) Heinz32

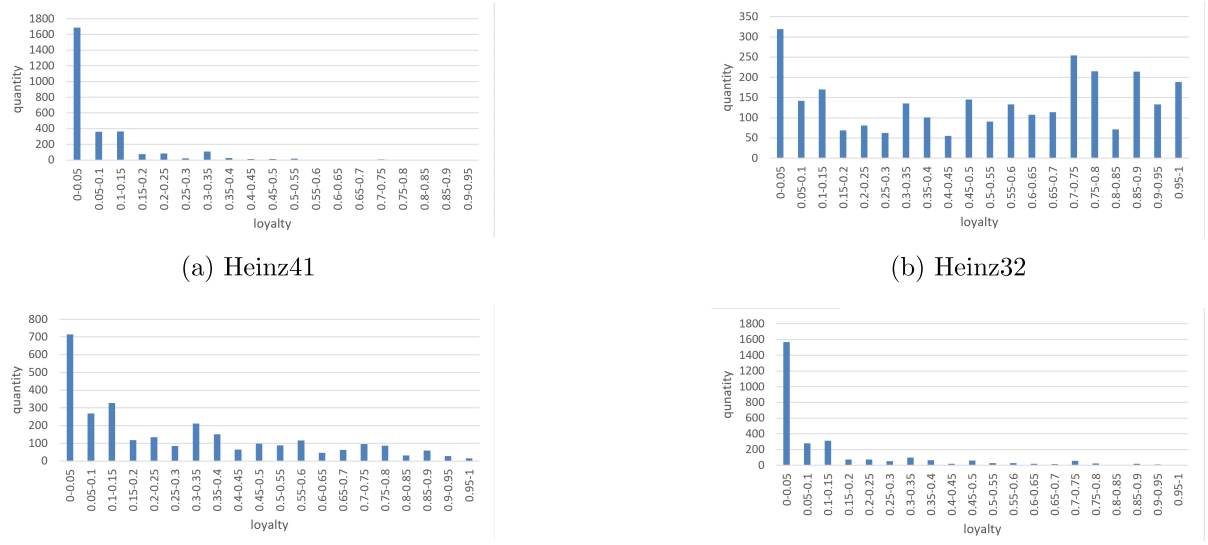
(c) Heinz28

(d) Hunts32

Figure 1: Histograms of price occurrence (blue) and purchase frequency (red) per catsup brand.

Individuals with less than five purchases are excluded from the data set, since we are mainly interested in frequent purchase behavior. As a result, 236 from the 300 individuals are left. The data will be divided into a calibration sample to fit the model and a holdout sample to assess the performance of the fit. The holdout sample consists of the last purchase of every individual, such that the rest of the 2242 observations become the calibration sample.

Furthermore, we create a *loyalty* variable, which takes a value between 0 and 1. This term represents the tendency of consumers to keep buying the same brand, where a larger value means a higher loyalty. It is created by means of the method Guadagni and Little (1983) used. The formula to estimate the brand loyalty for each individual is stated in Appendix A.1. It includes a so-called carry-over constant δ which is approximated to be 0.7 by trial-and-error. The spread of the loyalties of all 236 customers as a whole are given in Figure 2.



(a) Heinz41

(b) Heinz32

(c) Heinz28

(d) Hunts32

Figure 2: Histograms of loyalties per catsup brand.

As expected, in general the loyalties are higher for the brands that are bought more often (Heinz32 and Heinz28). For the other two less popular brands (Heinz41 and Hunts32) the loyalties are close to zero most of the time.

4 Methodology

Using the data set described in Section 3, we start with a model consisting of an alternative specific constant (asc) and include the two dummie variables $feature$ and $display$. This is the fixed parametric part of the model. Consequently, we add the $price$ variable and the $loyalty$ variable. Using these variables, we will construct eleven different model specifications by alternatively relaxing the parametric structure of $price$ and $loyalty$. The nonparametric functions will be estimated by the GAM model introduced by Abe (1999). The model is fitted using the local scoring and backfitting algorithm, which are stated in the Appendix A.2 and A.3. As smoothing technique we initially follow his choice of the knot-based approximation of TPS, which is described in Section 4.1. Subsequently, we implement PS instead, see Section 4.2, since this method could be a better choice in brand choice behaviour context. The smoothing parameter λ is in both cases determined by REML. First of all, the performances of both techniques are compared by means of the measures described in Section 4.4. After that, we vary the number of the knots to examine how the estimation accuracy changes.

