Applications of nonparametric methods in economic and political science

Dissertation

presented for the degree of Doctor rerum politicarum

at the Faculty of Economic Sciences

of the Georg-August-Universität Göttingen

by Nils-Bastian Heidenreich

from Wittmund, Germany

Göttingen, 2011

First Examiner: Prof. Dr. Stefan Sperlich
Second Examiner: Prof. Dr. Fred Böker

Disputation: 11.04.2011

Contents

List of Figures

List of Tables

Acknowledgement

I am grateful to my supervisor, Prof. Dr. Stefan Sperlich, for his support, helpful comments and the many valuable suggestions. Especially, I would like to thank for the light pressure at the end of my thesis in order to finish the work expeditiously. My special thanks go to Prof. Dr. Fred Böker for his valuable help and the very pleasant collaboration during my work at the institute. Furthermore, I would like to mention Prof. Dr. Maik Hammerschmidt, whom I thank not only for his friendly cooperation, but also for his willingness to act as examiner.

I wish to extend my thanks to Prof. Dr. Walter Zucchini for his support and time when it mattered the most. In addition, I thank all other members of the Institute for Statistics and Econometrics for their friendly cooperation and practical help in various respects. Especially, I would like to thank my colleague Dr. Roland Langrock. Without his permanent support in professional as well as personal manner, this dissertation would not have seen the daylight.

Lastly, I would like to thank my family for their unconditional encouragement throughout my academic career and my girlfriend for her loving support and patience; without them it would not have been possible for me to get this thesis together.

1. Introduction

"Statistical inference is the process of drawing conclusions from data that are subject to random variation" (Upton and Cook, 2008). Inferential statistics requires assumptions which can be summarized in a statistical model. These models can be roughly split up into two main approaches. Both of these approaches make use of specific assumptions for statistical modeling to draw statistical inference.

On the one hand, the parametric world in which statistical models need assumptions about the underlying distribution, e.g. the normal distribution with mean μ and variance σ^2. The parameter of this distribution can be estimated out of the present sample. By substituting the estimates into the formula of the parametric distribution, the underlying function f can be calculated. Based on the assumption of a parametric distribution, statisticians can draw conclusions, calculate confidence intervals and test various hypotheses about the data.

On the other hand, the idea of nonparametric modeling requires less rigid a priori statements about the underlying structure of the distribution. Hence, to make inferential statements in nonparametric estimation, no assumption about a specific functional form is needed, except of smoothness. This becomes obvious, e.g. in terms of efficiency, if the distributional assumptions of a parametric model are not fulfilled (Büning and Trenkler, 1994). Although, the term 'nonparametric' indicates that actually no parameters exist, we need some particular parameters, e.g. the bandwidth parameter h to control for smoothness of the estimates. However, these parameters are not components of an underlying distribution and hence several authors use the term distribution-free instead of nonparametric (see, e.g. Kotz and Johnson, 1982).

Nonparametric methods are used in a wide range of statistical problems. A simple case is density estimation, in which we have a continuous random variable X with some unknown continuous distribution and want to estimate the probability density function (pdf) $f(x)$. Another fundamental example is nonparametric regression, in which the conditional expectation of a dependent variable Y given some explanatory covariates $\mathbf{X} = (X_1, X_2, ..., X_n)$ is modeled by an unknown, but smooth, function of the covariates. We can estimate the regression function without predefining a parametric structure, such as linear or cubic.

In the context of density estimation the most widely used nonparametric method is the histogram. To overcome some shortcomings of histograms, the idea of kernel density estimation was proposed by Rosenblatt (1956) and Parzen (1962). In nonparametric regression several approaches have been developed. The most traditional approach of nonparametric kernel regression is the Nadaraya-Watson estimator (Nadaraya, 1964 and Watson, 1964). Nowadays, the field of literature concerning kernel density estimation or kernel regression is very extensive. Several aspects are covered, ranging from theoretical considerations about statistical properties or asymptotics to applications for manifold real-data problems. Due to these developments, a lot of data structures, like categorical data, time-series data or panel data can be tackled with nonparametric methods.

In several real data problems the researcher is faced with the problem of how to describe the relationship between dependent and independent variables of the dataset at hand by a statistical model. Purely linear parametric models may be easy applicable but they are far from adequate in many situations. Fully nonparametric models are very flexible, yet not always appropriate. If many continuous covariates need to be included in the model, nonparametric methods run into the curse of dimensionality, i.e. the models require an exponential increase in observation points. Many real data applications impose a natural classification of the explanatory variables. Hence, one wants to separate the estimation problem into two or more groups of design variables. In order to capture these problems several models were developed in recent years. Altogether they try to reduce the high dimensionality of multivariate nonparametric models. Therefore, they divide the covariates into distinct groups some modeled parametrically and some nonparametrically using an additive or bivariate structure. Such models are usually called semiparametric models. A general introduction about non- and semiparametric models can be found in Härdle et. al. (2004).

This thesis is structured as a cumulative dissertation and combines three papers which are concerned with a particular topic discussed in the first part of this introduction. All papers are supposed to be published together with Prof. Dr. Stefan Sperlich and further co-authors, in different statistical journals. The three main chapters of this thesis (chapter 2 to 4) are based on the original text of the papers, including an individual introduction, conclusion as well as a reference list. Also the meaning of the notation is given by time in the corresponding paper. The rest of this thesis can be seen as a frame for these three projects. Thus, the thesis is organized as follows. After this short introduction, the second chapter will give an idea of how to choose an optimal bandwidth parameter in nonparametric density estimation problems. Chapter 3 presents a semiparametric approach for discrete choice modeling with an application in the context of political science. In the fourth chapter this semiparametric approach will be extended to a data problem with a

slightly different covariate structure in the context of an urban transport problem. Finally, an overall conclusion will summarize the main results of the three projects and will point out the contribution of this thesis to current research. In the appendix, the whole programming code is given.

The second chapter of this dissertation compares various automatic bandwidth selection methods in the context of kernel density estimation. Density estimation is a very basic instrument in nonparametric statistics. A lot of bandwidth selection methods and programming routines exists to calculate an optimal bandwidth. The essential idea is to find an optimal bandwidth selection method which is always suitable and can be used automatically for every dataset at hand. Although several authors are concerned with solving the problem of finding the optimal bandwidth, currently no procedure is established which is appropriate for all estimation problems. Most authors propose a more or less limited method which fits in their particular case, resulting in a variety of bandwidth selection methods. Contrary, researchers with less mathematical background trust the most established and traditional selection methods, such as cross-validation or Silverman's rule of thumb.

The examination and evaluation of different bandwidth selection methods is a never-ending story as several authors almost annually publish new ideas to tackle their problems. Due to these reasons, the discussion of bandwidth selection methods is a very ambitious task. On the one hand, we had to deal with several mathematical algorithms from different theoretical backgrounds; on the other hand, we faced an extensive development of new methods in a short time period, which makes it almost impossible to keep the discussion up to date. Therefore, the first essay summarizes the state of the art in the context of bandwidth selection and updates former review papers on bandwidth selection, like Jones, Marron and Sheather (1996). However, we had to restrict the consideration to methods which were available up to 2009.

Chapter 3, proposes a solution for modeling multicategorical data with the aid of non- and semiparametric methods. One goal of this project is the consideration of the distribution of the dependent variable which is assumed to be categorical. The classical linear approach models the expectation of an at least approximately normal distributed target variable as a function of covariates. This is often inappropriate, e.g. when observing a discrete target variable. The basic statistical framework to model almost arbitrarily distributed response variables was proposed by Nelder and Wedderburn (1972) in their work about generalized linear models. Thereby, a function of the expectation is modeled by a predictor of the covariates which are considered in the model. Afterwards, this concept found its way into the field of non- and semiparametric modeling. An overview about several generalized non-

and semiparametric regression models can be found in Härdle et.al. (2004). Müller (2001) proposed a model for binary response variable in the context of a specific semiparametric approach using a generalized partial linear model. As mentioned earlier, a researcher has to transform the structure of a dataset at hand into a statistical model to draw inferences. In several real-data problems individuals face a discrete set of alternatives. Such problems are usually handled with multinomial or conditional logit models. The classical multinomial logit model, which models the predictor in a linear manner, imposes rigid restrictions on the covariates. The model in this essay overcomes these deficiencies by allowing a very flexible semiparametric modeling of the predictor. Therefore, we can detect nonlinearities as well as strong interactions between covariates. We investigate the influence of several covariates on the political party affiliation in a multi-party system like Germany. The data represent a typical situation for a multinomial logit model, in which the individual is faced with the choice between unordered alternatives. The support of political parties is modeled by several covariates which are divided into two distinct groups. On the one hand, some covariates which are modeled parametrically and on the other hand, some of them influence the predictor in a nonparametric way. Thereby, we use an approach based on a generalized partial linear model and extend this approach to multicategorical response variable. We consider additive as well as bivariate model structures in the nonparametric part and by means of three-dimensional plots we can identify various voter profiles for the political parties which can be very useful for decision makers at head of different political parties.

The fourth chapter of this thesis considers an urban transport problem, in which the students of the University in Bilbao are faced with a discrete set of choices for a transport mode to travel to the university. The underlying data structure is similar to the second project. The data were obtained by a written query from students of the University in Bilbao. The aim of this paper is methodological. In this project, we compare different approaches: fully parametric, non- or semiparametric, binary as well as multinomial models, including only individual- or also mode-specific covariates. We compare the results of the different models and give recommendations as to which of the considered models provides the best conclusions in the context of transport policy.

2. Bandwidth Selection Methods for Kernel Density Estimation – A Review of Performance

Abstract

On the one hand, kernel density estimation is a common tool for empirical studies in any research area. This goes hand in hand with the fact that these estimators are provided by many software packages. On the other hand, since about three decades the discussion on bandwidth selection has been going on. A good part of the discussion is concerned about nonparametric regression, but this issue is by no means less problematic for density estimation. This becomes obvious when reading empirical studies in which practitioners made use of kernel densities. Unfortunately, software packages offer only simple cross validation or Silverman's rule of thumb. New contributions typically provide simulations limited to show that the own invention outperforms existing methods. We review existing methods and compare them on a set of designs that exhibits features like few bumps and exponentially falling tails concentrating thereby on small and moderate sample sizes. Our main focus is on practical issues like fully automatic procedures, implementation and performance where the latter one is measured in many ways.
This essay is based on a joint work with my colleague Anja Schindler and Prof. Dr. Stefan Sperlich. The main contribution of the author of this thesis is made in the evalutaion of the cross-validation methods and the presentation of all estimation results.

2.1. Introduction

Suppose we have observed i.i.d. data $X_1, X_2, \ldots, X_n$ from a common distribution with density $f(\cdot)$, and we aim to estimate this density using the standard kernel (i.e. the Parzen-Rosenblatt) estimator

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.1}$$

where K is a kernel and h the bandwidth parameter. The problem is to find a reliable data driven estimator of the optimal bandwidth. First one has to decide on a method of assessing the performance of $\widehat{f}_h$. The generally accepted performance measures are the integrated squared error

$$\text{ISE}(h) \quad = \quad \text{ISE}\{\widehat{f}_h(x)\} = \int \{\widehat{f}_h(x) - f(x)\}^2 \, dx \tag{2.2}$$

or alternatively, the mean integrated squared error, i.e.

$$\text{MISE}(h) = \text{MISE}\left[\hat{f}_h(x)\right] = \int \text{MSE}\left[\hat{f}_h(x)\right] \, dx. \tag{2.3}$$

Let us denote the minimizers of these two criteria by $\widehat{h}_0$ and h_0 respectively. The main difference is that $ISE(h)$ is a stochastic process indexed by $h > 0$, while $MISE(h)$ is a deterministic function of h, see Cao (1993). Therefore we distinguish two classes of methods: the cross-validation methods trying to estimate $\widehat{h}_0$ and therefore looking at the ISE, and the plug-in methods which try to minimize the MISE to find h_0. It is evident that asymptotically these criteria coincide.

The main part of the nonparametric statistical community has accepted that there may not be a perfect procedure to select the optimal bandwidth. However, we should be able to say which is a reasonable bandwidth selector, at least for a particular problem. SiZer tries to show the practitioner what is a range of reasonable bandwidths, and is therefore quite attractive for data snooping, see Chaudhuri and Marron (1999) for an introduction, Godtliebsen, Marron and Chaudhuri (2002) for an extension to the bivariate case. Hanning and Marron (2006) made an improvement using extreme value theory. However, SiZer does not give back one specific data driven bandwidth as practitioners typically ask for.

Since until now the development of bandwidth selectors has been continuing, we believe it is helpful to review and compare the existing selectors to get an idea of the objective and performance of each selector. As we have listed more than 30 bandwidth selectors - several of them being modifications for particular estimation problems - we decided to restrict this study in mainly two directions. Firstly, we considered independent observations. Secondly,

we looked at smooth densities, namely we use four underlying distributions which are mixtures of at most three different normal and/or gamma distributions. This type of smoothness covers a broad range of problems in any research area; it is clearly rather different from estimating sharp peaks or highly oscillating functions. However, the latter problems should not be tackled with kernels anyway. Density problems with extreme tails are not included. It is well known that those problems should be transformed, see e.g. Wand, Marron and Ruppert (1991) or Yang and Marron (1999) for parametric, Ruppert and Cline (1994) for nonparametric transformations. After an appropriate transformation the remaining estimation problem falls into the here considered class, too. Note that the limitation to global bandwidths is not very restrictive neither, and quite common in density estimation. Actually, when X is transformed, and similar smoothness is assumed over the whole transformed support, then global bandwidths are most reasonable. Finally, we have restricted our study to already published methods.

There are already several papers dealing with a comparison of different automatic data driven bandwidth selection methods though, most of them are older than ten years. In the seventies and early eighties survey papers about density estimation were published by Wegman (1972), Tartar and Kronmal (1976), Fryer (1977), Wertz and Schneider (1979) as well as Bean and Tsokos (1980). A short introduction to various methods of smoothing parameter selection without a simulation study was released by Marron (1988a) and Park and Marron (1990). Then, extensive simulations studies have been published by Park and Turlach (1992), Marron and Wand (1992), Cao, Cuevas and Gonzáles Manteiga (1994) and Chiu (1996). A brief survey is also announced by Jones, Marron and Sheather (1996a) and a more comprehensive one in the companion paper Jones, Marron and Sheather (1996b). A very exhaustive simulation study has been published by Devroye (1997). Furthermore, Loader (1999) has published a comparison paper. In recent years to our knowledge only Chacón, Montanero and Nogales (2008) have published a comparison paper on this topic. Therefore, our paper is also an update of the two main review papers of Jones, Marron and Sheather (1996b) and Devroye (1997) although the overlap is moderate.

The idea of cross validation methods (LSCV) goes back to Rudemo (1982) and Bowman (1984), but one could also mention the pseudo-likelihood CV-methods of Habbema, Hermans and van den Broek (1974) or of Duin (1976). Due to the lack of stability of this method, see Wand and Jones (1995), different modifications have been proposed like the stabilized bandwidth selector recommended by Chiu (1991), smoothed CV proposed by Hall, Marron and Park (1992), the modified CV (MCV) by Stute (1992), or the version by Feluch and Koronacki (1992), and most recently the one-sided CV by Martínez-Miranda, Nielsen and Sperlich (2009) and the indirect CV by Savchuk, Hart and Sheather (2010). The biased CV (BCV) of Scott and Terrell (1987) is minimizing the asymptotic MISE like

plug-in methods do but uses a jack-knife procedure (therefore called CV) to avoid the use of prior information. The recent kernel contrast method of Ahmad and Ran (2004) can be used for MISE minimization as well, but it is not really data adaptive (or fully automatic) and it performs particularly well rather for regression than for density estimation. In the most exhaustive former comparison papers, Jones, Marron and Sheather (1996b) considered only LSCV, BCV and Chiu in this class, whereas Devroye (1997) applied only the LSCV in his simulations.

Compared to CV the so-called plug-in methods do not only minimize a different objective function, MISE instead of ISE, they are less volatile but not entirely data adaptive as they require pilot information. In contrast, CV allows to choose the bandwidth without making assumptions about the smoothness (or the like) to which the unknown density belongs. Certainly, if we have an appropriate pilot bandwidth the performance of plug-in methods is pretty good. Although, they have a faster convergence rate compared to CV, they can heavily depend on the choice of pilots. Among them, Silverman's (1986) rule of thumb is probably the most popular one. Various refinements were introduced like for example by Park and Marron (1990), Sheather and Jones (1991), or by Hall, Sheather, Jones and Marron (1991). Also the bootstrap methods of Taylor (1989) as well as all its modifications, see e.g. Cao (1993) or Chacón, Montanero and Nogales (2008), we count to plug-in methods as they aim to minimize the MISE. As representatives of the plug-in class almost all former simulations studies concentrate on Silverman's rule of thumb and the plug-in versions of Park and Marron (1990) or alternatively Sheather and Jones (1991). Jones, Marron and Sheather (1996b) additionally applied HSJM and some further refinements, which proved to be unfavorable. Devroye (1997) includes various plug-in versions and bandwidth selectors using a reference density, based on minimizing the L1-norm. Only few comparison papers consider bootstrap methods in their simulation. The most recently is Chacón, Montanero and Nogales (2008). However, they restricted to Bootstrap methods and only compare LSCV and the plug-in version of Sheather and Jones (1991).

The general criticism against the two classes of selection methods can be summarized as follows: CV leads to undersmoothing and breaks down for large samples, whereas plug-in depends on prior information and often works bad for small data sets and much curvature.

For the statements about asymptotic theory, we make the following assumptions on kernel and density. For some methods we will modify them.

(A1) The kernel K is a compactly supported density function on $\mathbb{R}$, symmetric around zero with Hölder-continuous derivative, K'.

(A2) $\mu_2(K) < \infty$, where $\mu_l(K) = \int u^l K(u)du$.

(A3) The density, f, is bounded and twice differentiable, f' and f'' are bounded and integrable, and f'' is uniformly continuous.

In our simulation study we restrict on selection methods which consider no higher order kernels. The main motivation for the usage of higher order kernels is their theoretical advantage of faster asymptotic convergence rates. However, their substantial drawback is a loss in the practical interpretability as they involve negative weights and can even give back negative density estimates. A good illustration of the understanding of higher order kernels can be found in Marron (1994).

In the context of asymptotic theory we are aware of the trade-off between the classical plug-in method and standard cross-validation. The plug-in has always smaller asymptotic variance compared to cross-validation (Hall and Marron, 1987a). To our knowledge, no other bandwidth selection rule has outperformed the asymptotic properties of the plug-in method. Although Hall and Johnstone (1992) stated that such methods must theoretically exist, they couldn't give any practical example.

2.2. Cross-Validation methods in density estimation

Recall the used performance measure, i.e. the integrated squared error (ISE):

$$\mathrm{ISE}(h) = \int \widehat{f}_h^2(x)\,dx - 2\,E\{\widehat{f}_h(X)\} + \int f^2(x)\,dx. \tag{2.4}$$

Evidently, the first term can be calculated from the data, the second can be expressed as the expected value of $\widehat{f}_h(X)$, and the third term can be ignored since it does not depend on the bandwidth. Note that estimating $E\{\widehat{f}_h(X)\}$ by $\frac{1}{n}\sum_{i=1}^n \widehat{f}_h(X_i)$ is inadequate due to the implicit dependency ($\widehat{f}_h$ depends on X_i). So the different modifications of CV basically vary in the estimation of the problematic second part.

2.2.1. Ordinary least squares cross-validation

This is a straightforward approach by just dropping X_i when estimating $f(X_i)$, called jack-knife estimator and denoted by $\widehat{f}_{h,-i}(X_i)$. It yields the *least-squares CV criterion*

$$\min_h \ \mathrm{CV}(h) = \int \widehat{f}_h^2(x)\,dx - 2\,\frac{1}{n}\sum_{i=1}^n \widehat{f}_{h,-i}(X_i). \tag{2.5}$$

Stone (1984) showed that under the assumptions (A1)-(A3), the minimizing argument, $\widehat{h}_{CV}$, fulfills $\mathrm{ISE}(\widehat{h}_{CV})\,\{\min_h \mathrm{ISE}(h)\}^{-1} \xrightarrow{a.s.} 1$. However, Hall and Marron (1987a) stated

that this happens at the slow rate of $O_p(n^{-1/10})$. Many practitioners use this CV method because of its intuitive definition and practical flavor. But as mentioned, it is not stable, tends to undersmooth and often breaks down for large samples.