4.1 Thin Plate Splines

Let us focus on a GAM model which is to be estimated by a one-dimensional smoothing function $s(\mathbf{x})$. If we apply a knot-based approximation of TPS, the basis functions stated in formula 3 of Section 2.3 are formed by joining a set of polynomial functions. This is done during the smoothing process. It can be shown that maximizing the penalized log-likelihood of the TPS is approximated by solving:

$$\hat{s}(\mathbf{x}) = \sum_{q=1}^{m+d} \delta_q \eta_m(\|\mathbf{X} - \mathbf{B}_q(\mathbf{x})\|) + \sum_{q=m+d+1}^{Q} \alpha_q^2 \phi_q(\mathbf{x}). \tag{4}$$

Here, α and δ are the parameters to be estimated and $\mathbf{X}$ is the design matrix. The first part of $\mathbf{X}$ is the regular part containing all the introduced covariates. The second part includes the knots, so this part is being penalized. Besides, d gives the dimension of the vector $\mathbf{x}$, which is 1 in our case. If m is the order of the smoothing penalty, then the function η_m is defined as

$$\eta_m(r) = \frac{\Gamma(1/2 - m)}{2^{2m}\pi^{1/2}(m-1)!} r^{2m}. \tag{5}$$

This research assumes m equals 2, as higher order penalties are hard to bring into practice. Furthermore, $\phi_q(\mathbf{x})$ is obtained by a recursion, see Wood (2017) for the explicit formulation. In practice, this comes down to solving the PIRLS problem

$$\text{minimize} \quad \|\mathbf{W}(\mathbf{y} - \mathbf{X}\beta)\|^2 + \lambda\beta'\mathbf{S}\beta \quad \text{subject to} \quad \mathbf{C}\beta = \mathbf{0}, \quad \text{w.r.t. } \beta' = (\delta', \alpha'). \tag{6}$$

Here, the identification matrix $\mathbf{C}$ is needed when more than one covariate is smoothed. In that case, a shift of one component can be compensated by an equal shift in reversed direction of another component. Usually, the constraint is set such that each smoothing function has an average of zero.

To bring these method into practice, we apply the modified local scoring algorithm (Abe (1999)) in combination with the backfitting algorithm, since this approach is computationally more efficient.

4.2 P-splines

When PS are implemented instead of TPS, the basis functions take the form of so called "tent-functions", which are formed by joining linear functions. Unlike TPS, these are defined beforehand.

Furthermore, the penalties imply d-th order differences on the spline coefficients α. In this research we focus on second order differences, which can be written as: $D_2\alpha = \Delta(\Delta\alpha_q) = \alpha_q - 2\alpha_{q-1} + \alpha_{q-2}$. The penalty can then be summarized as follows:

$$Pen = \sum_{i=1}^{q-1}(\alpha_{i+1} - \alpha_i)^2 = \alpha'\mathbf{D}_2'\mathbf{D}_2\alpha, \tag{7}$$

where

$$\mathbf{D}_2 = \begin{bmatrix} -1 & 1 & 0 & \cdot & \cdot \\ 0 & -1 & 1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad \text{s.t.} \quad \begin{bmatrix} \alpha_2 - \alpha_1 \\ \alpha_3 - \alpha_2 \\ \cdot \\ \cdot \end{bmatrix} = \mathbf{D}_2\alpha$$

and hence

$$Pen = \alpha'\mathbf{D}_2'\mathbf{D}_2\alpha = \alpha'\mathbf{P}\alpha = \alpha' \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \alpha.$$

The PIRLS problem then becomes

$$\text{minimize} \quad \|\mathbf{W}(\mathbf{y} - \mathbf{B}\alpha)\|^2 + \lambda\alpha'\mathbf{P}\alpha. \tag{8}$$

Hypothetically, this method is more suitable than TPS, as it is not inappropriate to assume that the next purchase more similar to last purchase, with regard to the catsup data. Furthermore, it can make sense to change the number of knots, as we can speak of a structured price, meaning that you can see when there is a price discount. In other words, here will be some price classes that are very close together. In that case, it can make sense to have less knots.

4.3 Constructing the set of knots

The basis dimension q should match with k, the number of knots supplied. In case of TPS these values are equal to each other, while for PS this is not the case. To define q B-spline bases, $k + m + 2$ knots need to be provided, where m is the order of the smoothing penalty which is two in our case. Besides, the middle $k - m$ knots should include all values of the covariate. As the bases are strictly local, the first and last $m + 1$ knot locations are chosen arbitrarily, to avoid multiple knots at both ends of the domain. Figure 3 shows both correct and incorrect placement of boundary knots for B-splines in case they are used in P-splines.

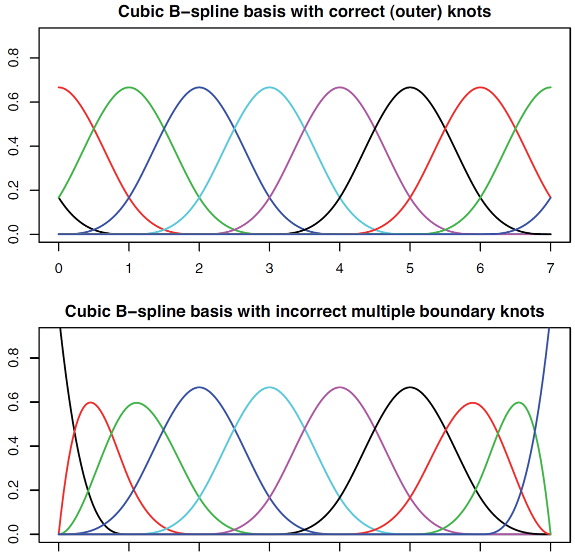


Figure 3: Different knot specifications for B-splines bases. Top: single knots at both ends of the domain, which is suitable for P-splines. Bottom: multiple knots at both ends of the domain, which is unsuitable for P-splines. Source: Eilers et al. (2015)

With regard to the number of knots, one should decide on this value carefully. When one degree of freedom is lost due to the identification constraint, then $k-1$ sets the upper limit on the degrees of freedom. Therefore, one should make sure that k is large enough such that there can be enough degrees of freedom to fit the true curve accurately. One the other hand, k should not be too large, since then the fitting becomes computationally costly.

4.4 Model evaluation

As described at the start of Section 4, different model specifications are constructed by alternatively relaxing the parametric structure of *price* and *loyalty*. The models are compared by means of four measures: the log-likelihood value, the BIC, the AIC and the hit-rate.

To evaluate changes in the log-likelihood, a the Likelihood Ratio Test is carried out. For example, one can evaluate whether the performance of a model significantly improves when a parametric function under the null hypothesis is replaced by a nonparametric specification.

$$LR = -2(l(\hat{\theta}_1) - l(\hat{\theta}_0)). \tag{9}$$

Here, $l(\hat{\theta}_0)$ en $l(\hat{\theta}_1)$ are the log-likelihood values under the null (H_0) and alternative (H_1) hypotheses, respectively. The test-statistic is g Chi-squared distributed. Here, g is the difference of degrees of freedom between the specification under H_0 and H_1. Note that this informal statistical inference is possible because Hastie and Tibshirani (1990) developed a method to approximate the degrees of freedom in the class of GAM's. Defining the degrees of freedom in terms of dimensionality, as is done in Ordinary Least Squares, is not useful here. However, since a nonparametric model is still linear in the observations, it can be presented as $\hat{y} = Sy$, where S is the design matrix. In that case, the effective degrees of freedom can be approximated by