2.2.2. Modified cross-validation

Stute (1992) proposed a so-called modified CV (MCV). He approximated the problematic term by the aid of a Hajek projection. In fact, he showed that under some regularity assumptions given below, $2E[f_h(x)]$ is the projection of

$$\begin{aligned} S + \frac{1}{h} E\left[K\left(\frac{X_1 - X_2}{h}\right)\right] &= S + \frac{1}{h}\int\int K\left(\frac{x-y}{h}\right) f(x)\, f(y)\, dx\, dy \\ &= S + \int f^2(y)dy + \frac{1}{2}h^2 \int t^2 K(t)dt \int f(y) f''(y)dy + O(h^3) \\ \text{for} \qquad S &:= \frac{1}{n(n-1)h}\sum_{i\neq j} K\left(\frac{X_i - X_j}{h}\right). \end{aligned}$$

This gives the criterion

$$\min_h \; MCV(h) = \int \widehat{f}_h^2(x)dx - S - \frac{\mu_2(K)}{2n(n-1)h}\sum_{i\neq j} K''\left(\frac{X_i - X_j}{h}\right). \tag{2.6}$$

It can be shown then that under assumptions (A1),

(A2') K is three times differentiable, with $\int t^4|K(t)|\,dt < \infty$, $\int t^4|K''(t)|\,dt < \infty$, $\int t^4[K'(t)]^2\,dt < \infty$, and $\int t^2[K'''(t)]^2\,dt < \infty$,

(A3') f four times continuously differentiable, the derivatives being bounded and integrable,

you get the following consistency result:

$$\frac{\mathrm{ISE}(\widehat{h}_0)}{\mathrm{ISE}(\widehat{h}_{\mathrm{MCV}})} \xrightarrow{P} 1, \quad \textit{and} \quad \frac{\widehat{h}_0}{\widehat{h}_{\mathrm{MCV}}} \xrightarrow{P} 1 \quad \textit{as} \quad n \to \infty.$$

2.2.3. Stabilized bandwidth selection

Based on characteristic functions Chiu (1991) gave an expression for h_{CV} which reveals the source of variation. Note that the CV criterion is approximately equal to

$$\frac{1}{\pi}\int_0^\infty |\tilde{\phi}(\lambda)|^2 \left\{w^2(h\lambda) - 2w(h\lambda)\right\} d\lambda + 2K(0)/(nh), \tag{2.7}$$

with $\tilde{\phi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} e^{i\lambda X_j}$ and $w(\lambda) = \int e^{i\lambda u}K(u)du$. The noise in the CV estimate is mainly contributed by $|\tilde{\phi}(\lambda)|^2$ at high frequencies, which does not contain much information about f. To mitigate this problem, he looks at the difference of the CV criterion and the MISE. Chiu defines Λ as the first λ fulfilling $|\tilde{\phi}(\lambda)|^2 \leq 3/n$ and replaces $|\tilde{\phi}(\lambda)|^2$ by $1/n$ for $\lambda > \Lambda$. This gives his criterion:

$$\begin{aligned} \min_h \; S_n(h) &= \int_0^{\Lambda} |\tilde{\phi}(\lambda)|^2 \left\{w^2(h\lambda) - 2w(h\lambda)\right\} d\lambda \\ &\quad + \frac{1}{n}\int_{\Lambda}^{\infty} \left\{w^2(h\lambda) - 2w(h\lambda)\right\} d\lambda + 2\pi K(0)/(nh), \qquad (2.8) \\ &= \frac{\pi}{nh}||K||_2^2 + \int_0^{\Lambda} \left\{|\tilde{\phi}(\lambda)|^2 - \frac{1}{n}\right\} \{w^2(h\lambda) - 2w(h\lambda)\} d\lambda, \qquad (2.9) \end{aligned}$$

with $||g||_2^2 = \int g^2(u)\, du$. For the minimizer, $\widehat{h}_{ST}$, of this criterion, it can be shown that $\widehat{h}_{ST} \xrightarrow{a.s.} \widehat{h}_0$, and it converges to h_0 at the optimal $n^{-1/2}$-rate. In the calculation of Λ we came across with the computation of square roots of negative terms in our simulations. To avoid complex numbers we calculated the absolut value of the radicand. Note that in the literature this procedure is often counted among the plug-in methods as it minimizes the MISE.

2.2.4. One-sided cross-validation

Marron (1986) made the point that the harder the estimation problem the better CV works. Based on this idea Hart and Yi (1998) introduced an estimation procedure called one-sided cross-validation in the regression context. They concluded that one-sided cross validation (OSCV) in regression clearly outperforms the ordinary CV. Martínez-Miranda, Nielsen and Sperlich (2009) extended OSCV to density estimation. They apply estimator (2.1) but with a local linear version of a one sided kernel,

$$\bar{K}(u) = \frac{\mu_2(K) - u\left(2\int_{-\infty}^{0} tK(t)\, dt\right)}{\mu_2(K) - \left(2\int_{-\infty}^{0} tK(t)\, dt\right)^2} 2K(u)\mathbf{1}_{\{u<0\}}. \qquad (2.10)$$

Respectively to ISE and MISE they define the one-sided versions OISE and MOISE, with their minimizers $\widehat{b}_0$ and b_0. The one-sided CV criterion is

$$\min_b \; \mathrm{OSCV}(b) = \int \widehat{f}^2_{left,b}(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{left,b}(X_i), \qquad (2.11)$$

where $\widehat{f}_{left,b}$ is the one-sided (to the left) kernel density estimator. Then they define the corresponding bandwidth for the "real" estimation problem by

$$\widehat{h}_{OSCV} := C \cdot \widehat{b}_{OSCV} \quad \text{with} \quad C = h_0/b_0. \qquad (2.12)$$

Note that C is deterministic and depends only on kernel K since

$$h_0 = \left(\frac{||K||_2^2}{(\mu_2(K))^2 ||f''||_2^2 n} \right)^{1/5} \quad , \quad b_0 = \left(\frac{||\bar{K}||_2^2}{(\mu_2(\bar{K}))^2 ||f''||_2^2 n} \right)^{1/5} . \tag{2.13}$$

This gives, for example $C \approx 0.537$ for the Epanechnikov kernel. The theoretical justification for the improved convergence rate of one-sided CV is based on the result of Hall and Marron (1987a) that under the assumptions (A1) - (A3)

$$n^{3/10}(\widehat{h}_{CV} - \widehat{h}_0) \longrightarrow N(0, \sigma^2 c^{-2}). \tag{2.14}$$

with known terms σ and c depending only on f and K. From this we can calculate the variance reduction of OSCV compared to CV by $\{C\bar{\sigma}c/(\bar{c}\sigma)\}^2$ where $\bar{c}$, $\bar{\sigma}$ are just as c, σ but with $\bar{K}$ instead of K. The reduction of the variance for the Epanechnikov kernel is at least 35% and 50% for the Gaussian kernel. Note that $\bar{K}$ can also be constructed as a one sided kernel to the right.

2.2.5. Further cross-validation methods

Feluch and Koronacki (1992) proposed to cut out not only X_i when estimating $f(X_i)$ but rather dropping also the $m < n$ nearest neighbors with $m \to \infty$ such that $m/n \to 0$. They called this version *modified CV*. Unfortunately, it turned out that the quality of this method crucially depends on m. Therefore it cannot be considered as *automatic* or *data driven*, and will not be considered further. This idea is similar to the CV selection for time series data, see Härdle and Vieu (1992).

Scott and Terrell (1987) introduced the Biased CV. They worried about unreliable small-sample results, i.e. the high variability while using the cross-validation criterion. However, they directly focused on minimizing the asymptotic MISE and estimated the unknown term $||f''(x)||_2^2$ via jack-knife methods. Already in their own paper they admitted a poor performance for small samples and mixtures of densities, see also Chiu (1996). In their simulation study, Jones, Marron and Sheather (1996b) underlined the deficient performance from 'quite good' to 'very poor'.

The smoothed cross-validation (SCV) was evolved by Hall, Marron and Park (1992). The general idea is a kind of presmoothing of the data before applying the CV-criterion. This procedure of presmoothing results in smaller sample variability, but enlarges the bias. Therefore the resulting bandwidth is often oversmoothing and cuts off some features of the underlying density. With this method it is possible to achieve a relative order of convergence of $n^{-1/2}$ but only when using a kernel of order ≥ 6. In total, it seems to

be appropriate - if at all - only for huge samples. Without using a higher order kernel Jones, Marron and Sheather (1996b) stated, that there exists an $n^{-1/10}$ convergent version of SCV that is identical to Taylor's bootstrap, see Taylor (1989). Additionally, with a special choice for g SCV results in an $n^{-5/14}$ version similar to a diagonal-in version of Park and Marron's plug-in, see Park and Marron (1990). Note that finally the SCV is closely related to the bootstrap method of Cao (1993). These three methods do not belong to the cross-validation methods, and hence, they are discussed later. In conclusion, we have not implemented these methods, because either it is very similar to other methods or it is necessary to use a higher order kernel.

Similar to the idea of one-sided cross validation, Savchuk, Hart, and Sheather (2010) introduce three classes of selection kernels, all being different from one-sided kernels: $(1+\alpha)\phi(u) - \alpha/\sigma\ \phi(u/\sigma)$ with ϕ being the standard normal density, and for different combinations of α and σ. Then, like in one-sided cv, LSCV is performed on an estimator with a selection kernel, and afterwards the bandwidth can be derived that corresponds to the kernel used in estimator (2.1). When looking at the MISE minimizing properties, this method exhibits excellent theoretical properties. For our implementation with Epanechnikov kernels it nevertheless worked well only for large samples and proper choices of α and σ. It seems to us that for obtaining a practical, fully automatic selection procedure, some additional work is necessary.

The partitioned cross-validation (PCV) was suggested by Marron (1988b). He modified the CV-criterion by splitting the sample of size n into m subsamples. Then, the PCV is calculated by minimizing the average of the score functions of the CV-score for all subsamples. In a final step the resulting bandwidth needs to be rescaled. The number of subsamples affects the trade off between variance and bias. Hence the choice of m is the smoothing problem in this case. As Park and Marron (1990) noticed: "this method ... is not quite fully objective". Another drawback is the required separation of the subsamples.

The pseudo-likelihood (also called the Kullback-Leibler) cross-validation (invented by Habbema, Hermans and van den Broek (1974) and by Duin (1976)) aims to find the bandwidth maximizing a pseudo-likelihood criterion with leaving out the observation X_i. Due to the fact that lot of authors criticize this method being inappropriate for density estimation, we skipped also this method in our simulation study.

Wegkamp (1999) suggests a method being very much related to the cross-validation technique providing quasi-universal bandwidth selection for bounded densities. Nonetheless, his paper stays on a rather technical level but is not suitable for practitioners.

Recently, Ahmad and Ran (2004) proposed a kernel contrast method for choosing bandwidths either minimizing ISE or alternatively the MISE. While it turned out to work quite

well for regression, the results for density estimation were less promising. A major problem is that one needs two series of contrast coefficients which have a serious impact on the performance of the method. As we are not aware of an automatic data driven and well performing method to choose them, we will not consider this method further.

2.3. Plug-in methods in density estimation

Under (A1)-(A3) the MISE can be written for $n \to \infty$, $h \to 0$ as

$$\text{MISE}\left[\hat{f}_h(x)\right] = \frac{h^4}{4}\mu_2^2(K)||f''(x)||_2^2 + \frac{1}{nh}||K||_2^2 + o\left(\frac{1}{nh}\right) + o(h^4), \tag{2.15}$$

such that the asymptotically optimal bandwidth is

$$h_0 = ||K||_2^{2/5}\left(||f''||_2^2\left[\mu_2(K)\right]^2 n\right)^{-1/5}, \tag{2.16}$$

where only $||f''||_2^2$ is unknown and has to be estimated. The most popular method is the "rule-of-thumb" of Silverman (1986). He uses the normal density as a prior for approximating $||f''||_2^2$. For the estimation of the standard deviation of X he proposes a robust version. If the true underlying density is unimodal, fairly symmetric and does not have fat tails, it works very well.

2.3.1. Park and Marron's refined plug-in

Natural refinements consist of using nonparametric estimates for $||f''||_2^2$. Let us consider

$$\widehat{f}_g''(x) = \frac{1}{ng^3}\sum_{i=1}^{n} K''\left(\frac{x - X_i}{g}\right),$$

where g is a prior bandwidth. Hall and Marron (1987b) proposed several estimators for $||f''||_2^2$, all containing double sums over the sample. They pointed out that the diagonal elements give a non-stochastic term which does not depend on the sample and increases the bias. They therefore proposed the bias corrected estimator

$$\widehat{||f''||_2^2} = ||\widehat{f}_g''||_2^2 - \frac{1}{ng^5}||K''||_2^2. \tag{2.17}$$

The question which arises is how to obtain a proper prior bandwidth g. In Park and Marron (1990) g is the minimizer for the asymptotic mean squared error of $\widehat{||f''||_2^2}$. With (2.16) one gets a prior bandwidth in terms of h (using the notation in the original paper):

$$g = C_3(K)C_4(f)h^{10/13},$$

where $C_3(K)$ contains the fourth derivative and convolutions of K, and $C_4(f)$ the second and third derivatives of f. Substituting the normal with estimated variance for f gives

$$h = \left(\frac{||K||_2^2}{\widehat{||f''||_2^2}\mu_2^2(K)n} \right)^{1/5}. \tag{2.18}$$

The optimal bandwidth is then obtained by numerical solution of (2.18). The relative rate of convergence to h_0 is of order $n^{-4/13}$, which is suboptimal compared to the optimal $n^{-1/2}$-rate, cf. Hall and Marron (1991).

2.3.2. Refined plug-in

For small samples and small bandwidths, the above estimator $\widehat{||f''||_2^2}$ can easily fail in practice. Also, to find a numerical solution may become involved in practice. To avoid these problems and to offer a quick and easy solution, we propose to first take Silverman's rule-of-thumb bandwidth for Gaussian kernels, $h_S = 1.06 \min\{1.34^{-1}IR, s_n\}n^{-1/5}$ with IR =interquartile range of X, and s_n the sample standard deviation, adjusted to Quartic kernels. This is done via the idea of canonical kernels and equivalence bandwidths, see Härdle, Müller, Sperlich and Werwartz (2004). The Quartic which comes close to the Epanechnikov kernel but allows for second derivative estimation. Finally, we adjust for the slower optimal rate for second derivative estimation and obtain as a prior

$$g = h_S \frac{2.0362}{0.7764} n^{1/5-1/9} \tag{2.19}$$

for (2.17). This bandwidth leads to very reasonable estimates of the second derivative of f, and hence of $\widehat{||f''||_2^2}$. A further advantage is that this prior g is rather easily obtained. As the idea actually goes back to Park and Marron (1990) we will call the final bandwidth $\hat{h}_{PM}$.

2.3.3. Bootstrap methods

The idea of these methods is to select the bandwidth along bootstrap estimates of the ISE or the MISE. For a general description of this idea in nonparametric problems, see Hall (1990). Imagine, for a given pilot bandwidth g we have a Parzen-Rosenblatt estimate, $\widehat{f}_g$, from which we can draw bootstrap samples $(X_1^*, X_2^*, \ldots, X_n^*)$. Then, defining the bootstrap kernel density

$$\widehat{f}_h^*(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h} \right), \tag{2.20}$$

the (mean) integrated squared error to be minimized can be approximated by

$$ISE^*(h) := \int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 dx \,, \tag{2.21}$$

$$MISE^*(h) := E_* \left[\int \left(\widehat{f}_h^*(x) - \widehat{f}_g(x)\right)^2 dx\right] \,. \tag{2.22}$$

It can be shown that the expectation E_* and so the $MISE^*$ depends only on the original sample but not on the bootstrap samples. Consequently, there is actually no need to do resampling to obtain the $MISE^*$. More specific, using Fubini's theorem and decomposing the $MISE^* = V^* + SB^*$ into integrated variance

$$V^*(h) = \frac{1}{nh} \cdot ||K||_2^2 + \frac{1}{n} \cdot \int \left(\int K(u) \cdot \widehat{f}_g(x - hu)\, du\right)^2 dx \tag{2.23}$$

and squared bias

$$SB^*(h) = \int \left(\int K(u) \cdot (\widehat{f}_g(x - hu) - \widehat{f}_g(x))\, du\right)^2 dx \tag{2.24}$$

gives (where $\star$ denotes convolution)

$$V^*(h) = \frac{1}{nh}||K||_2^2 + \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(K_h \star K_g) \star (K_h \star K_g)\right](X_i - X_j) \tag{2.25}$$

and

$$SB^*(h) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[(K_h \star K_g - K_g) \star (K_h \star K_g - K_g)\right](X_i - X_j). \tag{2.26}$$

In practice, it is hard to get explicit formulae for these integrals when kernels have bounded support. However, using the Gaussian kernel in the formulae (2.25) and (2.26) we can directly calculate the optimal bandwidth as the minimizer of

$$\begin{aligned} \text{MISE}^*(h) \;=\; & \frac{1}{2nh\sqrt{\pi}} + \frac{1}{\sqrt{2\pi}} \left[\frac{\sum_{i,j} \left(\exp\left(-\frac{1}{2} \left(\frac{X_i - X_j}{g\sqrt{2}} \right)^2 \right) \right)}{\sqrt{2g^2} \cdot n^2} \right. \\ & \left. - \frac{2 \cdot \sum_{i,j} \left(\exp\left(-\frac{1}{2} \left(\frac{X_i - X_j}{\sqrt{h^2 + 2g^2}} \right)^2 \right) \right)}{\sqrt{h^2 + 2g^2} \cdot n^2} + \frac{(n+1) \sum_{i,j} \left(\exp\left(-\frac{1}{2} \left(\frac{X_i - X_j}{\sqrt{2(h^2 + g^2)}} \right)^2 \right) \right)}{\sqrt{2(h^2 + g^2)} \cdot n^3} \right] \end{aligned} \tag{2.27}$$

The equivalent bandwidth for any other kernel can be obtained as described in Marron and Nolan (1988).

The bootstrap approach in kernel density estimation was first presented by Taylor (1989). Many modified versions were published in the following, e.g. Faraway and Jhun (1990), Hall (1990) or Cao (1993). The crucial difference between these versions is the choice of the pilot bandwidth and the procedure to generate the resampling distribution.

Taylor (1989) suggested to take $g = h$ and used a Gaussian kernel. Several authors pointed out that this procedure has no finite minimum and hence chooses a local minimum or the upper limit of the bandwidths grid as its optimum, see Marron (1992). Differing from this approach, Faraway and Jhun (1990) proposed a least-square cross-validation estimate to find g. Hall (1990) recommended to use the empirical distribution to draw bootstrap samples of size $m < n$, proposed $m \simeq n^{1/2}$, $h = g(m/n)^{1/5}$, and minimized $MISE^*$ with respect to g. Cao, Cuevas and Gonzáles-Manteiga (1994) demonstrated that the bootstrap version of Hall is quite unstable and shows a bad behavior especially for mixtures of normal distributions, which make up the biggest part of our simulation study. They found also that the methods of Faraway and Jhun (1990) and Hall (1990) are outperformed by the method of Cao (1993), see below.

In the smoothed bootstrap version of Cao (1993) the pilot bandwidth g is estimated by asymptotic expressions of the minimizer of the dominant part of the mean squared error. For further details see Cao (1993). He noticed that in (2.26), for $i = j$ these terms will inflate the bias artificially. He therefore proposed a modified bootstrap integrated squared bias MB*

$$MB^*(h) = \frac{1}{n^2} \sum_{i \neq j} \left[(K_h \star K_g - K_g) \star (K_h \star K_g - K_g) \right] (X_i - X_j). \tag{2.28}$$

As to what concerns the convergence rates, he showed for his bandwidth h_0^*

$$\frac{\mathrm{MISE}(h_0^*) - \mathrm{MISE}(h_0)}{\mathrm{MISE}(h_0)} \quad = \quad O_P(n^{-5/7}) \tag{2.29}$$

and

$$\frac{\mathrm{MISE}(h_{0_M}^*) - \mathrm{MISE}(h_0)}{\mathrm{MISE}(h_0)} \quad = \quad O_P(n^{-8/13}) \tag{2.30}$$

The convergence rate for the original bootstrap version is slightly faster than that for his modified version.

Recently, Chacón, Montanero and Nogales (2008) published a bootstrap version quite similar to Cao's (1993). They showed that the asymptotic expressions of his bandwidth estimates might be inadequate and defined an expression $g(h)$ for every fixed h. Their estimation procedure allows different kernels L and K for the bandwidths g and h. They

calculated the optimal pilot bandwidth $g(h)$ using first the common way of reverting to a reference distribution, and afterwards via estimation. In their simulation study they stated that the former version outperforms the empirical approach, and is a good compromise between classical cross-validation and plug-in. However, it depends seriously on the reference density. On the contrary, the empirical version suffered from sample variability even more than classical CV. Exploring the asymptotics, they achieved root-n convergence under the use of higher-order kernels.