$$\text{dof} = \text{tr}(2S - SS'). \tag{10}$$

Despite the fact that the models can be improved by adding more parameters, it can also cause an overfitting. Therefore, the Bayesian Information Criteria (BIC) is investigated as well, to compare different model specifications. Note that the likelihood-ratio test is only applicable when the models are nested, while this is not the case for the BIC. This measure is based on the estimated log-likelihood $l(\hat{\theta})$ and a penalty is added for the number of parameters r. The model with the lowest value is preferred. If n is the sample size then the statistic is formulated as:

$$\text{BIC} = -2l(\hat{\theta}) + log(n)r. \tag{11}$$

When we zoom in on one model specification and increase the number of knots, the change in the log-likelihood value will be small. Put differently, the coefficient of the penalty will probably be disproportionately big, even though the change in number of parameters will be small as well. Therefore, we also look at the Akaike Information Criteria (AIC) to compare models with different number of knots, since the coefficient of the penalty is smaller:

$$\text{AIC} = -2l(\hat{\theta}) + 2r. \tag{12}$$

Note that the AIC en BIC only give information about the quality relative to other models. Therefore, to assess the absolute performance of the fit, we will perform a one-step-ahead forecast for each specification using the holdout sample. These predictions will be evaluated using the sample hit-rate, which is a measure for the model accuracy. If each individual chooses the alternative with the highest predicted probability, then the hit-rate can be seen as the fraction of correct predicted observations. The hit-rate for the out-of-sample predictions should be about as good as the in-sample predictions.

5 Results

The performance measures and predictions described in Section 4.4 will be performed on eleven different model specifications. First of all, Section 5.1 provides a summary of the performances for all (non)parametric specifications. After that, the smooth functions of best performing models are investigated in Section 5.2, as well as the number of knots in Section 5.3.

5.1 Performances of (non)parametric specifications

To start with, Table 3 shows the results of all parametric specifications for both the calibration and holdout sample. Consequently, Tables 4 and 5 present the results of the nonparametric counterparts using either TPS or PS, respectively. The basis dimension q is set to 45, as this seems to give best fits overall.

Table 3: Performances of all parametric specifications using the catsup data set.

		Calibration sample			Holdout sample	
Specification	dof	BIC	logLik	hit-rate	logLik	hit-rate
M1. asc	3.00	4675.78	-2326.32	0.54	-285.93	0.46
M2. asc, disp	6.00	4395.27	-2174.49	0.59	-268.81	0.54
M3. asc, disp, feat	9.00	4349.83	-2140.20	0.60	-267.31	0.54
M4. asc, disp, feat, price	12.00	3784.20	-1845.81	0.65	-201.40	0.64
M6. asc, disp, feat, loy	12.00	1535.12	-721.27	0.84	-107.62	0.83
M8. asc, disp, feat, price, loy	15.00	1332.27	-608.27	0.86	-82.57	0.86

Table 4: Performances of all non parametric specifications using TPS with 45 knots.

		Calibration sample			Holdout sample	
Specification	dof	BIC	logLik	hit-rate	logLik	hit-rate
M5. asc, disp, feat, f(price)	39.97	3796.83	-1734.65	0.66	-193.39	0.64
M7. asc, disp, feat, f(loy)	56.37	1196.89	-368.69	0.90	-104.88	0.79
M9. asc, disp, feat, f(price), loy	35.12	1358.97	-528.48	0.87	-79.06	0.86
M10. asc, disp, feat, price, f(loy)	49.45	1070.75	-327.19	0.90	-71.37	0.88
M11. asc, disp, feat, f(price), f(loy)	61.97	1118.57	-285.42	0.92	-67.44	0.86

Table 5: Performances of all non parametric specifications using PS with 45 knots.

		Calibration sample			Holdout sample	
Specification	dof	BIC	logLik	hit-rate	logLik	hit-rate
M5. asc, disp, feat, f(price)	33.74	3798.38	-1751.13	0.65	-194.57	0.64
M7. asc, disp, feat, f(loy)	32.26	1199.98	-461.98	0.88	-102.20	0.77
M9. asc, disp, feat, f(price), loy	28.21	1320.16	-540.51	0.87	-77.74	0.88
M10. asc, disp, feat, price, f(loy)	41.21	1044.27	-349.76	0.90	-71.59	0.87
M11. asc, disp, feat, f(price), f(loy)	53.75	1088.17	-308.12	0.91	-68.13	0.87

First of all, in Table 3 one can clearly see the performance improving when more variables are added to the parametric model: the log-likelihood and hit-rate of the calibration sample increase, while the BIC decreases. For the holdout sample similar results are shown and the hit-rates are slightly smaller than the hit-rates of the calibration sample, as expected.

When the nonparametric relaxation of either *price* (M5) or *loyalty* (M7) using TPS in Table 4 are compared to their linear-in-parameter counterpart (models M4 and M6 respectively), the likelihood ratio test easily rejects the latter models at $\alpha = 0.001$. Likewise, models M9 and M10 are preferred in comparison to the parametric model M8 in which all variables are included. Furthermore, when models M9 and M10 are compared to the fully nonparametric model M11, the latter has a significantly better fit based on the log-likelihood. However, when models M8-M11 are compared based on the BIC and the hit-rate, then M10 is preferred. This is a sign that M11 is slightly overfitted.

The results for the models with smooth functions based on PS are stated in Table 5. Again, we conclude that M10 fits the data best. Furthermore, if the specifications are compared to the TPS in Table 4 by means of the log-likelihood, we conclude TPS give better fits. However, if we focus on the BIC then PS is preferred for models M9 to M11 due to a slight decrease in degrees of freedom.

With regard to the degrees of freedom, note that this value substantially increase when the covariates are estimated by smoothing functions instead of linear-in-parameters. This makes sense, as the degrees of freedom give an indication of the complexity of the model. When going from parametric to nonparametric implementation, more information about the data points is used, so the complexity of the model increases as do the degrees of freedom.

5.2 Smooth curves of best fitting models

Figure 4 shows the smooth curves of the three *loyalty* covariates for model M10 using PS, where Heinz41 is the base category.

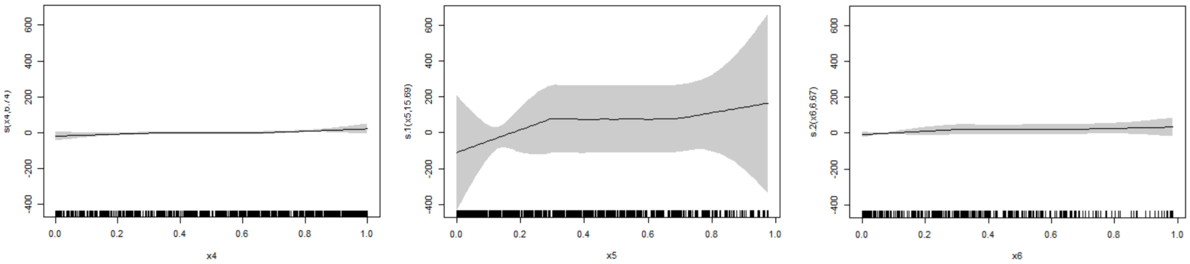


Figure 4: Smooth curves for model M10 using PS , where the shaded areas are the confidence intervals and $x4 = loy.heinz32$, $x5 = loy.heinz28$, $x6 = loy.hunts32$. The black stripes on the x-as represent the observation values.

In general, the fits try to stay as close to zero as possible. Not surprisingly, the curve of *loy.heinz*32 (left) is most accurately estimated, as there the loyalty values are evenly spread over the domain (also compare to Figure 2). On the other hand, *loy.heinz*28 (middle) has relatively wide confidence intervals in comparison to *loy.hunts*32 (right). We expected this to be the other way around, as *loy.hunts* has few values close to 1 which would make the estimate more uncertain.

When the variable *price* is also estimated by smooth functions using PS, this results in model M11. In that case, the smooth curves of *loyalty* do not look different from those in Figure 4, while the smooth curves of the three *price* covariates are shown in Figure 5. Here, one can clearly see that when there are more observations on a certain part of the domain, the estimated curve becomes more accurate there. Note that when the smooth functions are based on TPS instead, this results in similar looking curves.