A bias corrected bootstrap estimate was developed by Grund and Polzehl (1997). They obtained an root-n convergent estimate which attained very good results for larger sample sizes, but only in few cases for moderate and small sample sizes. Moreover, to derive their asymptotic theory they used extraordinary strong assumptions, compared to other methods discussed here. In their simulation study Grund and Polzehl showed that the performance heavily depends on the choice of g. They stated that using their oversmoothing bandwidth, which provides a root-n convergence, seems to be far from optimal for smaller sample size. In contrast, using $g = h$ would achieve better performance in practical applications, but results in very slow convergence rate, namely of order $n^{-1/10}$. Summing up, they remarked that higher rates of convergence do not result in better practical performance, especially for small samples.

In sum, in our simulation study we concentrate on just one representative of the class of bootstrap estimates, going back to Cao (1993). He proved that the pilot bandwidth g as the minimizer of (2.22) coincides with the minimizer of the dominant part of the mean squared error. Concretely, it is given by

$$g = \left(\frac{||K||_2^2}{\widehat{||f'''||_2^2} \mu_2^2(K) n} \right)^{1/7}. \tag{2.31}$$

This formula is used for the pilot bandwidth g in the calculation of (2.27). In our simulations, we additionally ran the bootstrap for the Epanechnikov kernel calculating formulae (2.23) and (2.24) numerically. As this was much slower and gave uniformly worse results, we will neglect it for the rest of the paper.

2.3.4. Further plug-in methods

Many other plug-in methods have been developed. Some of them show better asymptotic properties and others a better performance in particular small sample simulations. However, most of them have not become (widely) accepted or even known.

An often cited method is the so-called Sheather and Jones (1991) bandwidth, see also

Jones and Sheather (1991). They used the same idea like Park and Marron (1990) but replaced the "diagonal-out" estimator of $||f''||_2^2$ by their "diagonal-in" version to avoid the problem that the estimator $\widehat{||f''||_2^2}$ (see (2.17)) may give negative results. They stated that the non-stochastic term in (2.17) is subducted because of its positive effect on the bias in estimating $||f''||_2^2$. The idea is to choose the prior bandwidth g such that the negative bias due to the smoothing compensates the impact of the diagonal-in term. As a result they estimated $||f''||_2^2$ by $||\widehat{f}_g''||_2^2$ which is always positive, and obtained

$$g = C(K, L) \left(\frac{||f''||_2^2}{||f'''||_2^2} \right)^{1/7} h^{5/7},$$

where $C(K, L)$ depends on L, a kernel introduced to estimate $||f''||_2^2$, and K, the kernel in the original estimation. Then, $||f''||_2^2$ and $||f'''||_2^2$ were estimated using $||\widehat{f}_a''||_2^2$ and $||\widehat{f}_b'''||_2^2$, where a and b were estimated via the rule-of-thumb. Sheather and Jones (1991) showed that their optimal bandwidth has a relative order of convergence to h_0 of $O_p(n^{-5/14})$ which is only slightly better than that of Park and Marron (1990). Jones, Marron and Sheather (1996b) indicates the closeness of h_{PM} to h_{SJ} for practical purposes in their real data application. Hence, without beating h_{PM} in practical performance, having only a slightly better convergence rate but being computationally much more expensive, we favor h_{PM} to h_{SJ}.

Hall, Sheather, Jones and Marron (1991) introduced a plug-in method giving back a bandwidth $\widehat{h}_{HSJM}$ which achieves the optimal rate of convergence, i.e. $n^{-1/2}$. The problem with $\widehat{h}_{HSJM}$ is that they use higher order kernels to ensure the $n^{-1/2}$ convergence (actually a kernel of order 6). Marron and Wand (1992) showed that albeit their theoretical advantages, higher order kernels have a surprisingly bad performance in practice, at least for moderate samples. Furthermore, in the simulation study of Park and Turlach (1992) $\widehat{h}_{HSJM}$ behaved very bad for bi- and trimodal densities, i.e. those we plan to study.

Jones, Marron and Park (1991) developed a plug-in method based on the smooth CV idea. They used the prior bandwidth $g = C(f)n^p h^m$, where the normal is used as reference distribution to calculate the unknown $C(f)$. The advantage of this estimator is the $n^{-1/2}$ convergence rate if $m = -2$, $p = \frac{23}{45}$ even if the kernels are of order 2. However, in simulation studies Turlach (1994) and Chiu (1996) observed a small variance compared to the LSCV, but an unacceptable large bias.

Kim, Park and Marron (1994) also showed the existence of a $n^{-1/2}$ convergent method without using higher order kernels. The main idea of obtaining asymptotically best bandwidth selectors is based on an exact MISE expansion. But primarily the results of this paper are provided for "theoretical completeness" because the practical performance in simulation studies for moderate sample sizes is rather disappointing, which was already

explicitly mentioned in their own paper and as well shown in the exhaustive simulation study of Jones, Marron and Sheather (1996b).

Not very well known is the "Double Kernel method" based on the L_2 loss function, see Jones (1998). He explores a modification of the L_1 based method proposed by Devroye (1989), see also Berlinet and Devroye (1994). This method is claimed to be quite universal. Under special assumptions it reduces to Taylor's bootstrap respectively biased CV. However, as already mentioned, these two methods have several disadvantages and also the Double Kernel method requires the use of higher order kernels. In Jones (1998) the performance of the Double Kernel method is assessed by comparing asymptotic convergence rates, but it does not provide the expected improvement in the estimation of h_0 (MISE optimal bandwidth), e.g. compared to SCV.

Finally, for further MISE minimizing selection methods recall Ahmad and Ran (2004), Savchuk, Hart, and Sheather (2010), and the so-called biased CV, all having been introduced in the section about cross validation methods.

2.4. Mixtures of methods

Recall that all authors criticize that the cross-validation criterion tends to undersmooth and suffers from high sample variability. At the same time, the plug-in estimates deliver a much more stable estimate but often oversmooth the density. We therefore also consider mixtures of classical cross-validation methods and plug-in estimates. Depending on the weighting factor $\alpha \in (0, 1)$, the mixed methods are denoted by Mix(α), with $\alpha \cdot \widehat{h}_{CV} + (1 - \alpha) \cdot \widehat{h}_{PM}$. We mix in three different proportions: Mix(1/2), Mix(1/3) and Mix(2/3). For the resulting mixed bandwidths we calculate the according ISE-value to assess the performance of the respective mix proportion.

We are aware of different approaches which combine various density estimators by using a mixture of their smoothing parameters. In the literature several papers address the problem of linear and/or convex aggregation, e.g. Rigollet and Tsybakov (2007), Samarov and Tsybakov (2007) as well as Yang (2000). However, as the main focus of this paper is not on the aggregation of different bandwidth estimators, we will not investigate this much in detail, but instead consider our mixtures as representatives.

2.5. Finite sample performance

The small sample performance of the different cross-validation methods, plug-in and bootstrap methods is compared, including Chiu (1991). For obvious reasons we limited the study to data adaptive methods without boundary correction. Although we tried many different designs we summarize here the results for four densities. We have compared the performance by different measures based on the integrated squared error (ISE) of the resulting density estimate (not the bandwidth estimate), and on the distance to the real optimal bandwidth $\widehat{h}_0$ (of each simulation run, as it is sample-dependent). There are a lot of measures assessing the quality of the estimators. We will concentrate on the most meaningful ones, that are:

m_1: $mean\left[ISE(\hat{h})\right]$, the average (or expected) ISE

m_2: $std\left[(ISE(\hat{h})\right]$, the volatility of the ISE

m_3: $mean(\hat{h} - \widehat{h}_0)$, bias of the bandwidth selectors

m_4: $mean\left(\left[ISE(\hat{h}) - ISE(\widehat{h}_0)\right]^2\right)$, squared L_2 distance of the ISEs

m_5: $mean\left[\mid ISE(\hat{h}) - ISE(\widehat{h}_0) \mid\right]$, L_1-distance of the ISEs.

Further, we considered various densities for our simulation study, but for sake of presentation we give only the results for the following ones:

1. Simple normal distribution, $\mathcal{N}(0.5, 0.2^2)$ with only one mode
2. Mixture of $\mathcal{N}(0.35, 0.1^2)$ and $\mathcal{N}(0.65, 0.1^2)$ with two modes
3. Mixture of $\mathcal{N}(0.25, 0.075^2)$, $\mathcal{N}(0.5, 0.075^2)$, $\mathcal{N}(0.75, 0.075^2)$ with three modes
4. Mixture of three gamma, $Gamma(a_j, b_j)$, $a_j = b_j^2$, $b_1 = 1.5$, $b_2 = 3$ and $b_3 = 6$ applied on $8x$ giving two bumps and one plateau

As can be seen in Figure 2.1, all densities have the main mass in $[0, 1]$ with exponentially decreasing tails. This way we can neglect possible boundary effects. Moreover, it is assumed that the empirical researcher has no knowledge on possible boundaries. We also simulated estimators with boundary corrections getting results very close to what we found in the present study.

We studied almost all selection methods, excluding the non-automatic ones and those having proved to perform uniformly worse than their competitors. In the presentation of the results we concentrate on the methods which delivered the best results at least for one density. Hence, some methods were dropped, e.g. the MCV sometimes provides

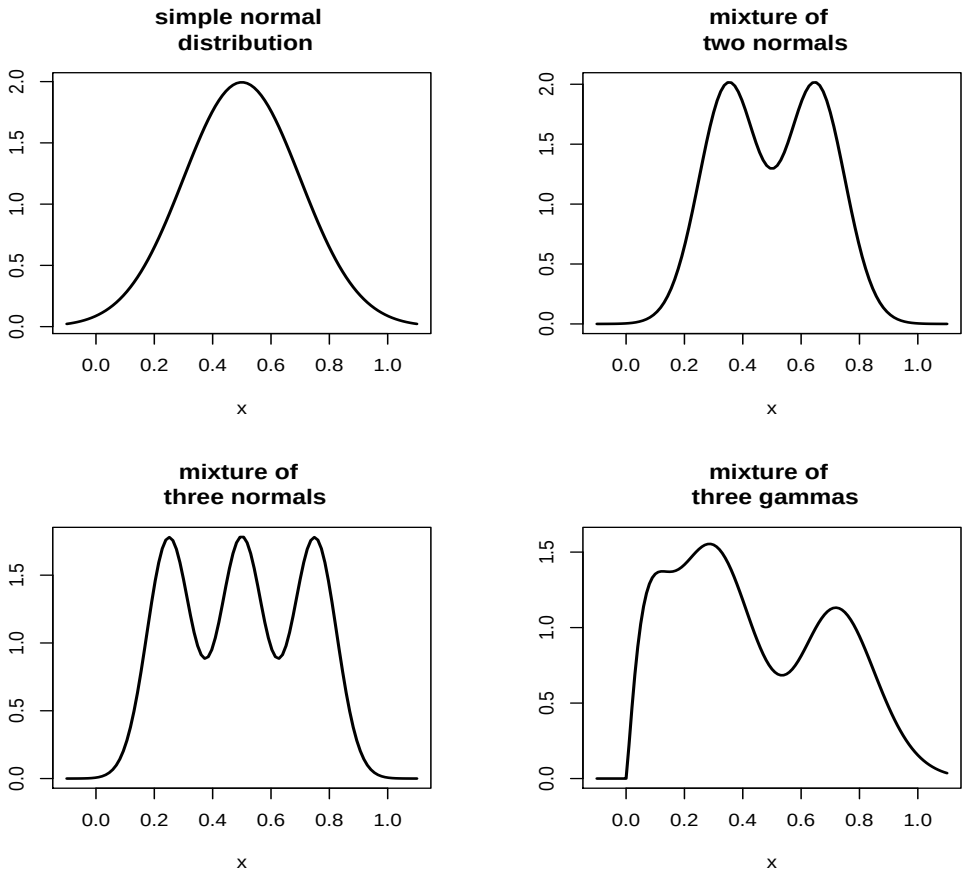

Figure 2.1.: *The data generating densities: design 1 to 4 from upper left to lower right.*

multiple minima with a global minima far outside the range of reasonable bandwidths. In the range of bootstrap methods we concentrate on the presentation of the version (2.27) of the Smoothed Bootstrap which obtained the best results among all bootstrap methods. For our mixed version (CV with refined plug-in) we first concentrate on Mix(1/2) when comparing it to the other methods, and later sketch the results of all mixed versions.

To conclude, we present the following methods: CV (cross validation), OSCV-l (one-sided CV to the left), OSCV-r (oscv to the right), STAB (stabilized), RPI (refined plug-in), SBG (smooth bootstrap with Gaussian kernel - the results refer to the equivalent bandwidth for the Epanechnikov kernel), Mix (mixed method for $\alpha = (1/2)$), and as a benchmark the ISE (infeasible ISE minimizing $\widehat{h}_0$).

Simulation results

In order to summarize the different methods of choosing the optimal bandwidth, we first consider the selected bandwidths and the corresponding biases for each method separately.

Afterwards, we compare the methods by various measures. The shown results are based on 250 simulation runs.

2.5.1. Comparison of the bias for the different bandwidths

In Figure 2.2 and 2.3 we illustrate the Bias (m_3) of the different methods for the mixture of three normal distributions varying sample size and distribution.

Let us consider the cross-validation method (CV). Many authors have mentioned the lack of stability of the CV-criterion and the tendency to undersmooth. In Figures 2.2 and 2.3 we see that CV has the smallest bias for all sample sizes and densities due to the fact that it chooses the smallest bandwidth. When the ISE optimal bandwidth is indeed very small, CV certainly does very well therefore. However, CV clearly undersmooths in the case of the simple normal distribution.

In contrast, the one-sided versions (OSCV) are more stable. Regarding the bias they are neither the best nor the worst in all sample sizes and models. As already stated by the authors, the OSCV tends to overestimate the bandwidth a little bit. While for $n = 25$

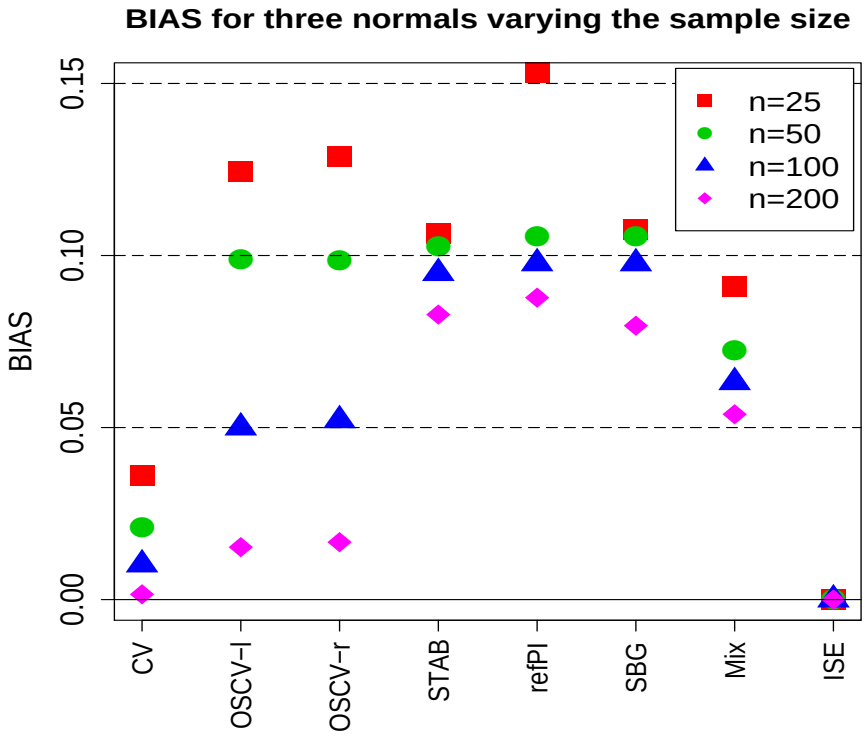

Figure 2.2.: *Comparison of the BIAS for different n for a mixture of three normals*

OSCV is outperformed by almost all other methods, this bias problem disappears rapidly for increasing n. In Figure 2.2 we see that their biases are much smaller than for the other methods except CV, and STAB in the simple normal case. Moreover, their behavior is quite stable and they do not fail as dramatically as the other methods in one or more cases. This feature is an intuitive benefit of this method when in practice the underlying density is completely unknown. For the densities studied, the differences between the left-(OSCV-l) and the right-sided (OSCV-r) versions are negligible except for the gamma distributions because of the boundary effect that is present on the left side.

The stabilized procedure of Chiu (STAB) is excellent for the simple normal case but it falls short when estimating rougher densities: "when the true density is not smooth enough, the stabilized procedure is more biased towards oversmoothing than CV" (Chiu ,1991). This fact can be seen in both Figures (2.2 and 2.3) where STAB has increasing difficulties with an increasing number of bumps. Even though this method demonstrates here a reasonable performance, the results should be interpreted with care, since in the derivation of Λ one has to deal with complex numbers, a problem we solved in favor of this method for this simulation study such that all presented results are slightly biased in favor of STAB.

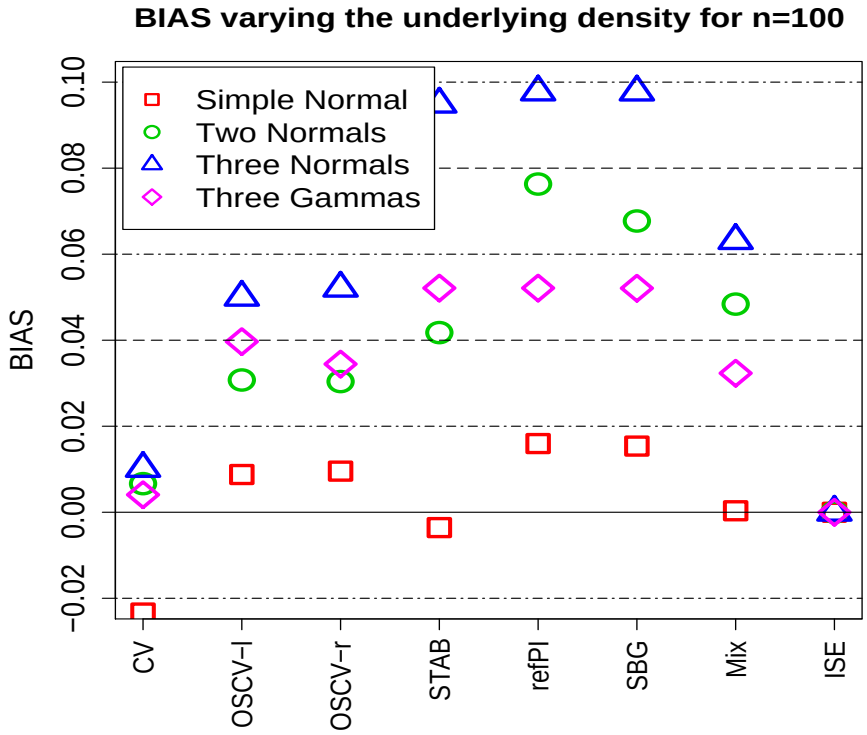

Figure 2.3.: *Comparison of the BIAS for different densities for a sample size of n=100.*

The refined plug-in (refPI) and the smoothed bootstrap SBG show a similar behavior as the stabilized procedure for $n = 100$, though the bias is much worse for refPI in small samples. Not surprisingly, in general, the bias for these MISE minimizing methods is larger than for all others. This partly results from the fact that we assume for the prior bandwidth that the second or third derivative comes from a simple normal distribution. Note that the bias of the SBG bandwidth is not as big as for the refPI.

The mixture of CV and plug-in is a compromise with biases lying between the ISE and the MISE minimizing methods. It will be interesting whether this leads also to a more stable performance (see below). Note that there is only a slight difference between the three versions of mixtures (not shown). Clearly, the larger the share of the respective method, the bigger their impact on the resulting estimate.