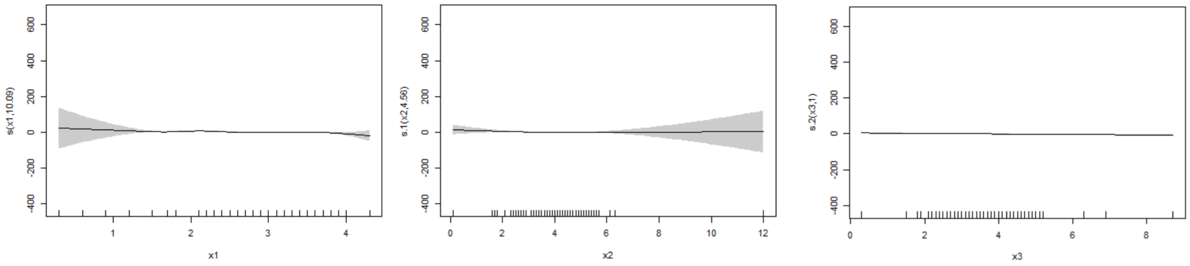
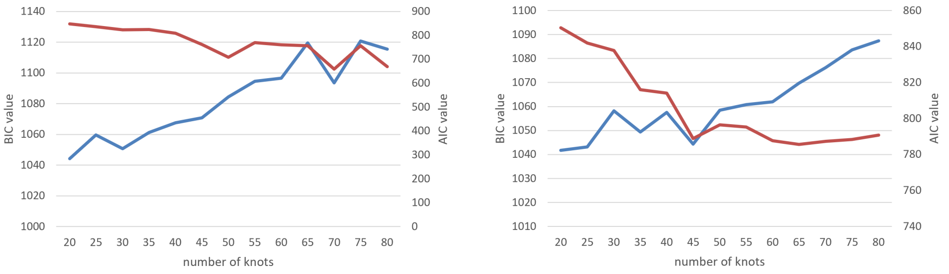


Figure 5: Smooth curves for the variable *price* of model M11 using PS, where the shaded areas are the confidence intervals and $x1 = price.heinz32$, $x2 = price.heinz28$, $x3 = price.hunts32$. The black stripes on the x-as represent the observation values.

5.3 Changing the number of knots

Since model M10 seems to represent the data most accurate, we concentrate on this model and change the number of knots supplied for the smoothing function of the variable *loy*. The results are shown in Table 6.



(a) TPS

(b) PS

Figure 6: BIC value (blue) and AIC value (red) when changing the number of knots for model M10 based on the calibration sample.

First of all, one can see that the AIC is decreasing while the BIC is not, as expected. This is due to the fact that the increase in log-likelihood is smaller than the penalty that is added in case of the BIC. If we focus on the models with PS, the AIC is steadily decreasing for the splines with up to 45 knots. Likewise, the BIC is between 1040 and 1060 for the splines with up to 45 knots, after that the BIC starts increasing. In other words, adding more knots does not result in a better model. For the models with TPS this trend is less clear. When M11 is investigated instead, the same aspects can be detected. These figures are represented in Appendix A.5.

As stated in Section 4.3, when the number of knots increases, so does the computation time to construct the model. For models M10 and M11, the average computation time is under twenty seconds for the splines with up to 45 knots, regardless of the smoothing technique. After that, the average computation time increases and for more than 65 knots it takes between 100 and 200 seconds to estimate the concerning model. We can conclude that not only the performances do not improve anymore after 45 knots, also the estimation becomes computationally costly. In addition, note that especially for models with more than 45 knots the computation time varies a lot. This is probably due to the fact that we are using Maximum Likelihood estimation, which has randomly chosen starting values. When the starting value is close to the optimal solution, the computation time can appear relatively low.

6 Conclusion

The utility functions used in the MNL model to describe consumer brand choice data are usually specified as linear in the parameters. This can be an unrealistic assumption due to the fact that individuals have a subjective perception of price. As a result, the function of price will be nonlinear. Considering this, it is appealing to implement nonparametric functions instead, which can be done by spline smoothing. In this context, we address the following research questions: "*How do P-splines compare in performance to Thin Plate Splines?*" and "*How does the number of knots influence the estimation accuracy?*".

To address these questions, we construct different model specifications, which are evaluated by means of their performances. First of all, as expected, the non-parametric relaxations outperform their parametric counterparts regardless the technique, since more information about the data points is used. However, it can possibly result in an overfitting. In other words, it is not always better to use smoothing splines for every covariate.

With regard to the first research question, we conclude that one technique is not significantly better in performance than the other. In other words, we can not confirm the hypothesis that PS would be more suitable than TPS in brand choice related context. Note, that these conclusions are all taken within the context of the used consumer data set about catsup brand choices.

One could repeat the research with a data set where the time dependence between purchases is stronger, to be able to draw more precise conclusions.
Furthermore, more research can be done on the comparison of other smoothing techniques. For example, the approximated knot-based TPS we used can be compared in performance to the *full* TPS approach.

Concerning the second question, we conclude that the number of knots matters, but does not seem to differ per spline smoothing technique. When including more knots, the performance of the fit increases. However, after a certain point, the computation time increases significantly, while the performance does hardly improve anymore. In other words, we can not confirm the hypothesis that one can use less knots when there is a structured price. Again, one could repeat the research with a data set where the price classes per category are more distinct, to be able to draw better conclusions.
Another suggestion could be to vary the number of knots per variable. In this research, we assumed the same number of knots for both variables *price* and *loyalty*. This is not self-evident as their domains are different.
Furthermore, the knots were evenly placed over the domain for simplicity. Possibly, the fits can be improved when the knots are placed depending on the density of the data points over the domain.

References

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

R. A. Briesch, P. K. Chintagunta, and R. L. Matzkin. Semiparametric estimation of brand choice behavior. *Journal of the American Statistical Association*, 97(460):973–982, 2002.

A. Buja, T. Hastie, R. Tibshirani, et al. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510, 1989.

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

P. H. Eilers and B. D. Marx. Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(6):637–653, 2010.

P. H. Eilers, B. D. Marx, and M. Durbán. Twenty years of p-splines. *SORT: statistics and operations research transactions*, 39(2):0149–186, 2015.

C. Gu and G. Wahba. Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing*, 12(2):383–398, 1991.

P. M. Guadagni and J. D. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

S. Gupta and L. G. Cooper. The discounting of discounts and promotion thresholds. *Journal of consumer research*, 19(3):401–411, 1992.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

T. Hastie and R. Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

T. Hastie and R. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016, 1990.

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

G. Kalyanaram and J. D. Little. An empirical analysis of latitude of price acceptance in consumer package goods. *Journal of consumer research*, 21(3):408–418, 1994.

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

S. N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.

S. N. Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

S. N. Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.

A Appendix

A.1 Creating brand loyalty variable

Source: Guadagni and Little (1983)

The brand loyalty variable is an exponentially weighted average of past purchase decisions. For individual i and brand j at time unit t the brand loyalty $b_{i,j,t}$ is defined as

$$b_{i,j,t} = \delta b_{i,j,t-1} + (1-\delta)I[y_{i,t} = j], \quad \text{with} \quad 0 < \delta < 1. \tag{13}$$

Here, δ is the carry-over constant which can be chosen by trial-and-error. In case there is a total of J alternatives, to initialize the brand loyalty, use

$$\begin{aligned} b_{i,j,1} &= \delta && \text{if} \quad y_{i,1} = j \\ b_{i,j,1} &= (1-\delta)/(J-1) && \text{if} \quad y_{i,1} \neq j. \end{aligned} \tag{14}$$

A.2 Modified local scoring algorithm

Source: Abe (1999)

Initial estimate by the linear model, $\eta(x_{nj}) = \beta' x_{nj}$ for all n and j