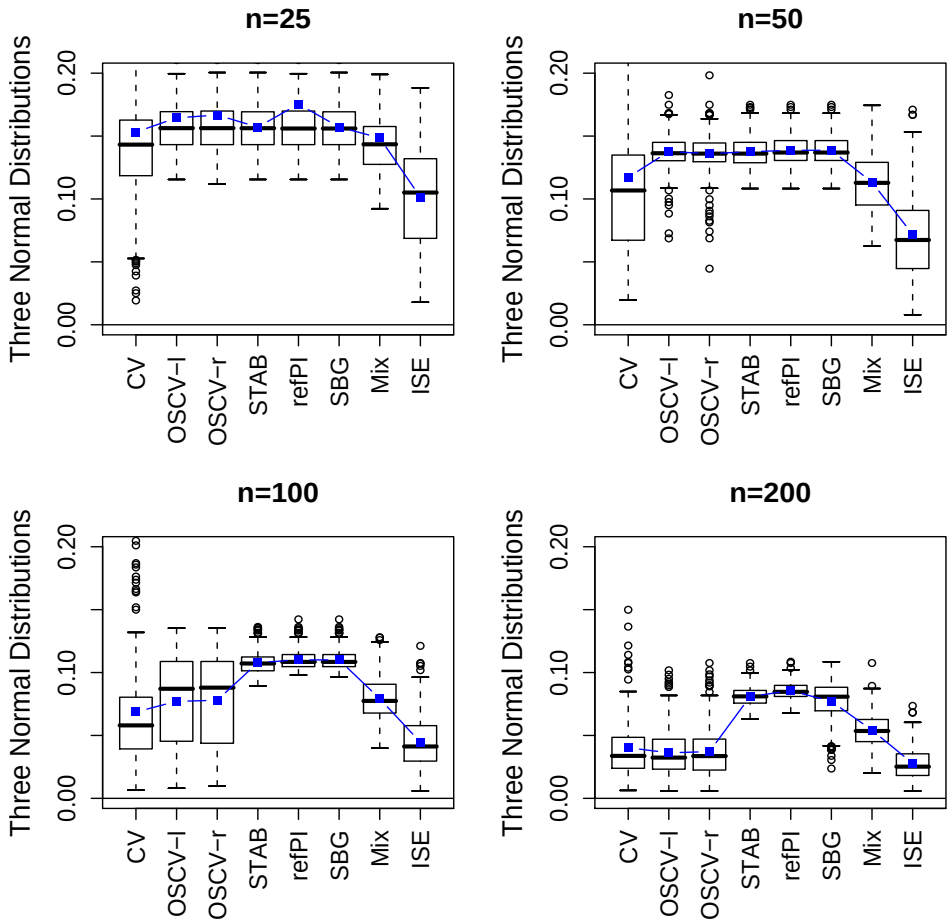

Figure 2.4.: *Box-plots and means (■) of the ISE-values for the mixture of three normal densities with different sample sizes.*

2.5.2. Comparison of the ISE-values

Next we compare the ISE-values of the density estimates based on the selected bandwidths. The results are given in form of boxplots plus the mean (linked filled squares) displaying the distribution of the ISEs after 250 simulation runs such that we get an idea of measures m_1 and m_2, m_4, and m_5 in one figure. In Figure 2.4 we consider the mixture of three normal distributions (model 3) and compare different sample sizes, whereas in figure 2.5 the sample size is fixed to $n = 100$ while the distribution varies.

Certainly, for all methods the ISE values increase with the complexity of the estimation problem. As expected, the classical CV-criterion shows a high variation for all cases (upper

extreme values not shown for the sake of presentation), doing somewhat better for more complex densities. The one-sided and the mixed versions do considerably better, though the least variation is achieved by the MISE minimizing methods (STAB, refPI and SBG). The drawback of these three methods becomes obvious when looking at the size of its ISE-values; they are clearly smaller for the CV-based methods for $n \geq 25$. Moreover, for increasing sample size their ISE values decrease very slowly whereas for the CV-methods these values come close to the optimal achievable ISE-values. Note that in the sense of minimizing the ISE, the one-sided and the Mix(1/2) versions show the best performance. They do not vary as much as the classical CV-criterion and their mean value is almost always smaller than for the other methods, see Figure 2.5.

The stabilized procedure of Chiu (STAB) delivers - as the name suggests - a very stable

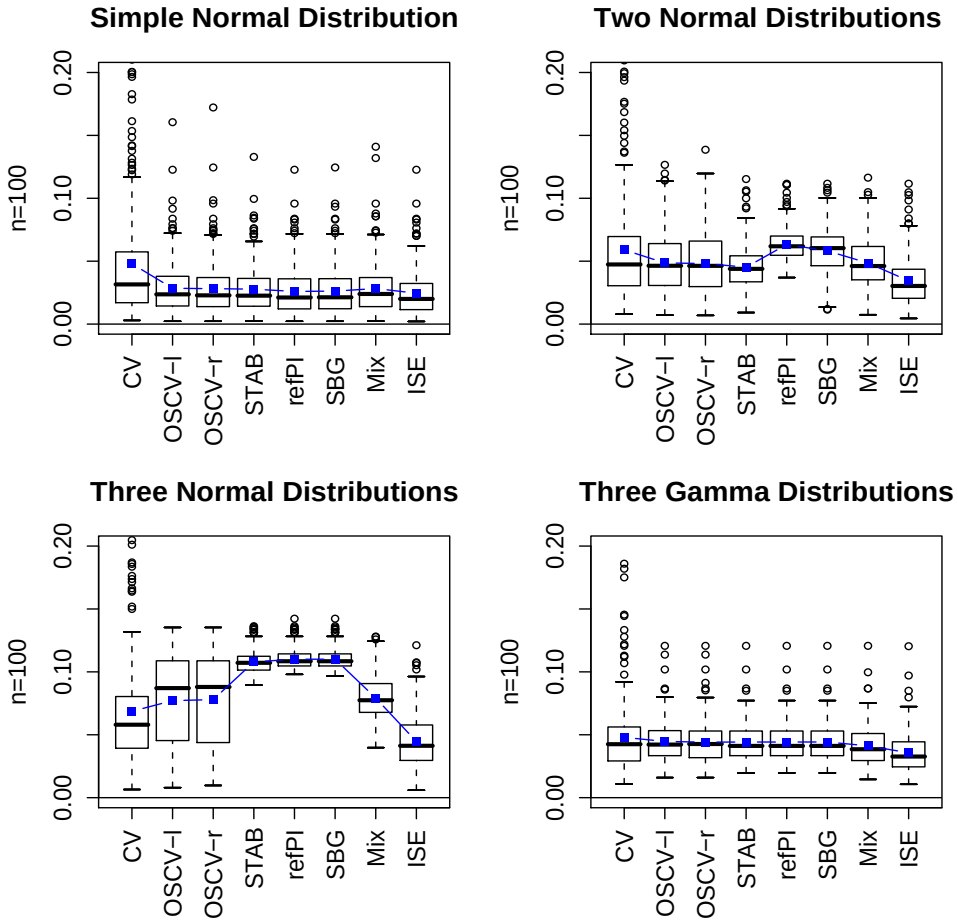

Figure 2.5.: *Box-plots and means (■) of the ISE-values for different distributions with sample size 100.*

estimate for the bandwidth. But in the end it is hardly more stable than the one-sided CV methods but much worse in the mean and median. We else see confirmed what we already discussed in the context of biases above. The mixture of CV and plug-in lowers the negative impacts of both versions and does surprisingly well; they deliver a more stable estimate, and gives good density estimates (looking at the ISE).

2.5.3. Comparison of the L1- and L2-distance of the ISE

To get an even better idea of the distance between the achieved ISE values of the selection methods and the ISE optimal (i.e. achievable) values, we have a closer look at m_5 and m_4, i.e. the L_1 and L_2 distances. In our opinion, these measures should be the most interesting for practitioners. Figure 2.6 and 2.7 show the L1-distance, and Figure 2.8 and 2.9 the L2-distance, respectively, for different sample sizes and models.

The pictures show that for CV, the m_5 are really big if the underlying density is not wiggly. This obviously is due to the the high variability of the selected bandwidths. Here, it does especially apply for small sample sizes (the value for $n = 25$ is even out of the range of the pictures); but for large samples like $n = 500$ the classical CV does not work at all (not shown). However, for the mixture of three normals the CV delivers almost the smallest m_5.

While both OSCV have problems with particularly small sample sizes, they easily compete with all other selectors. One may say that again, for the normal densities the OSCV methods are neither the best nor the worst methods, but always close to the best method. This corroborates our statement from above that the OSCV-criteria could be used if we do not know anything about the underlying density. Another conspicuous finding in Figure 2.7 is the difference between the two one-sided versions for the gamma distributions. Because of missing boundary correction on the left, the OSCV-l behaves very badly especially for a small sample size of $n = 25$ (out of the range) and $n = 50$. We get a similar result for $n = 25$ when looking at the L2-distances (out of the displayed range in Figure 2.9).

The three MISE minimizing methods do very well for the simple normal distribution, but else we observe a behavior for L1 and L2 which can be traced back to the fact of the prior problem described above. Even for bigger sample sizes all three methods deliver a relative big L1-distance for the mixture models. They further do not benefit as much from increasing n as other methods do. Within this MISE minimizing group, the STAB shows a better L1-distance for more complex densities. Actually, for the mixture of the three Gamma distributions we can see that the L1-distances are always very small, except for the refPI with $n = 25$ (out of the plotted range).

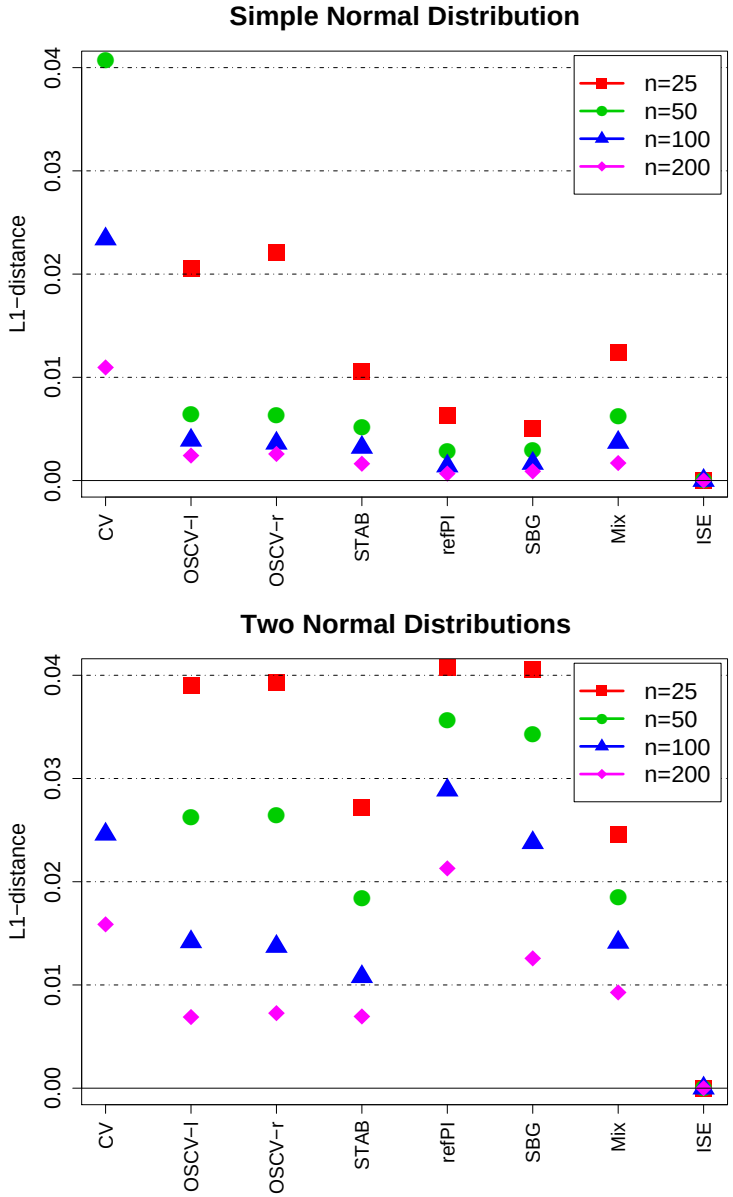

Figure 2.6.: *L1-distance for different sample sizes of simple normal distribution and mixture of two normal distributions.*

The mixture of CV and refined plug-in reflects the negative attributes of the CV, but for larger samples it is often in the range of the best methods. A further advantage of the mixed version is that it is much more stable than the CV or refPI when varying the sample size.

We obtain not the same but similar results for the L2-distance given in the Figures 2.8 and 2.9. We skipped the values for $n = 25$ because they were too big for most of the methods. CV obtains very large values for small sample sizes, so that they fall out of the range of

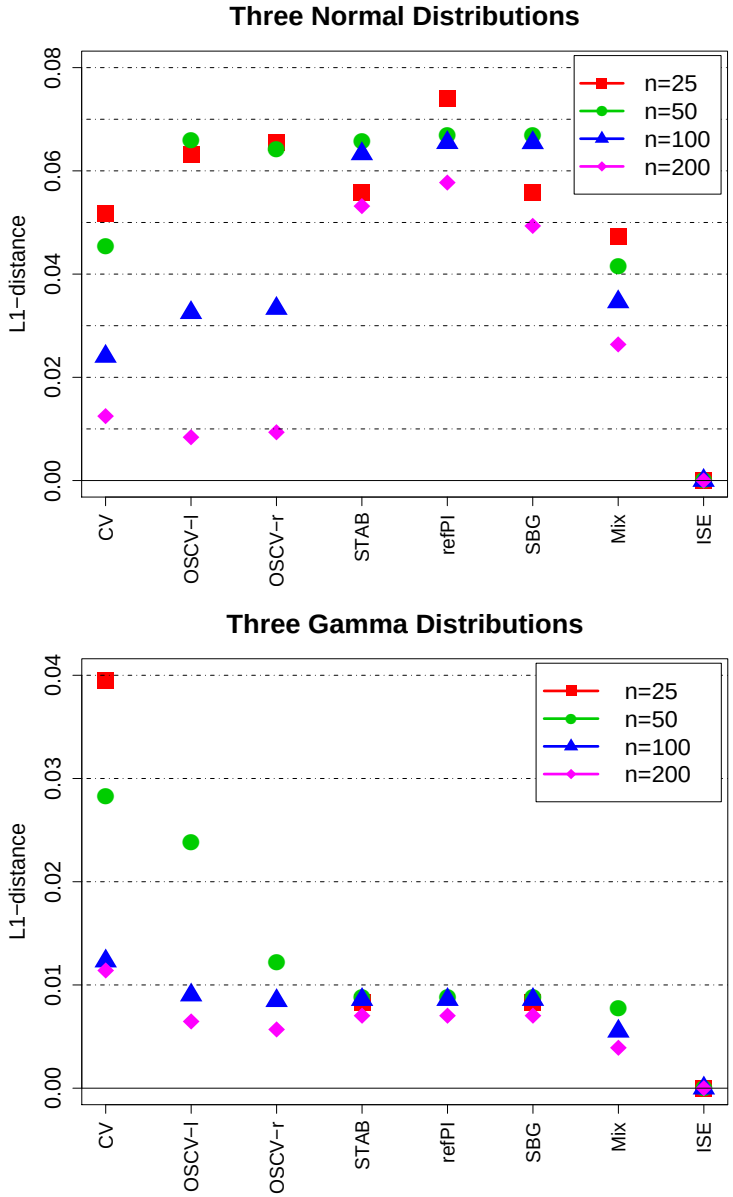

Figure 2.7.: *L1-distance for different sample sizes of mixture of three normal distributions resp. three gamma distributions.*

the pictures in many cases. The one-sided versions show an important improvement. The three MISE minimizing methods are excellent for the simple normal (not surprisingly) and the mixture of gammas. Among them, the STAB shows the smallest L2 distance. For sample sizes $n > 50$ the one sided CV versions outperform the others - for simple normal and gamma mixtures giving the same results as STAB but else having much smaller L2 distances. A huge difference between the left and the right one-sided occurs because of the boundary problem.

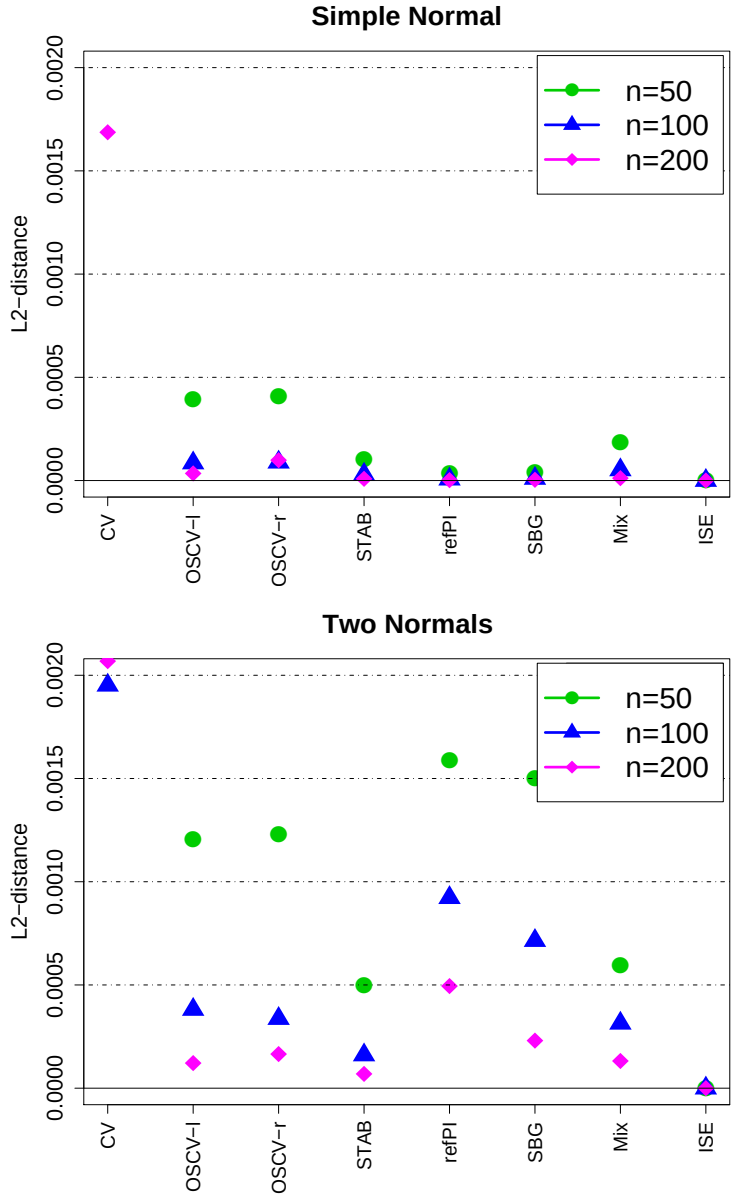

Figure 2.8.: *L2-distance for different sample sizes of simple normal distribution and mixture of two normal distributions.*

A comparison of the L1- and the L2-distance for $n = 100$ varying the distributions is shown in Figure 2.10. As can be seen in both pictures, the performance of all measures (without CV) for the simple normal distribution and the mixture of the three gamma distributions is pretty good. Also for the mixture of two normals most of the methods deliver good results, only the values for CV, refPI and the SBG become larger.

For more complex densities like the mixtures of three normals, the pictures show that the MISE minimizing measures deliver worse results, because of the large biases. The most

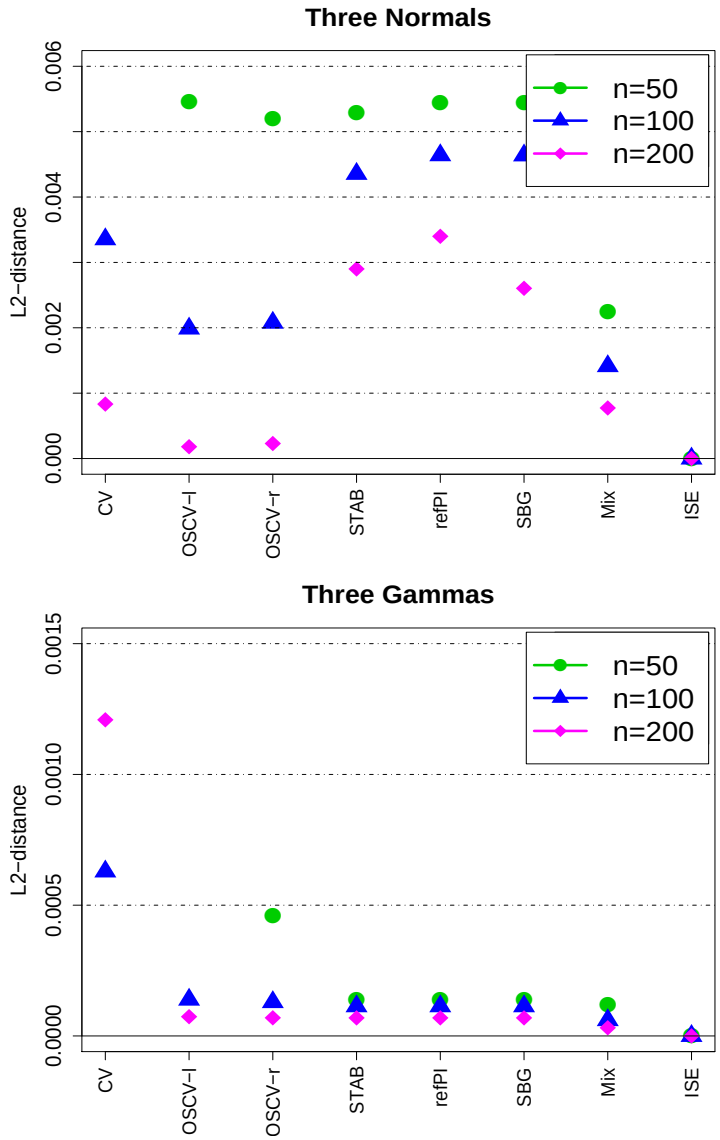

Figure 2.9.: *L2-distance for different sample sizes of mixture of three normal distributions resp. three gamma distributions.*

stable versions are the OSCV and the Mix(1/2). For smaller sample sizes (not shown) the pictures are quite similar, but the tendencies are strengthened and only the Mix(1/2) version delivers stable results for all distributions.