Repeat

Compute the current estimate of μ_{nj} from η as

$$\mu_{nj} = \frac{e^{\eta(x_{nj})}}{\sum_k e^{\eta(x_{nk})}}$$

Compute the adjusted dependent variable z_{nj} and the weight w_{nj}, where

$$z_{nj} = \eta(x_{nj}) + \frac{y_{nj} - \mu_{nj}}{\mu_{nj}(1-\mu_{nj})}$$

$$w_{nj} = \mu_{nj}(1-\mu_{nj})$$

Obtain $f_p(x_p)(p = 1, ..., P)$ by nonparametric regression of z_{nj} on x_{nj} with weight w_{nj} (by the backfitting procedure)

Until log-likelihood converges

A.3 Backfitting algorithm

Source: Friedman and Stuetzle

Initialization by the linear model: $f_p(x_p) = \beta_p x_p$ for all $p = 1, ..., P$

Cycle over the explanatory variables: $p = 1, ..., P, 1, ..., P, 1...$

$$f_p(x_p) = E\left\{z - \sum_{q \neq p} f_q(x_q)|x_p\right\}$$

Until change in the functions, $f_p(x_p)$, is sufficiently small.

A.4 Estimation results for model M10 using PS

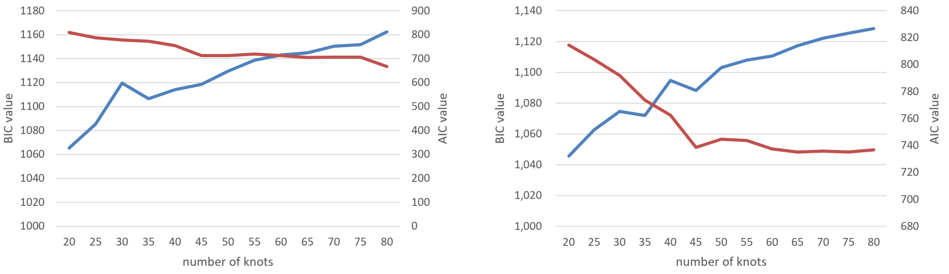
Table 6: Significance of parametric coefficients of model M10 using PS.
Significance codes: * is p-value less than 0.05, ** is p-value less than 0.01, *** is p-value less than 0.001.

Covariate	Estimate	Std.Error	z value	$Pr(>\|z\|)$
c.heinz32	10.52	3.68	2.86	4.30e-03 **
disp.heinz32	0.85	0.41	2.09	3.70e-02 *
feat.heinz32	1.13	0.48	2.36	1.82e-02 *
price.heinz32	-2.43	0.27	-8.99	< 2e-16 ***
c.heinz28	-68.09	94.99	-0.72	0.47
disp.heinz28	1.61	0.48	3.38	7.33e-04 ***
feat.heinz28	2.05	0.47	4.35	1.39e-05 ***
price.heinz28	-0.93	0.18	-5.15	2.54e-07 ***
c.hunts32	-12.81	12.86	-1.00	0.32
disp.hunts32	1.58	0.67	2.34	1.92e-02 *
feat.hunts32	1.79	0.79	2.26	2.38e-02 *
price.hunts32	-1.31	0.37	-3.53	4.12e-04 ***

Table 7: Approximate significance of smooth terms of model M10 using PS.
Significance codes: * is p-value less than 0.05, ** is p-value less than 0.01, *** is p-value less than 0.001.

smooth term	edf	Chi.sq	p-value
s(loy.heinz32)	6.87	74.21	1.55e-12 ***
s(loy.heinz28)	15.81	88.61	8.30e-12 ***
s(loy.hunts32)	6.53	24.45	2.35e-03 **

A.5 Changing the number of knots for model M11



(a) TPS

(b) PS

Figure 7: BIC value (blue) and AIC value (red) when changing the number of knots for model M11 based on the calibration sample.

A.6 R-code

- **reproduction_TPS**: data descriptives, parametric model specifications, nonparametric model specifications using TPS, changing knots M10 using TPS, model evaluation.
- **extension_PS**: nonparametric model specifications using P-splines, changing knots M10 using PS, model evaluation.