2.5.4. Comparison of the mixed methods

Finally we have a closer look to the quite promising results obtained by mixing CV with refPI. We did this in different proportions as described above. In Table 2.1 and Table 2.2

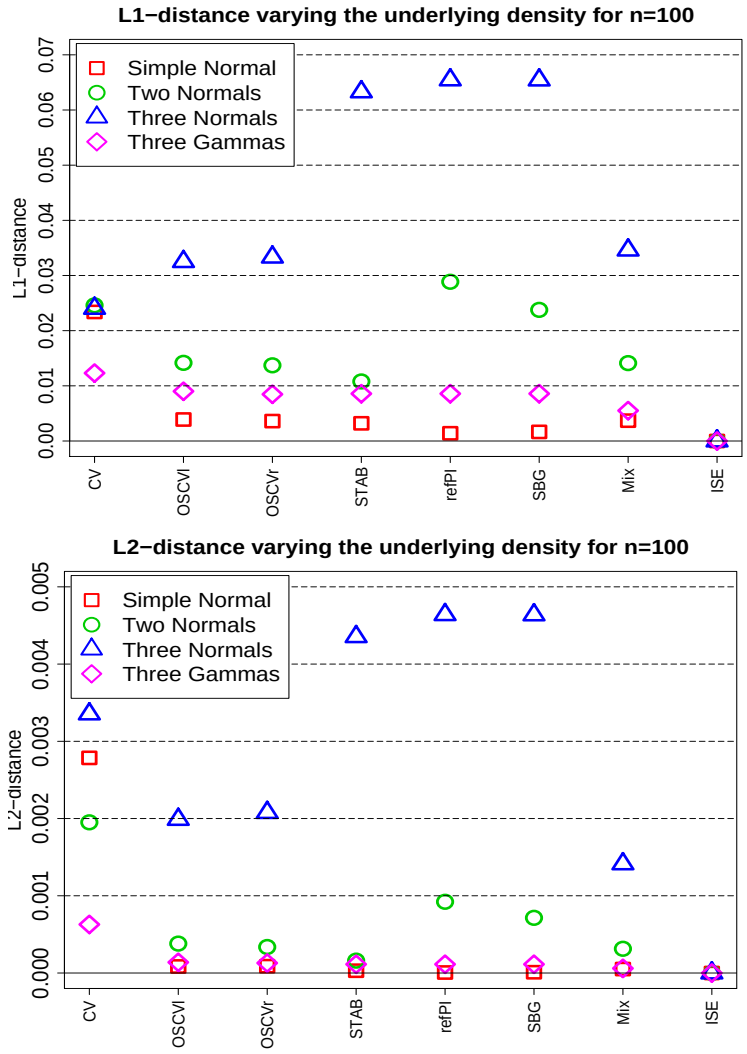

Figure 2.10.: *L1- and L2-distances for different underlying densities whit* $n = 100$

we have tabulated different measures looking at the ISE, the bias of the chosen bandwidth as well as the L1- and L2-distances for the four densities. We also give the values for the infeasible ISE-minimizing bandwidths.

For design 1 (simple normal density) in Table 2.1 the Mix(1/3) is the best. This is an expected result because we know from above that the refPI works very well for this distribution. The only measure in which this mix is not the best is the bias (m_3). The reason is that CV gives the smallest bias here. For design 2 (mixture of two normal densities) in Table 2.1 the Mix(2/3) wins except for the standard deviation of the ISE values (m_2) where Mix(1/3) is superior. This is explained by the very large sample variation typical

for CV.

For design 3 (trimodal distribution) in Table 2.2, Mix(2/3) does best except for the standard deviation of the ISE-values (m_2). This is not surprising because above we have always stated that for more complex distributions the CV works very well while the refPI performs poorly. For the mixture of the three gammas (design 4) we observe that the values of the different measures are nearly the same, especially for the L2-distance. The main differences occur for small sample sizes. The best is Mix(2/3). As we can see from the results, sometimes Mix(2/3) is the best and sometimes Mix(1/3). The Mix(1/2) lies in between. Consequently, the main conclusion is that the mix yields very stable results and is an attractive competitor to the other bandwidth selection methods.

		Design 1				Design 2			
n	Crit.	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)
25	m_1	.0605	.0802	.0699	.0730	.0810	.1017	.1104	.1055
	m_2	.0571	.0646	.0604	.0610	.0426	.0447	.0342	.0390
	m_3	0	-.0112	.0048	-.0018	0	.0459	.0663	.0576
	m_4	0	.0014	4e-04	6e-04	0	.001	.0013	.0011
	m_5	0	.0197	.0094	.0124	0	.0207	.0294	.0246
50	m_1	.0374	.0471	.0420	.0436	.0561	.0706	.0793	.0745
	m_2	.0298	.0365	.0320	.0333	.0325	.0338	.0265	.0303
	m_3	0	-.0050	.0083	.0029	0	.0385	.0613	.0519
	m_4	0	4e-04	1e-04	2e-04	0	5e-04	8e-04	6e-04
	m_5	0	.0097	.0046	.0062	0	.0146	.0233	.0185
100	m_1	.0246	.0307	.0271	.0282	.0344	.0452	.0525	.0485
	m_2	.0184	.0226	.0193	.0203	.0197	.0217	.018	.0199
	m_3	0	-.0070	.0056	3e-04	0	.0359	.0578	.0484
	m_4	0	1e-04	2e-05	5e-05	0	2e-04	4e-04	3e-04
	m_5	0	.0061	.0025	.0037	0	.0108	.0181	.0141
200	m_1	.0146	.0173	.0158	.0163	.0225	.0289	.0349	.0318
	m_2	.0106	.0127	.0113	.0117	.0135	.0148	.0135	.0143
	m_3	0	-.0028	.0055	.0021	0	.0283	.0491	.0404
	m_4	0	3e-05	5e-06	1e-05	0	1e-04	2e-04	1e-04
	m_5	0	.0027	.0012	.0017	0	.0064	.0124	.0093

Table 2.1.: *Values of the criteria m_1 to m_5 for mixed methods.*

2.6. Conclusions

This review and comparison study tries to give an idea of the state of the art in bandwidth selection for density estimation, about fifteen years after the last large reviews of Jones, Marron and Sheather (1996b) and Devroye (1997).

Though our review necessarily (otherwise it would not be a review but just an update) overlaps with them, we looked at different aspects, performance measures, included many more and to the best of our knowledge all new selection methods which have been proposed in the literature.

General findings about LSCV and classic plug-in methods are certainly the same as in other studies. For a better understanding of the performances, however, we looked at

		Design 3				Design 4			
n	Crit.	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)	ISE	MIX(2/3)	MIX(1/3)	MIX(1/2)
25	m_1	.1013	.1355	.1543	.1486	.0777	.088	.1326	.1331
	m_2	.0407	.0302	.0490	.0506	.0397	.0417	.1116	.1118
	m_3	0	.0666	.1003	.0909	0	.0163	.1987	.1922
	m_4	0	.0021	.0063	.0056	0	3e-04	.0141	.0141
	m_5	0	.0342	.0530	.0472	0	.0103	.0549	.0554
50	m_1	.0718	.1010	.1230	.1133	.0527	.0618	.0602	.0604
	m_2	.0342	.0309	.0165	.0224	.0245	.0266	.0233	.0243
	m_3	0	.0573	.0839	.0725	0	.0227	.0384	.0319
	m_4	0	.0014	.0032	.0022	0	2e-04	1e-04	1e-04
	m_5	0	.0292	.0512	.0415	0	.0091	.0075	.0077
100	m_1	.0446	.0667	.0898	.0792	.0357	.0409	.0419	.0412
	m_2	.0217	.0226	.0132	.0177	.0152	.0159	.015	.0153
	m_3	0	.0472	.075	.0632	0	.0234	.0387	.0323
	m_4	0	7e-04	.0023	.0014	0	6e-05	6e-05	6e-05
	m_5	0	.0221	.0452	.0346	0	.0052	.0062	.0055
200	m_1	.0278	.0432	.0642	.0542	.0245	.0280	.0291	.0284
	m_2	.0126	.0154	.0112	.0135	.0095	.0095	.0092	.0093
	m_3	0	.0387	.0654	.0539	0	.0204	.0365	.0297
	m_4	0	3e-04	.0014	8e-04	0	2e-05	3e-05	3e-05
	m_5	0	.0154	.0364	.0264	0	.0035	.0046	.0039

Table 2.2.: *Values of the criteria m_1 to m_5 for mixed methods.*

both L_1 and L_2 measures of the ISE, but also at the bias of the bandwidth estimate.

As well known, the CV leads to a small bias but large variance. It works well for rather wiggly densities and moderate sample size. However, it neither behaves well for rather small nor for rather large samples. The quality is unfortunately dominated by its variability. A fully automatic alternative is the one sided version. In contrast to the classical CV, the OSCV methods show a behavior which is very stable. Moreover, they are maybe not uniformly the best but quite often, and never the worst. We believe therefore that indirect CV with particular selection kernels, see also Savchuk, Hart, and Sheather (2010) is the most promising approach. Presently, the choice of the optimal selection kernel is a problem of prior knowledge which can heavily impact the final performance of the method.

This is similar for the mix-methods (combining CV and plug-in). While they show an excellent - maybe the best - behavior, one can certainly not identify a "best mix" in advance; this would require prior knowledge. A further evident computational disadvantage is that we first have to apply two other methods (CV and refPI) to achieve good results. Nevertheless, it is maybe a little bit surprising that simple mixtures have not been considered so far.

The refPI and the SBG show a, similar to OSCV, stable behavior due to the fact that they are minimizing the MISE and depend on prior information. The need of prior knowledge is the main disadvantage of these methods, and – as explained above – typically require a smooth underlying density. The worst case for these methods is when trying to estimate a trimodal normal density.

Also the STAB method is quite stable as suggested by its name. Although the full name refers to cross validation, it actually minimizes the MISE like refPI and SBG do. Consequently, it performs particularly well for the estimation of rather smooth densities but else not. The STAB method shows again the worst behavior for trimodal densities, indeed.

Our conclusion is therefore that among all existing (automatic) methods for kernel density estimation, to the best of our knowledge the OSCVs seem to outperform all competitors when no (or almost no) prior knowledge is available – maybe except the one about possible boundary problems. Depending on the boundary, one would apply left- or right-hand OSCV. For moderate sample sizes however, the mixture of CV and refPI seems to be an attractive alternative until n becomes large and CV fails completely.

2.7. References

Ahmad, I.A. and Ran, I.S. (2004). Data based bandwidth selection in kernel density estimation with paramteric start via kernel contrasts, *Journal of Nonparametric Statistics* **16**(6): 841-877.

Bean, S.J. and Tsokos, C.P. (1980). Developments in nonparametric density estimation, *International Statitstical Review* **48**: 267-287.

Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates, *Publications de l'Institut de Statistique de l'Université de Paris* **38(3)**: 3-59.

Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika* **71**: 353-360.

Cao, R. (1993). Bootstrapping the Mean Integrated Squared Error, *Journal of multivariate analysis* **45**: 137-160.

Cao. R.; Cuevas, A. and González Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation, *Computational Statistics and Data Analysis* **17**: 153-176.

Chacón, J.E.; Montanero, J. and Nogales, A.G. (2008). Bootstrap bandwidth selection using an h-dependent pilot bandwidth, *Scandinavian Journal of Statistics* **35**: 139-157.

Chaudhuri, P. and Marron, J.S. (1999). SiZer for Exploration of Structures in Curves, *Journal of the American Statistical Association* **94**(447): 807-823.

Chiu, S.T. (1991). Bandwidth Selection for Kernel Density Estimation, *The Annals of Statistics* **19**: 1883-1905.

Chiu, S.T. (1996). A comparative Review of Bandwidth Selection for Kernel Density Estimation, *Statistica Sinica* **6**: 129-145.

Devroye, L. (1989). The double kernel method in density estimation, *Ann. Inst. Henri Poincaré* **25**: 533-580.

Devroye, L. (1997). Universal smoothing factor selection in density estimation: theory and practice (with discussion), *Test* **6**: 223-320.

Duin, R.P.W. (1976). On the choice of smoothing parameters of Parzen estimators of probability density functions, *IEEE Transactions on Computers* **25**: 1175-1179.

Faraway, J.J. and Jhun, M. (1990). Bootstrap Choice of Bandwidth for Density Estimation, *Journal of the American Statistical Association* **85/412**: 1119-1122.

Feluch, W. and Koronacki, J. (1992). A note on modified cross-validation in density estimation, *Computational Statistics & Data Analysis* **13**: 143-151.

Fryer, M.J. (1977). A review of some non-parametric methods of density estimation, *Journal of Applied Mathematics* **20**(3): 335-354.

Godtliebsen, F.; Marron, J.S. and Chaudhuri, P. (2002). Significance in Scale Space for Bivariate Density Estimation, *Journal of Computational and Graphical Statistics* **11**: 1-21.

Grund, B. and Polzehl, J. (1997). Bias corrected bootstrap bandwidth selection, *Journal of nonparametric statistics* **8**: 97-126.

Habbema, J.D.F.; Hermans, J. and van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation, in: Bruckman, G. (Ed.), *COMPSTAT '74. Proceedings in Computational Statistics,* Physica, Vienna: 101-110.

Hall, P. (1990). Using the bootstrap to estimate mean square error and select smoothing parameters in nonparametric problems, *Journal of Multivariate Analysis* **32**: 177-203.

Hall, P. and Johnstone, I. (1992). Empirical Functionals and Efficient Smoothing Parameter Selection, *Journal of the Royal Statistical Socienty B* **54** (2): 475-530.

Hall, P. and Marron, J.S. (1987a). Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation, *Probability Theory and Related Fields* **74**: 567-581.

Hall, P. and Marron, J.S. (1987b). Estimation of integrated squared density derivatives, *Statistics & Probability Letters* **6**: 109-115.

Hall, P. and Marron, J.S. (1991). Lower bounds for bandwidth selection in density estimation, *Probability Theory and Related Fields* **90**: 149-173.

Hall, P.; Marron, J.S. and Park, B.U. (1992). Smoothed cross-validation, *Probability Theory and Related Fields* **92**: 1-20.

Hall, P.; Sheater, S.J.; Jones, M.C. and Marron, J.S. (1991). On optimal databased bandwidth selection in kernel density estimation, *Biometrika* **78**: 263-269.

Hanning, J. and Marron, J.S. (2006). Advanced Distribution Theory for SiZer, *Journal of the American Statistical Association* **101**: 484-499.

Härdle, W.; Müller, M.; Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models, *Springer Series in Statistics*, Berlin.

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series, *Journal of Time Series Analysis* **13**: 209-232.

Hart, J.D. and Yi, S. (1998). One-sided cross-validation, *Journal of the American Statistical Association* **93**: 620-631.

Jones, M.C. (1991). The roles of ISE and MISE in density estimation, *Statistics & Probability Letters* **12**: 51-56.

Jones, M.C. (1998). On some kernel density estimation bandwidth selectors related to the double kernel method, *Sankhyā Ser. A* **60**: 249-264.

Jones, M.C., Marron, J.S. and Park, B.U. (1991). Asimple root n bandwidth selector, *The annals of statistics* **19**(4): 1919-1932.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996a). A brief survey of bandwidth selection for density estimation, *Journal of the American Statistical Association* **91**: 401-407.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996b). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics* **11**: 337-381.

Jones, M.C. and Sheather, S.J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statistics & Probability Letters* **11**: 511-514.

Kim, W.C., Park, B.U. and Marron, J.S. (1994). Asymptotically best bandwidth selectors in kernel density estimation, *Statistics & Probability Letters* **19**: 119-127.

Loader, C.R. (1999). Bandwidth selection: Classical or Plug-In?, *The annals of statistics* **27**(2): 415-438.

Marron, J.S. (1986). Convergence properties of an empirical error criterion for multivariate density estimation, *Journal of Multivariate Analysis* **19**: 1-13.

Marron, J.S. (1988a). Automatic Smoothing parameter selection: A survey, *Empirical Economics* **13**: 187-208.

Marron, J.S. (1988b). Partitioned cross-validation, *Economic Reviews* **6**: 271-283.

Marron, J.S. (1992). Bootstrap Bandwidth Selection, in: *Exploring the limits of bootstrap, eds. R. LePage and L. Billard, Wiley, New York*: 249-262.

Marron, J.S. (1994). Visual understanding of higher order kernels, *Journal of Computational and Graphical Statistics* **3**: 447-458.

Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation, *Statistics and Probability Letters* **7(3)**: 195-199.

Marron, J.S. and Wand, M.P. (1992). Exact mean integrated squared errors, *Annals of statistics* **20**: 712-736.

Martínez-Miranda, M.D.; Nielsen, J. and Sperlich, S. (2009). *One sided Cross Validation in density estimation*, In "Operational Risk Towards Basel III: Best Practices and Issues in Modeling, Management and Regulation", ed. G.N.Gregoriou; John Wiley and Sons, Hoboken, New Jersey, 177-196.

Park, B.U. and Marron, J.S. (1990). Comparison of Data-Driven Bandwidth Selectors, *Journal of the American Statistical Association* **85**: 66-72.

Park, B.U. and Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors, *CORE Discussion Paper* **9205**.

Rigollet, P. and Tsybakov, A. (2007). Linear and convex aggregation of density estimators, *Mathematical Methods of Statistics* **16**: 260-280.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics* **9**: 65-78.

Ruppert, D. and B.H. Cline, B.H. (1994). Bias Reduction in Kernel Density Estimation by Smoothed Empirical Transformations, *The Annals of Statistics* **22**: 185-210.

Samarov, A. and Tsybakov, A. (2007). Aggregation of density estimators and dimension reduction, In:*Advances in Statistical Modeling and Inference: essays in honor of Kjell A. Doksum*, ed. V. Nair,233-251.

Savchuk, O.J., Hart, J.D. and Sheather, S.J. (2010). Indirect Cross-Validation for Density Estimation, *Journal of the American Statistical Association* **105(489)**: 415-423.

Scott, D.W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association* **82**(400): 1131-1146.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B* **53**: 683-690.

Silverman, B.W. (1986). Density estimation for Statistics and Data Analysis, Vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics* **12**(4): 1285-1297.

Stute, W. (1992). Modified cross validation in density estimation, *Journal of statistical planning and Inferenz* **30**: 293-305.

Tartar, M.E. and Kronmal, R.A. (1976). An introduction to the implementation and theory of nonparametric density estimation, *The American Statistician* **30**: 105-112.

Taylor, C.C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation,
Biometrika **76**: 705-712.

Turlach, B.A. (1994). Bandwidth selection in Kernel density estimation: A Review, *Working Paper*.

Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing, Vol. 60 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Wand, M.P.; Marron, J.S. and Ruppert, D. (1991). Transformations in Density Estimation, *Journal of the American Statistical Association* **Vol. 86**, No. 414: 343-353.

Wegkamp, M.H. (1999). Quasi universal bandwidth selection for kernel density estimators, *Canadian Journal of Statistics* **27**: 409-420.

Wegman, E.J.(1972). Nonparametric probability density estimation: I. A summary of available methods, *Technometrics* **14**: 533-546.

Wertz, W. and Schneider, B. (1979). Statistical density estimation: a bibliography, *International Statistical Review* **47**: 155-175.

Yang, Y. (2000). Mixing strategies for density estimation, *Annals of Statistics* **28(1)**: 75-87.

Yang, L. and Marron, S. (1999). Iterated Transformation-Kernel Density Estimation. *Journal of the American Statistical Association*, **94(446)**: 580-589.

3. Semiparametric voter profiling in a multi-party system – new insights via flexible modeling

Abstract

Due to the German reunification and the preceding political process both the composition of the electorate and the list of influential parties has changed substantially. Nowadays, classic multinomial logit models seem to be no longer sufficient – if they ever were – for analyzing voter profiles. In a more complex world with frequent structural changes of societies, the outcomes of these models simplify too much and partly contradict general beliefs. We develop and provide a smoothed likelihood estimator that allows for flexible functional forms and interactions between covariates. It reveals strong interactions of age and income, as well as highly nonlinear and rather different shapes of the factor impacts for each party's likelihood to be voted.

This chapter is the result of a collaborative project with my colleague Roland Langrock and Prof. Dr. Stefan Sperlich. The main contribution of the author of this thesis is made in the development of the semiparametric approach and the implementation of all considered models in R. The presentation as well as interpretation of the results is also the task of the author of this thesis.

3.1. Introduction and Motivation

The multinomial logit model (MNL) allows to investigate the influence of a vector of covariates on more than two possibly unordered outcomes of categorical response variables. We introduce a semiparametric extension via kernel smoothing providing a profile likelihood algorithm. We understand this mainly as an explorative tool, maybe even to find an appropriate parametric specification. By means of two- and three-dimensional plots we achieve comprehensive insights into the data structure. We apply our model to study political party affiliation in Germany and to identify various electorate profiles for the different political parties. Such informations are of great use for policy makers and analysts - not only to design their campaigns for targeted voter groups.

The MNL has become popular in econometrics by the work on brand choice behaviour by McFadden (1974) and on urban travel demand by Domencich and McFadden (1975), respectively. Since then the model has been used in a wide field of applications, but still especially in studies of consumer behaviour. Different trials were undertaken to include nonlinear effects of the explanatory variables. Krishnamurthi and Raj (1988) used logarithmic transformations. Ben-Akiva and Lerman (1985) as well as Kalyanaram and Little (1994) proposed piecewise linear (utility) functions on predetermined (sub)intervals. More recently, some authors developed nonparametric and semiparametric methods for these kind of data. Yee and Wild (1996) considered a backfitting algorithm on a class of multivariate additive models using smoothing splines. Abe (1998, 1999) proposed a special class of generalized additive models which accommodates to a multinomial qualitative response to study consumer demand. His algorithm is based on a penalized likelihood function and modified local scoring (Hastie and Tibshirani, 1986). Tutz and Scholz (2004) approximate unspecified additive functions by a finite number of basis functions which are penalized with respect to their localization. Kneib, Baumgartner and Steiner (2007) modified this using penalized B-splines and a Bayesian approach for their estimation, again for studying consumer choices.

In this work, we prefer to stick to a likelihood specification but localize it via kernels. More specific, we do profile likelihood estimation in the spirit of Severini and Wong (1992). Statistical inference on the parameters can therefore be derived from the marginal Fisher information. For doing inference on the nonparametric part one can apply a (semi-)parametric bootstrap along Härdle et al. (2004a). We allow for an additive structure of the nonparametric part but will first look at the multivariate impact function as we believe interaction plays an important role in political affiliations. The model includes individual-specific and mode-specific variables, whereas the former represent characteristics of the observed individuals and the latter properties of the alternatives (political parties in our case). The

impact of all mode- and individual-specific variables will be estimated nonparametrically in a first step. Further decomposition can be done in consequent steps.

Our main interest is in profiling specific voter groups with respect to age, income and gender. Nowadays, the German party system comprises five main political parties: the Christian Democratic Union (CDU) respectively its Bavarian counterpart (the Christian Social Union, CSU but for simplicity subsumed with the CDU in this paper), the Social Democratic Party (SPD), the Free Democratic Party (FDP), the Left Party (LP) and the Alliance '90/The Greens (A90G). Some authors raise the question whether it is appropriate to speak of only one party system (e.g. Roberts, 1997) as the electoral behaviour in Eastern and Western Germany is quite different. While in the East three parties, namely CDU, SPD and LP contend with each other for the leading position, in Western Germany only the two big parties (CDU and SPD) have had enough authority to form ruling coalitions with one of the three smaller parties. This dichotomy is founded in the historical development, see Section 3.3. Nonetheless, after having controlled for the location, i.e. East or West, of the voter, one might expect that on average left parties are supported by the lower (working) class, the green movement by young people from the middle class, the Christian democrats by elderly population, the liberals by the richest, etc. A linear MNL might provide this information and therefore confirm what is known anyway. However, as it only gives back the average linear effects, it does not provide any insight into the details and does therefore not allow for further conclusions. In contrast, we will see that our semiparametric models provide well interpretable estimates.

Section 3.2 describes the general though theoretical model specification, and an appropriate estimation procedure. In Section 3.3 we will introduce the data, give some discussion of the parties' historical background and provide the classic MNL estimates. Section 3.4 is dedicated to the detailed semiparametric analysis applying our procedure from Section 3.2. We conclude in Section 3.5.

3.2. Model and Estimation

Consider a semiparametric multinomial logit model with K different outcome categories that have no natural order. The conditional probability of outcome $Y = k$, $k = 1, \ldots, K$, given the individual covariate vectors $\mathbf{X} = (X_1, \ldots, X_p)^t \in \mathbb{R}^p$ and $\mathbf{T} = (T_1, \ldots, T_q)^t \in \mathbb{R}^q$ is assumed to be given by

$$\mathbb{P}(Y = k \,|\, \mathbf{X}, \mathbf{T}) = \frac{\exp\left(\mathbf{X}^t \boldsymbol{\beta}_k + m_k(\mathbf{T})\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{X}^t \boldsymbol{\beta}_j + m_j(\mathbf{T})\right)}\,. \tag{3.1}$$

We set for identification $\boldsymbol{\beta}_K = \mathbf{0}$, and $m_K(\cdot) \equiv 0$, i.e. K is the reference mode. Each $m_k(\cdot)$, $k = 1, \ldots, K$, is assumed to be a smooth function with domain $\mathbb{R}^q$ and each $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^t$, $k = 1, \ldots, K-1$, denotes an unknown parameter vector. Variables that depend on both modes and individuals could be considered as well. Note that the nonparametric functions m_k also capture any mode-specific effect, see further discussion below. So $\mathbf{X}$ must not contain mode-specific dummies.

If the functions $m_k(\cdot)$ were known it would be easy to find estimators for the vectors $\boldsymbol{\beta}_k$, and vice versa. Following the ideas of profiled likelihood by Severini and Wong (1992), the functions $m_k(\cdot)$ are regarded as nuisance when estimating the finite-dimensional parameters $\boldsymbol{\beta}_k$. The functions $m_k(\cdot)$ themselves can be estimated via kernel smoothing. Note that the estimate of $m_{k,\boldsymbol{\beta}.}(\cdot)$ will depend on all $\boldsymbol{\beta}_j$, $j = 1, \ldots, K-1$, indicated by the index '$\boldsymbol{\beta}.$'. This yields asymptotically normal, $\sqrt{n}$-consistent and efficient estimators for the vectors $\boldsymbol{\beta}_k$ owing to likelihood estimation. For the m_k one obtains consistent estimators with statistical properties typical for nonparametric kernel smoothing, see also Rodríguez-Póo et al. (2003).

In order to estimate the so-called least favorable curve $m_{k,\boldsymbol{\beta}.}(t)$ at point $\mathbf{t} := (t_1, \ldots, t_q)$ for given $\boldsymbol{\beta}_k$, $k = 1, \ldots, K-1$, take a q-dimensional kernel $K : \mathbb{R}^q \to \mathbb{R}$, bandwidth matrix $\mathbf{H} \in \mathbb{R}_+^{q \times q}$, and consider the local likelihood

$$\begin{aligned} \mathcal{L}_s(m_{k,\boldsymbol{\beta}.}(\mathbf{t})) &= \sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t} - \mathbf{t}_i)\big) \mathcal{L}(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t})), y_i), \qquad (3.2) \\ \text{with} \quad \boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t})) &:= (\eta_{1i}, \ldots, \eta_{ki}(m_{k,\boldsymbol{\beta}.}(\mathbf{t})), \ldots, \eta_{Ki}), \\ \text{where} \quad \eta_{ki}(m_{k,\boldsymbol{\beta}.}(\mathbf{t})) &:= \mathbf{x}_i^t \boldsymbol{\beta}_k + m_{k,\boldsymbol{\beta}.}(\mathbf{t}) \\ \text{and} \quad \eta_{ji} &:= \mathbf{x}_i^t \boldsymbol{\beta}_j + m_{j,\boldsymbol{\beta}.}(\mathbf{t}_i) \quad \text{for } j \neq k, \end{aligned}$$

with $\mathbf{x}_i^t = (x_{i1}, \ldots, x_{ip})$, $\mathbf{t}_i^t = (t_{i1}, \ldots, t_{iq})$. Here, $\mathcal{L}(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t})), y_i)$ denotes the log-likelihood of (3.1) of the ith observation with predictor $\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t}))$ wherein $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ and $m_{j,\boldsymbol{\beta}.}(\mathbf{t}_i)$ for $j \neq k$ are treated as fixed such that $\boldsymbol{\eta}_i$ is only a function of $m_{k,\boldsymbol{\beta}.}$ in (3.2).

With $\widehat{m}_{k,\boldsymbol{\beta}.}(\cdot)$ at hand, we can compute the profile likelihood

$$\begin{aligned} \mathcal{L}_p(\boldsymbol{\beta}_k) &= \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k), y_i), \qquad (3.3) \\ \text{where now} \quad \boldsymbol{\eta}_i(\boldsymbol{\beta}_k) &:= (\eta_{1i}, \ldots, \eta_{ki}(\boldsymbol{\beta}_k), \ldots, \eta_{Ki}), \\ \text{with} \quad \eta_{ki}(\boldsymbol{\beta}_k) &:= \mathbf{x}_i^t \boldsymbol{\beta}_k + m_{k,\boldsymbol{\beta}.}(\mathbf{t}_i) \end{aligned}$$

and η_{ji}, $j \neq k$, as before. Notice that, in (3.3), $\boldsymbol{\eta}_i(\cdot)$ is a function of $\boldsymbol{\beta}_k$.

In order to understand the estimation procedure we need the first two derivatives of

$l_i(\boldsymbol{\eta}) := \mathcal{L}(\boldsymbol{\eta}, y_i)$ with respect to η_k, $\eta_k = \mathbf{x}_i^t \boldsymbol{\beta}_k + m_k(\mathbf{t}_i)$. First, note that

$$l_i(\boldsymbol{\eta}) = \sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \eta_{ki} - \log \sum_{j=1}^{K} \exp\left(\eta_{ji}\right) . \tag{3.4}$$

Then, it follows immediately that

$$\begin{aligned} l'_{ik}(\boldsymbol{\eta}) &= \mathbb{1}_{\{y_i=k\}} - \frac{\exp(\eta_{ki})}{\sum_{j=1}^{K} \exp(\eta_{ji})} \\ l''_{ik}(\boldsymbol{\eta}) &= -\frac{\exp(\eta_{ki}) \cdot \sum_{j=1}^{K} \exp(\eta_{ji}) - \exp(\eta_{ki})^2}{\sum_{j=1}^{K} \exp(\eta_{ji})} . \end{aligned}$$

To obtain the maximum of the smoothed likelihood $\mathcal{L}_s(m_{k,\boldsymbol{\beta}.}(\mathbf{t}))$, successively from mode 1 to mode K, we have to solve the first order condition

$$\sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t} - \mathbf{t}_i)\big) l'_{ik}\big(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t}))\big) = 0 \tag{3.5}$$

with respect to $m_{k,\boldsymbol{\beta}.}(\mathbf{t})$. For $\boldsymbol{\beta}_k$ the equation system to solve is

$$\sum_{i=1}^{n} l'_{ik}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k))(\mathbf{x}_i + m'_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)) = \mathbf{0} , \tag{3.6}$$

wherein $m'_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)$ denotes the gradient of $m_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)$ with respect to $\boldsymbol{\beta}_k$. By deriving equation (3.5) with respect to $\boldsymbol{\beta}_k$ one obtains

$$m'_{k,\boldsymbol{\beta}.}(\mathbf{t}) = \frac{\sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t} - \mathbf{t}_i)\big) l''_{ik}\big(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t}))\big) \mathbf{x}_i}{\sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t} - \mathbf{t}_i)\big) l''_{ik}\big(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t}))\big)} . \tag{3.7}$$

Equations (3.5) to (3.7) can be used to implement a Newton-Raphson-type algorithm:

1. *Find appropriate starting values* $\boldsymbol{\beta}_k^{(0)}$, $m_k^{(0)}(\cdot)$, $k = 1, \ldots, K-1$ *(e.g. by fitting an appropriate parametric MNL) and set* $j = 0$.

2. *For* $k = 1, 2, \ldots, K-1$, *compute*

$$\begin{aligned} \boldsymbol{\beta}_k^{(j+1)} &= \boldsymbol{\beta}_k^{(j)} - \mathcal{B}^{-1} \sum_{i=1}^{n} l'_{ik}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k^{(j)}))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)) \\ \textit{with} \qquad \mathcal{B} &= \sum_{i=1}^{n} l''_{ik}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k^{(j)}))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i))^t \end{aligned}$$

and $m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)$ *as in (3.7).*

3. *For* $k = 1, 2, \ldots, K-1$, *compute*

$$m_{k,\boldsymbol{\beta}\cdot}^{(j+1)}(\mathbf{t}) = m_{k,\boldsymbol{\beta}\cdot}^{(j)}(\mathbf{t}) - \frac{\sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big) l'_{ik}\big(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}\cdot}^{(j)}(\mathbf{t}))\big)}{\sum_{i=1}^{n} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big) l''_{ik}\big(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}\cdot}^{(j)}(\mathbf{t}))\big)}$$

for all points $\mathbf{t}$ *at which the function* $m_{k,\boldsymbol{\beta}\cdot}(\cdot)$ *is to be estimated.*

4. *Repeat steps* **2.**–**3.** *for* $j = 1, 2, \ldots$ *until convergence.*

It is convenient to estimate the functions $m_{k,\boldsymbol{\beta}\cdot}(\cdot)$ in step **3** at the observation points $\mathbf{t}_i$, $i = 1, \ldots, n$, as this guarantees that independent of the bandwidth choice at least for one observation $K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big)$ is nonzero. As described for example in Härdle et al. (2004b) these steps simplify slightly if a Speckman-type algorithm is applied. We implemented both versions of the algorithm and obtained almost the same results.

Recall that the m_k will automatically capture any mode-specific effect. If there is also a vector of mode characteristics $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{rk})^t \in \mathbb{R}^r$ available, then one might want to regress them on the m_k. Note that a nonparametric modeling of their influence would not make much sense as the support is discrete consisting of few values. Therefore, the influence of the mode-specific covariate vector should be modeled by a simple linear relation with unknown parameter $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_r)^t$ resulting in an optional fifth step:

5. *After convergence of* $\boldsymbol{\beta}_k$ *and* $m_{k,\boldsymbol{\beta}\cdot}$, *perform an additional regression*

$$\sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot m_{k,\boldsymbol{\beta}\cdot} = \gamma_0 + \sum_{j=1}^{r} \gamma_j \cdot \sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot Z_{jk}\,.$$

3.3. Data and Prior Parametric Approach to the Political Party Affiliation Data

The aim is to understand the determinants of voter's choice, and thereby to identify typical voter groups of the dominant political parties in the multi-party system of Germany. Let us begin by briefly sketching the historical background of the German party system. After the second world war, Western Germany was governed either by the CDU or the SPD, with absolute majority or in a coalition with the FDP, respectively. In the seventies diverse groups of alternative green activists contested at various local elections. In 1980 a green organization was confounded at federal level. It was composed of distinct groups like the anti-nuclear movement, the student-movement, feminist groups and the peace movement (for details see Lösche, 1993). They won the first seats in the German Bundestag in 1983 and from 1998 to 2005 they joined the federal government in a coalition with the SPD. In

the East German dictatorship the Socialist Unity Party (SED) had sole political power, although small and well-controlled Christian and liberal parties co-existed to give the system a semblance of legitimacy. After reunification the so-called PDS was confounded as the heir of the SED. In 2005, the PDS entered an alliance with the just founded West-German party "Labour and Social Justice - Electoral Alternative (WASG)". Since 2007 the alliance is simply called "The Left" (LP in the following). It achieved 8.7% in the 2005 election to the German Bundestag (26% in East Germany).

The data for the upcoming analysis was taken from the German Socio-Economic Panel (SOEP) of the year 2006. The variable of interest is political party affiliation, i.e. the answer to the question "Toward which party do you lean?". In accordance with the standard literature, the socio-economic factors that are taken into account are age, log-income (monthly net total household income), region, and gender of the voter (cf. Quinn et al. 1999, Dow and Enderby, 2004). For a couple of reasons, we have not included the elsewhere often considered covariates education and religion. In fact, the reported years of education as well as vocational qualifications are hardly comparable between Eastern and Western Germany. This is actually also true for reported religion; while in Western Germany, the majority of the people officially still belongs to either the protestant or Roman catholic church no matter if they are practitioners or not, in Eastern Germany Christians form rather an avowed minority as in the socialist system the affiliation to a church could easily entail serious negative consequences for the family. Concerning the region we included a dummy 'east', indicating whether a person was resided in Eastern or Western Germany before reunification. This way we especially account for several of the aspects discussed above.

All together, 8787 people reported their party affiliation in the original data set. Out of them, 376 favored a different party than the here considered ones, 227 lived abroad before reunification or did not report their regional provenience, and 383 persons made no, 32 an implausible declaration about their income. Note that contrary to parametric methods, for local smoothers like our kernel approach, a trimming of the covariates that enter nonparametrically has no impact on the final outcome, except if the bandwidth is chosen data adaptively. So our semiparametric estimator will not suffer from a selection bias after such a trimming. We restrict the data to a monthly income of max. 10 000 €, and people being younger than 65, the official retirement age. For the interpretation this means we concentrate on a more homogeneous, mostly professionally active group (about 75% of the people) but without top earners (109 out of 8787, i.e. 1.2%). This has the additional advantage that our kernel smoother will not suffer from data sparse areas due to extreme tails in the covariates' distribution. The resulting data set consists of $n = 5343$ observations with the descriptive statistics summarized in Table 3.1 and 3.2. As expected,

income is strongly skewed to the right whereas age is somewhat skewed to the left. Insiders may wonder why in Table 3.2 the affiliation to CDU in 2006 seems to be smaller than for the SPD. Note that this changes when one re-includes people being older than 65.

Variable			Yes - 1	No - 0	Yes (in %)	No (in %)
X_1	Gender	1 if female	2 630	2 713	49.22	50.78
X_2	Region	1 if East	1 165	4 178	21.80	78.20
			Min	Max	Mean	Median
T_1	Income	Euro/Month	400	10 000	3 223	3 000
T_2	Age	Years	21	64	45.63	47.00

Table 3.1.: *Descriptive statistics for the considered covariates.*

Political Party		CDU	SPD	A90G	LP	FDP
Affiliation	(in %)	37.66	38.84	11.87	6.42	5.22

Table 3.2.: *Percentages of reported political affiliation.*

Principally, there are two ways to analyze voter groups: *(i)* using purely descriptive statistics based on public-opinion polls, as routinely published by market research institutes such as Infratest dimap or Forsa in Germany, and *(ii)* employing inferential statistics by fitting adequate models. When it comes to voter profiling, one of the main drawbacks of *(i)* is that the distribution of the voters choice can not be quantified based on statistical laws and hence can not be used to support inferential statements about the population. Furthermore, such analyses typically focus on only one or two covariates at a time. Models such as the multinomial logit attempt to overcome this deficiency by modeling the voter's party affiliation as outcome of a distribution that depends on a number of covariates. However, the multinomial logit and similar parametric models have limitations as well: they are based on assumptions concerning the specific functional form that links the covariates to the outcome. Another limitation is the additive separability and the implied neglect of possible interactions between different covariates. The proposed semiparametric model for multicategorical data attempts to overcome those deficiencies. These aspects will be studied in detail in the course of the subsequent section.

Before starting with the semiparametric analysis, we consider a fully parametric MNL. The main purpose of doing so is to show its deficiencies compared to the semiparametric model we propose. The estimated coefficients of the log-odds are given in Table 3.3. Note that these parametric estimates suffer from a trimming bias as it is supposed that among

the people older than 65 the CDU affiliation is above average, as is the FDP affiliation among top earners.

	Mode effect	Female/Male	East/West	log(Income)	Age/10
SPD	3.863(0.484)**	0.084(0.063)	-0.458(0.083)**	-0.426(0.060)**	-0.088(0.028)**
A90G	1.039(0.701)	0.365(0.092)**	-0.584(0.129)**	-0.129(0.087)	-0.278(0.040)**
LP	3.649(0.909)**	-0.089(0.124)	2.127(0.136)**	-0.851(0.113)**	0.064(0.051)
FDP	-3.794(1.060)**	-0.535(0.134)**	0.223(0.155)	0.410(0.131)**	-0.291(0.055)**

Table 3.3.: *Parameter estimates for a fully parametric MNL with CDU as the reference category (standard errors in brackets).*

The reference mode is the largest party, i.e. the CDU. Relatively to the reference mode, being from the East substantially raises the likelihood of preferring the LP. The CDU is quite strong among older people, and thus it is not surprising that the impact of age is significantly negative for all other parties except of the LP for which we find a positive however insignificant coefficient. Female voters are more likely to support A90G and SPD. Being a young and female Western German resident is the typical characterization of an A90G-voter (cf. Walter, 2008). On average, presence of high income decreases the likelihood of supporting the LP and the SPD, while it increases that of supporting the FDP.

3.4. Semiparametric Analysis of Voter Profiles

Since the turn of the century the influence of the big parties is declining and the number of floating voters is increasing. In general, the identification with the different political parties has decreased (cf. Alemann, 2003). This development demands a more detailed view on the different voter profiles than the purely parametric MNL can offer. We thus propose to use an alternative model, namely the semiparametric MNL that has been introduced in Section 3.2. The details of our implementation are as follows. While the two dummies gender and region enter parametrically, age and log-income go to the nonparametric part:

$$\mathbb{P}(Y = k) = \frac{\exp\left(\beta_{1,k} \cdot \mathbf{Sex} + \beta_{2,k} \cdot \mathbf{East} + m_k(\mathbf{Inc}, \mathbf{Age})\right)}{\sum_{j=1}^{5} \exp\left(\beta_{1,j} \cdot \mathbf{Sex} + \beta_{2,j} \cdot \mathbf{East} + m_j(\mathbf{Inc}, \mathbf{Age})\right)} \tag{3.8}$$

The smoothing parameter for the nonparametric part was chosen on a grid of bandwidths from 0.5 to 1 times the standard deviation of age and log-income, respectively. For the

presentation of the results, we selected $h = 0.6$. We start by looking at the fitted parametric part of the model. Then we check for possible mode-specific effects. Finally, we give a detailed discussion of the fitted nonparametric part of the model.

The estimated impacts of gender and region are given in Table 3.4. As could be expected, they are quite close to those in Table 3.3 for the parametric MNL.

Mode	SPD	A90G	LP	FDP
Sex	0.097(0.057)	0.384(0.086)**	-0.084(0.119)	-0.531(0.133)**
East/West	-0.463(0.074)**	-0.532(0.121)**	2.138(0.130)**	0.229(0.152)

Table 3.4.: *Estimated coefficients for the parametric variables of the GPLM with CDU as the reference category*

For the purpose of identifiability, intercepts are not explicitly given in the proposed model – the functions m_k already account for any intercept effects. Such intercept effects describe the unexplained heterogeneity over modes. Thus, it sometimes may be of interest to find an indicator that can substitute and in this way explain part of the m_k. Typically one speaks here of persistency or loyalty of the voters. In order to instrument this, we consider the mode-specific variable 'member' (= number of party members in thousand), as given in Table 3.5.

Political Party		CDU	SPD	A90G	LP	FDP
Members	(in Thsd.)	720.792	561.239	44.677	60.338	64.880

Table 3.5.: *Size of political parties in Germany in 2006, measured in party members. Source: Niedermayer (2007)*

After the model given in (3.8) has been fitted, we performed the following additional linear regression to determine the influence of the mode-specific variable 'member':

$$\sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot m_k(\mathbf{Inc}, \mathbf{Age}) = \gamma_0 + \gamma_1 \cdot \sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot \mathbf{Mem}_k \tag{3.9}$$

The estimated coefficient of 'member' is given by $\hat{\gamma_1} = 0.00266$ (with standard error 0.000023, thus highly significant). The estimated intercept, $\hat{\gamma_0} = -1.681$, is highly significant, too. This reflects the non-explained 'mode' (intercept) effect.

Concerning the nonparametric functions m_k we consider two different model specifications: At first, in Subsection 3.4.1, we show the results under the assumption of an additive separability, i.e.

$$m_k(\mathbf{Inc}, \mathbf{Age}) = m_{1,k}(\mathbf{Inc}) + m_{2,k}(\mathbf{Age}), \quad k = 1, 2, 3, 4. \tag{3.10}$$

Neglecting the statistical discussion of dimensionality issues, the main advantage here is that the additive structure leads to an easier interpretation. However, in Subsection 3.4.2, we will see that due to strong interactions the bivariate functions comprise much more meaningful information about voter profiles. For a detailed discussion on nonparametric additive modeling with and without interaction we refer to Sperlich et. al. (2002).

3.4.1. Additive decomposition of the nonparametric part

To obtain an additive decomposition of the bivariate estimates, we applied standard R-routines for spline-based backfitting. For the reference group CDU these functions are constant, Figure 3.1 thus displays the marginal impacts of log-income and age only for the other parties. Recall that the additive functions are only identified up to an additive constant, so they could arbitrarily be shifted up or down without changing the slope. Obviously, the average linear impact should be and is the same as for the parametric linear model. Most of the rest, however, is quite different.

Over all income groups, SPD, LP and A90G find their strongest support in the low income group; even for the FDP the impact in this group is downwards sloped. For A90G and FDP this changes at a monthly income of about 1 500 € ($\approx \exp(7.3)$) where it starts to increase for both. This effect is stronger for the FDP, for which it increases up to the top earners, while for the A90G it changes again direction at monthly incomes higher than 8 000 €($\approx \exp(8.9)$). The main difference between SPD and LP here is that for the LP the slope is twice as steep as for the SPD; this is in agreement with the results in the parametric case.

Similarly, the influence of age over all modes is downwards sloped at first (recall that this is in comparison with the reference party CDU). This effect is especially strong for SPD and LP until the age of 30, where this effect flattens for the SPD and reverses for the LP whose support increases than steadily until the age of about 55. For the FDP and the A90G the support steadily decreases with age – over almost the same range but in different ways. For the SPD and the A90G the support is relatively stable in the age class from 30 to 45.

Summarizing, for the A90G the influence of age as well as of log-income turns out to be clearly nonlinear. The nonlinearities are mainly caused by the upper middle class and the

strong voter block related to the anti-nuclear movement – typically middle-aged people who became politically interested in the 80's. Furthermore, in Section 3.3 the effect of age has turned out insignificant in case of the LP. The picture drawn here is more precise as a considerable valley around the age of 32, mainly caused by voters from Eastern Germany, can be recognized. This generation was especially involved in and affected by the end of the GDR-regime during their youth and thus turns away from the heir of the Socialist Unity Party. Considering the other covariate, we see that the support in the lowest social class is high. Finally, recall that for the FDP the income effect was positive significant in the parametric model. The picture drawn here apparently is more informative.

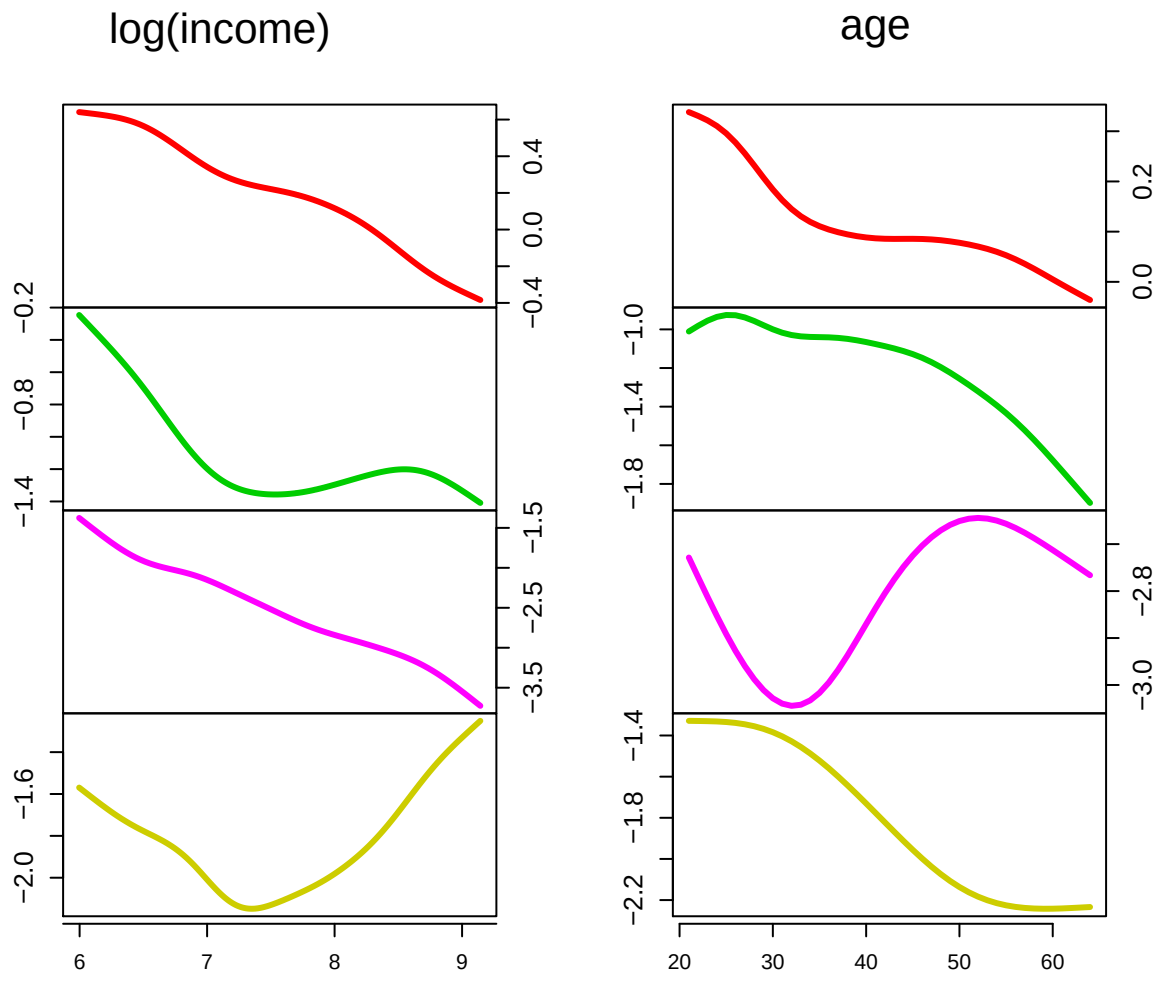

Figure 3.1.: *Additive decomposition of the impact functions* $m_{j,k}$, $j = 1, 2$, $k = 1, \ldots, 4$ *(SPD, A90G, LP and FDP top down).*

Further interpretation and explanations will be given when considering bivariate impacts in

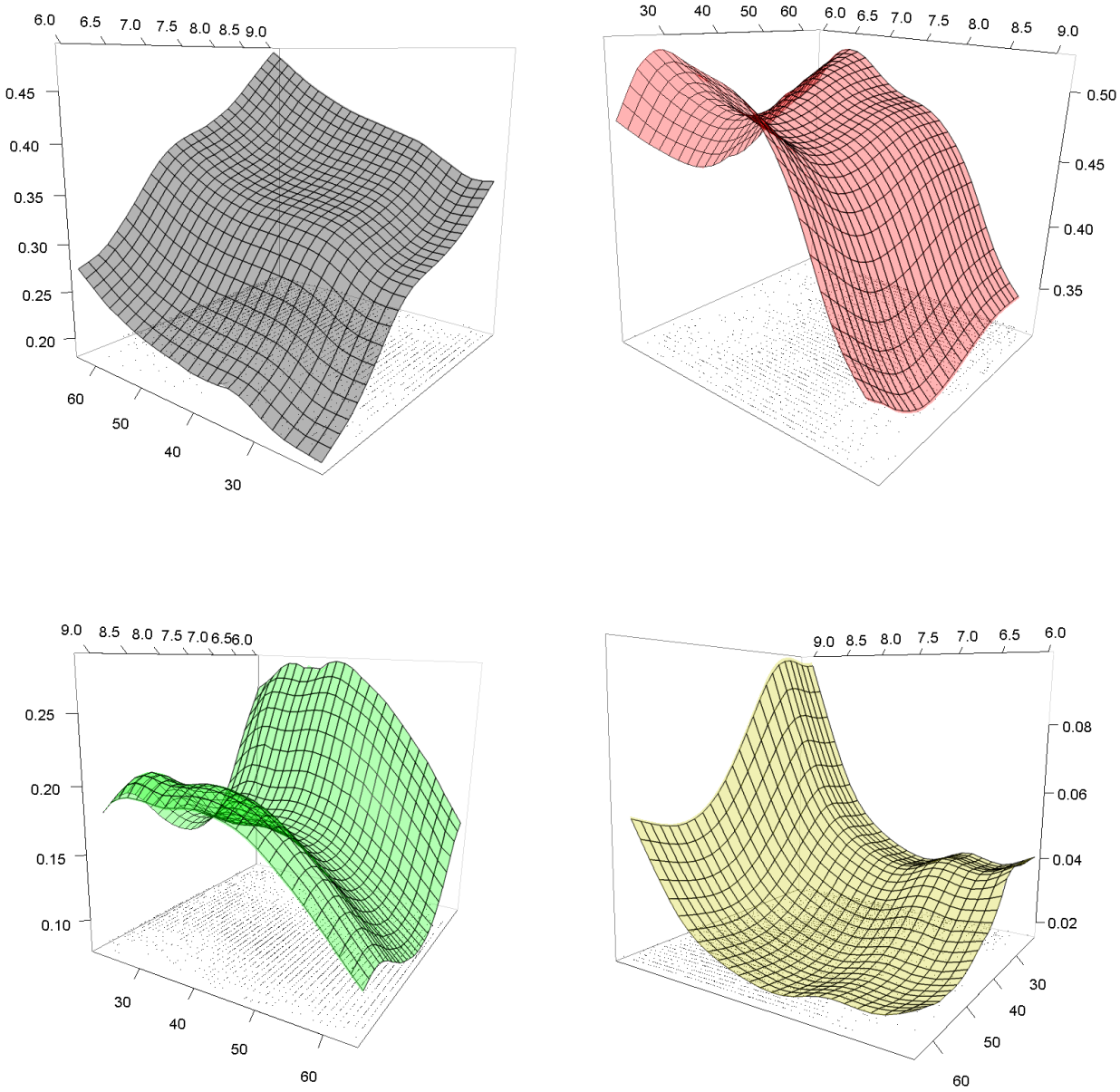

Figure 3.2.: *Probabilities of supporting a political party for women in Western Germany when the m_k are additive separable (CDU, SPD, A90G, and FDP top down).*

the subsequent section. For a better understanding of the estimation outcome, we conclude this section with plots of the affiliation probabilities determined by the model specifications (3.8) and (3.10), see Figure 3.2. When considering probabilities on comparable scales for West Germany, the LP surface looks almost flat and has therefore been skipped here. Instead, in Figure 3.2 we have added the outcome for the reference mode CDU which

has the highest chances to get voted by old people that are well-off. The plot reveals stronger support at high ages for all income-levels; this is due to the fact that the group of pensioners builds the strongest block of CDU-supporters (see Jesse, 2007). Certainly, most of the pensioners are not included in this sample. Nevertheless, it is most plausible that the upward trend continues beyond the age of 65. For the other three parties one can see how the impacts of $m_{1,k}$ and $m_{2,k}$ get reflected in the probabilities of votes. However, we will see in the next section that some of the simplifications implied by the assumed additivity are quite misleading. We used the R-package `RGL` (Adler and Murdoch, 2009) to illustrate the corresponding three-dimensional plots of the bivariate estimation. The tiny black points at the bottom indicate the observations. As the viewpoint can be changed arbitrarily, it is possible to turn the attention towards the axis referring to age, the axis referring to log-income or something in between. In Figures 3.3 and 3.4, the surfaces have been rotated such that the main features can most easily be recognized. Thus, the reader should be aware that the directions of the axes are not always the same. Furthermore, the R-package `akima` (based on Akima, 1978) has been used to generate a smooth surface via bivariate interpolation for irregularly spaced input data.

3.4.2. Bivariate nonparametric part

In this section we study how the results change when profile interaction between income and age is present. Figures 3.3 and 3.4 display the probabilities of supporting particular parties as a function of age and log-income. Keep in mind that for other values of the dummies for 'Sex' and 'East' the surfaces change (as the probability function is not linear) but only in the sense that some slopes become somewhat flatter or steeper; the general pattern remains the same.

As the figures change substantially compared to those from the last section, it is obvious that interaction does play an important role for voter profiles. Moreover, it is very difficult to find adequate parametric models that can appropriately reflect these interactions. This finding is nontrivial as it basically implies that most of the standard techniques usually applied are insufficient for a correct inference and interpretation.

In Figure 3.3, women in Western Germany are considered, these plots thus are the counterparts to Figure 3.2. The upper left plot shows the probabilities of supporting the CDU. Note that the general upward trend in age has disappeared for incomes larger than 1 500 €($\approx \exp(7.3)$), i.e. the major part of the sample population. With respect to the other covariate the likelihood of supporting the CDU increases as the income increases, no matter for what age. This is in sharp contrast to the SPD where the curve corroborates

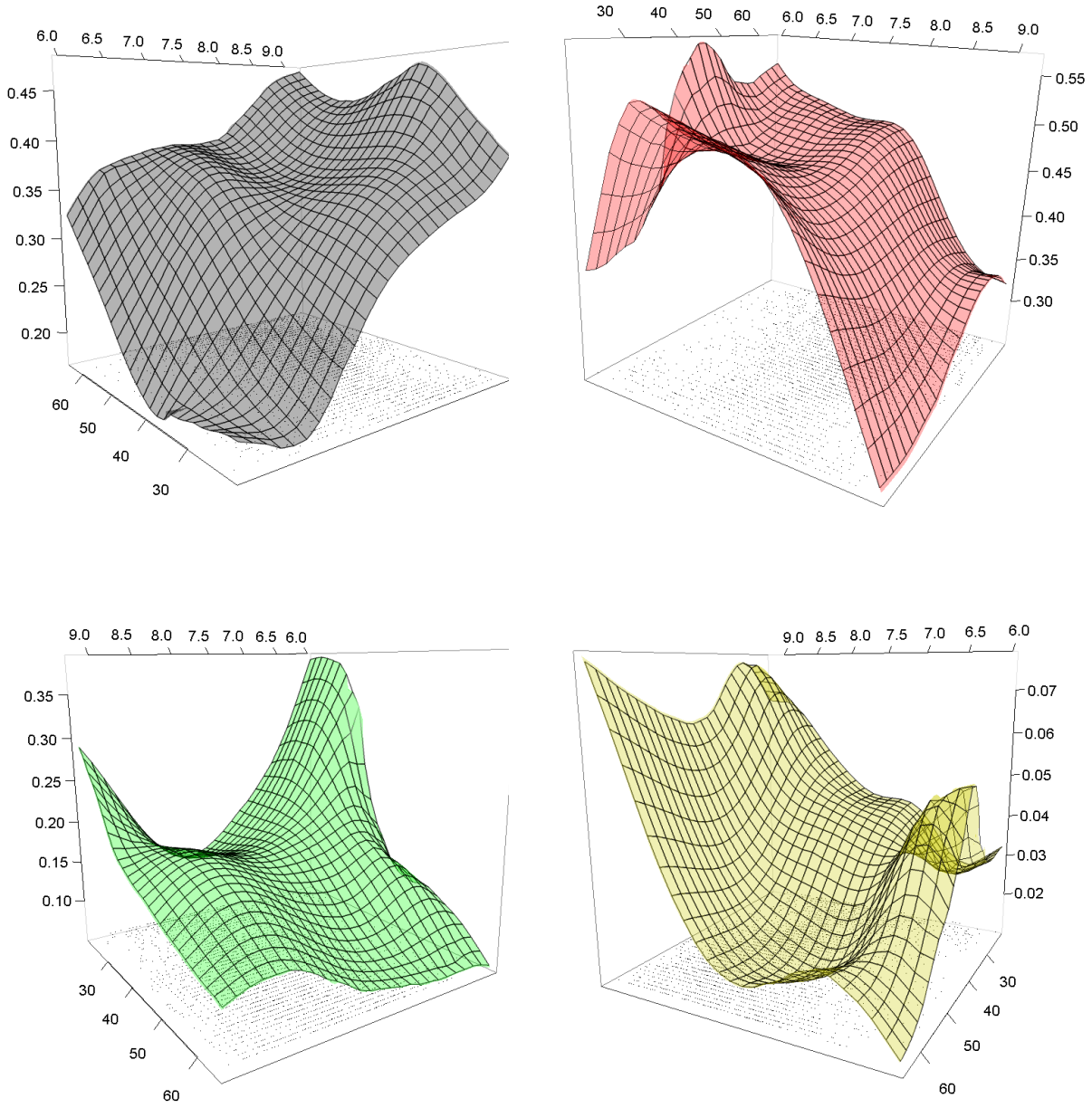

Figure 3.3.: *Probability for women in Western Germany of being a supporter of the different parties in a bivariate setting (CDU, SPD, A90G and FDP from the upper left to lower right)*

the statement that the SPD is a party of the working class across all ages. Apart from some changes for very low incomes and people aged over 40 (where we have only few data), this is the only plot that hardly changes compared to the additive counterpart.

In the bottom left picture, the stylized facts for A90G get more accented but with several

nuances. The typical voter of the green party is generally believed to be young and poor (mostly students) or middle-aged and well-off. According to the additive decomposition, young people as well as low income groups indeed constitute particularly strong voter blocks. However, the expected prominent role of the upper middle class cannot be detected. In the case of a bivariate modeling, we recognize a strong U-shape only within a particular group; young people with low and with high income respectively build quite strong voter blocks. In contrast, for people older than 40 the U-shape disappears; the upper-middle income class is the strongest in this age group. This is due to a strong support in the group of middle-aged people with academic background. Hence, in Figure 3 we see that the composition of A90G-voters is much more complex than indicated by the additive approach. By means of the bivariate model we are able to capture both major features described above.

The FDP has the smallest data basis, the estimates in this case thus are more wiggly and should not be over-interpreted. However, it can be recognized that the voters of FDP are rather young and rather high-earning. Especially in the working class the support for FDP is low.

Figure 3.4 gives the probabilities of being affiliated to the left party (LP), on the left hand side for men in Eastern Germany, and on the right hand side for women in Western Germany. The general structure of the surface is the same for both, the dummies for 'Sex' and 'East' merely cause a variation in magnitudes of the probabilities. The range for LP in Western Germany is too small to allow for a meaningful evaluation.

Considering the left plot of Figure 3.4, we recognize for East Germans the facts that have already been described in Section 3.4.1: strong support in the group of young and low-income people. However, as in case of the green party A90G, the bivariate impact function gives again more insight into the voter profile. First, according to the univariate graph in Figure 3.1 high income on average leads to rather small affiliation with the LP (which is to be expected for a left-wing party). However, from Figure 3.4 it can be seen that in Eastern Germany there is a notable popularity of the LP in the lower-middle and the upper-middle class for people older than 50. Exactly here you will typically find those that benefit from the GDR-regime. However, the influence of this group is slightly decreasing since the elections of 2005 (cf. Niedermayer, 2006.) Secondly, also among those who lost their jobs and social status during the change towards a market economy, the support is expected to be high, and it is – see low incomes for people older than 50. These specific characterisations of voter groups could not be captured by the additive decomposition as it averages over all ages to derive the influence of the log-income (and vice versa). In order to capture these effects we need to allow for interaction between income and age.

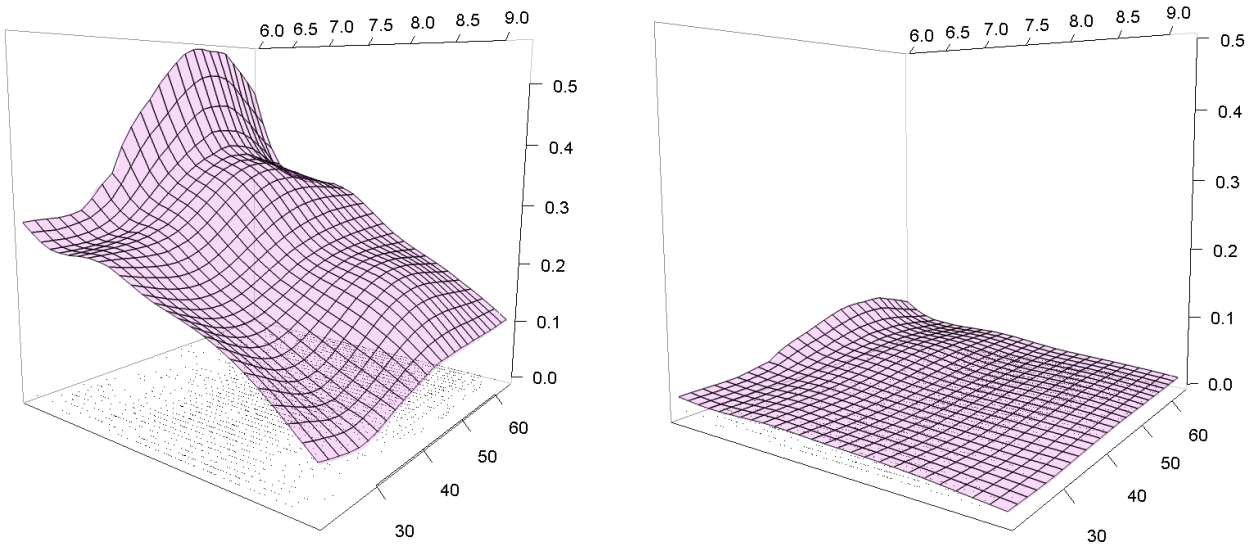

Figure 3.4.: *Probability for men in Eastern Germany (left picture) and women in Western Germany (right picture) of being a supporter of LP*

3.5. Conclusions and Outlook

We have introduced a semiparametric multinomial logit model in order to estimate political party affiliation in Germany, where we have a response variable with more than two possible outcomes. Our approach extends the GPLM model for the binary case as described by Müller (2001). The derivation of the model is done in compliance with the GPLM framework and hence the mathematical properties of asymptotic normality, consistency and efficiency of the estimators are fulfilled. The algorithm usually converges very fast.

We add flexibility to the standard MNL since we do not assume a specific functional form for the influence of a subset of the explanatory variables, and since we allow for interactions between covariates. We consider individual-specific as well as mode-specific variables. The model can capture binary and continuous variables, whereas the latter are most suitable to be modeled nonparametrically. It is possible to consider just a simple linear structure or one-dimensional additive as well as bivariate or multivariate nonparametric functions. Due to illustration purposes we recommend to restrict oneself to bivariate estimation. The flexibility of the model enables to get much more comprehensive insights, e.g. in the profiles of specific voter groups. Indeed, the presented results give a more detailed view on

the support of political parties than any other existing method. In particular, the three-dimensional plots and the possibility to change the viewpoint arbitrarily enable readers with little statistical background to interpret the results and provide a basis for further decisions.

Our model is directly applicable to further research concerned with similar problems, e.g. brand choice in marketing studies or choice of transportation modes. Due to the coverage of different types of covariates our model has a large scope of applications.

In the application of estimating political party affiliation we have seen that our model overcomes many deficiencies of both parametric modeling and nonparametric modeling assuming an additive separability. The strong nonlinearities and interactions between covariates show how complex voter groups nowadays are structured, and underline the need for more sophisticated modeling than the conventional MNL can offer. This becomes particularly obvious in regard of the results for the parties A90G and LP, where our analysis reveals the presence of strong interactions between age and income as well as highly pronounced nonlinearity. Possible reasons to support A90G are manifold. For young people the popularity is highest in the low income group and likewise for rich people. Additionally, middle-aged voters which are well-off support A90G. The voters of the LP are certainly quite heterogeneous due to its recent history: On the one hand, the old guard of the Eastern dictatorship and economic losers of the change in Eastern Germany belong to the group of PDS supporters. On the other hand, in West Germany the WASG has successfully increased their popularity as an electoral alternative for more social justice. Hence, there the support in the lowest social class is high. Especially in Eastern Germany, both big parties (CDU and SPD) suffer from the popularity of the LP. In sum, the kind of voters of the political parties are very distinct. A more or less stable base can only be detected by the mode intercepts.

3.6. References

Abe, M. (1998). Measuring consumers 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.

Adler, D. and Murdoch, D. (2009). Package 'rgl' — 3D visualization device system (OpenGL). *Available on: http://rgl.neoscientists.org.*

Akima, H. (1978). A Method of Bivariate Interpolation and Smooth Surface Fitting for Irregularly Distributed Data Points. *ACM Transactions on Mathematical Software* **4/2**: 148–164.

Alemann, U. v. (2003). *Das Parteiensystem der Bundesrepublik Deutschland*. Schriftenreihe / Bundeszentrale für Politische Bildung (ed.), Bonn.

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

Domencich, T. and McFadden, D. (1975). *Urban Travel Demand: A Behavioral Analysis*. North Holland, Amsterdam.

Dow, J. K. and Endersby, J. W. (2004). Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral studies,* **23/1**: 107–122.

Härdle, W., Huet, H., Mammen, E. and Sperlich, S. (2004a). Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, **20/2**: 265–300.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004b). *Nonparametric and Semiparametric Models*. Springer Series in Statistics, Berlin.

Hastie, T.J. and Tibshirani, R.J. (1986). Generalized Additive Models *Statistical Science*, **1/3**: 297–318.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*. Chapman and Hall, London. item Jesse, E. (2007). Die Bundestagswahl 2005 im Spiegel der repräsentativen Wahlstatistik. In *Statistisches Bundesamt (ed.): Wirtschaft und Statistik, Available on: www.destatis.de*.

Kalyanaram, D. and Little, J.D.C. (1994). An Empirical Analysis of. Latitude of Price Acceptance in Consumer Package Goods. *Journal of Consumer Research*, **21/3**: 408–418.

Kneib, T., Baumgartner, B. and Steiner, W.J. (2007). Semiparametric Multinomial Logit Models for Analysing Consumer Choice Behaviour. *Advances in Statistical Analysis*, **91/3**: 225–244.

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

Lösche, P. (1993). *Kleine Geschichte der deutschen Parteien.* Verlag W. Kohlhammer, Stuttgart.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Zarembka, P. (ed.): 105–142.

Müller, M. (2001). Estimation and testing in gerneralized partial linear models - A comparative study. *Statistics and Computing*, **11/4**: 299–309.

Niedermayer, O. (2006). Die Wählerschaft der Linkspartei. PDS bei der Bundestagswahl 2005. *Zeitschrift für Parlamentsfragen*, **37/3**: 525–538.

Niedermayer, O. (2007). Parteimitgliedschaften im Jahre 2006. *Zeitschrift für Parlamentsfragen*, **38/2**: 368–375.

Quinn, K. M., Martin, A. D. and Whitford, A. B. (1999). Voter choice in Multi-Party Democracies: A Test of Competing Theories and Models. *American Journal of Political Science*, **43/4**: 1231–1247.

Roberts, G. K. (1997). *Party politics in the New Germany.* A Cassell Imprint; The New Germany Series, London.

Rodríguez-Póo, J. M., Sperlich, S. and Vieu, P. (2003). Semiparametric Estimation of Weak and Strong Separable Models. *Econometric Theory*, **19/6**: 1008–1039.

Severini, T.A. and Wong, W.H. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of Statistics*, **20/4**: 1768–1802.

Sperlich, S., Tjøstheim, D., and Yang, L. (2002). Nonparametric Estimation and Testing of Interaction in Additive Models. *Econometric Theory*, **18/2**: 197–251.

Tutz, G. and Scholz, T. (2004). Semiparametric Modelling of Multicategorical Data. *Journal of Statistical Computation and Simulation*, **74/3**: 183–200.

Walter, F. (2008). *Baustelle Deutschland.* Suhrkamp Verlag, Frankfurt am Main.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models, *Journal of the Royal Statistical Society B*, **58/3**: 481–493.

4. A Semiparametric Model of Urban Transport Demand

Abstract

The aim of this work is to estimate a transport demand function for university students in the Bilbao area. We want to compare the estimation results based on different model approaches. To do this, we use on the one hand a simple parametric and on the other hand a semiparametric approach that allows to model the utility that individuals (i) get from the mode of transport they use (car, train, bus, underground).
We find significant differences between the two approaches in the multinomial case indicating the inappropriateness of the standard parametric methods compared to the estimation of the semiparametric model. The greatest differences seems to occur in the marginal effect of the continuous variable income as well as price.
This chapter of the thesis is a collaboration with Prof. Ana Fernández-Sainz and Prof. Javier Bilbao Ubillos from the University of Bilbao, Spain, as well as Prof. Dr. Stefan Sperlich. The main contribution of the author of this thesis is made in the development of the semiparametric approach as well as the implementation of all considered models in R. The presentation as well as interpretation of the results is also the task of the author of this thesis.

4.1. Introduction

The main objective of this paper is to explore different methodological approaches to specify and estimate a transport demand function. The principle goal is to check, if the distinction between the results deriving from different approaches are sufficient to change the recommendations for public policy makers. This objective is in line with a wide-ranging research program - developed by the Spanish authors of this paper - aimed at drawing up a more effective way of substituting private car by collective transport in congested areas. More concretely we try to estimate the effects of improving the available supply of public transport or cutting public transport prices to attract new passengers from private car.

Researchers would like to focus not only on the potential efficiency of these instruments but also on the best use. They want to state precisely what type of improvement in the availability of modes of collective transport (underground, bus and train) or which selected reduction of collective transport prices would have the greatest potential impact on reducing the use of private vehicles. To do this one must first analyse the determinants of the demand for public transport by students located in a densely populated urban area endowed with alternative means of public transport. We will concentrate on the role played by quality of service (in terms of trip length and frequency) in this demand.

The main aim of this study is to find model formulations which best describe the influence of various covariates and hence deliver a good basis to consider the possibilities of reducing congestion and pollution by acting on these student trips.

We propose a probabilistic function for demand of transport in which individuals are faced with the choice between the specific modes of transport available (car, underground, bus, train or a combination of modes). The probabilistic models we consider are based on maximization of individual utility, which we consider to be dependent on the characteristics of the means of transport (quality and price) and those of the individuals themselves. This model makes it possible to quantify individuals' responses to changes in features of the alternatives.

To analyse this model we need a sufficient data base. We have built up a base with a high statistical level, from 1 780 surveys filled in by students who travel daily to university and reside in areas surrounding the city of Bilbao. These observations make it possible to obtain conclusions about the actual influence that the different variables have on the configuration of the students' demand for transport.

These considerations could serve as a reference when defining public transport policy that would meet the needs of this large group of potential users.

The work is structured as follows. In Section 4.2 we introduce the model of the demand for public transport and its econometric formulation. Section 4.3 provides the theoretical basis of a semiparametric approach to handle the transport demand function. In section 4.4 we compare the different approaches and analyse the estimation results of the previous models. Finally, we give some conclusions.

4.2. The Economic Formulation and the Econometric Model

Discrete choice models have been used in different economic applications including choice of transportation mode (McFadden, 1974), choice of residence (McFadden, 1978), choice of vehicle type (Choo and Mokhtarian, 2004), etc. In these models, a set of individuals ($i = 1, ..., N$) are faced with a range of mutually exclusive alternatives, that is they have to make a choice (k) within the set of possible choices ($k = 1, .., K$). The traditional theory of rational choice asserts that individuals can rank possible alternatives in order of preference and make choices in a deterministic and coherent way. They will always choose from available alternatives the option they prefer and therefore, in two identical situations the optimal choice will be the same. The description of the econometric model is based on the remarks in the forerunner paper (cf. Bilbao-Ubillos and Fernández-Sainz, 2004).

It seems, however, that in practice human behaviour is not as rational as traditional economic theory assumes. For this reason *choice* has been analyzed as a probabilistic process rather than a deterministic one: an individual i chooses an option $k = 1, ..., K$ with probability $P_i(k)$.

Depending on the nature of the random mechanisms involved, different alternatives have been taken into account to obtain the probabilities of choice. We use the approach of McFadden (1979), which deals with the problem by assuming that decision rules are deterministic but utilities are stochastic. Probability choice and demands are obtained from the maximization of random utility.

$$P_i(k) = Pr(\widetilde{U}_{ik} = max_l\ \widetilde{U}_{il}) \quad \text{for} \quad l = 1, ..., K \tag{4.1}$$

where $\widetilde{U}_{ik}$ is individual i's utility level when he/she makes the choice k[1].

[1]Observe that utility depends only on the mode of transport chosen by the individual and not on the consumption of other goods. This assumption is due to the requirement of consistency between discrete choice models and the maximization of random utility: the indirect utility function must be additively separable in income (this determines the indirect utility from the good consumption) and the good

Given that urban population consists of a large number of individuals N, the expected population demand for alternative k, D_k, is given by:

$$D_k = \sum_{i=1}^{N} P_i(k) \tag{4.2}$$

Domenich and McFadden (1975) assume that the utility associated with each mode of transport is a function that depends on the mode characteristics (z) and on the individual's socioeconomic characteristics (w) plus an additive error term (e).

Then, if we assume a linear relationship, we have:

$$\widetilde{U}_{ik} = \alpha_i + z'_{ik}\beta + w'_i\gamma_k + e_{ik} \tag{4.3}$$

where z_{ik} are the characteristics of mode of transport k as individual i perceives them and w_i are socio-economic characteristics of individual i. In this way, individual i prefers (in expected terms) mode of transport k to mode of transport l if and only if $\widetilde{U}_{ik} > \widetilde{U}_{il}$:

$$(z'_{ik} - z'_{il})\beta + w'_i(\gamma_k - \gamma_l) + (e_{ik} - e_{il}) > 0 \tag{4.4}$$

Socio-economic characteristics will be determinant for choice if and only if $\gamma_k - \gamma_l \neq 0$. In this case individuals with different socio-economic characteristics have value modes of transport differently. When $\gamma_k - \gamma_l = 0$ socio-economic characteristics do not influence individuals' choice.

Therefore, assuming that individuals are rational and that they maximize their perceived utility subject to appropriate constraints, the econometric model that we use to quantify transport demand needs some preliminary definitions.

Let be y^*_{ik} the latent variable we use to denote the indirect utility level from mode of transport k :

$$y^*_{ik} = V_{ik}(z_{ik}, w_i) + \epsilon_{ik} \tag{4.5}$$

When the student chooses a mode of transport, he has six different possibilities: to drive his own car or to travel by car as a passenger, train, bus, underground and any combination of the foregoing.

In these terms the observable variable is:

$$Y_i \;=\; k \quad \Longleftrightarrow \quad y^*_{ik} = \text{Max}(y^*_{il}) \qquad \text{for} \quad l = 1, ..., 6$$

consumption) and the characteristics of the mode of transport chosen by the individual (this determines the indirect utility from the use of this mode). To compare the utility of two different alternatives it is sufficient to take into account only the last term.

In these terms, the probability, Pr_{ik}, that individual i will choose mode of transport k taking into account the latent variable and the distribution of ϵ_{ik} assuming that ϵ_{ik} are independently and identically distributed with the type I extreme-value distribution $F(\epsilon_{ik} < \epsilon) = exp(-e^{-\epsilon})$, is given by:

$$Pr_{ik} = Pr(Y_i = k) = \frac{exp(V_{ik})}{\sum_{l=1}^{6} exp(V_{il})} \tag{4.6}$$

whereas we can only identify the differences in indirect utilities. Therefore, we have to define a reference category K, where for this purpose the indirect utility V_{iK} is set equal to zero.

McFadden (1978) showed that this type of model can be derived from the theory of utility maximization as the multinomial logit model.

4.3. Semiparametric Approach

During the past two decades, the number of theoretical and empirical studies on nonparametric and semiparametric methods for estimating and testing microeconomic or macroeconomic models has grown rapidly. For a general introduction to nonparametric and semiparametric econometrics, see e.g. Härdle et. al. (2004).

Examples for nonparametric or semiparametric issues related to discrete choice modeling have been presented by Kneib, Baumgartner and Steiner (2007) or Tutz and Scholz (2004). Whereas the former paper studied consumer choice behaviour based on a Bayesian approach and the latter applied penalized basis functions on an example about hereditary diseases.

Huang and Nychka (2000) proposed a nonparametric multiple-choice model based on the penalized likelihood method within a Random Utility Maximization framework and applied it to the non market evaluation of recreational sites. Abe (1999) relaxed the traditional assumption in the standard multinomial logit (MNL) formulation by introducing the modified generalized additive models and applied it to panel choice data.

Our estimation procedure is based on the profile likelihood algorithm proposed by Severini and Wong (1992). If $K = 2$ the multinomial model reduces to the binary case. A detailed derivation of the algorithm for the simple binary case can be found in Müller (2001).

In our approach, we consider a semiparametric multinomial logit model with K different outcome categories that have no natural order. The conditional probability of outcome $Y_i = k$, $k = 1, \ldots, K$, given the individual covariate vectors $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})^t \in \mathbb{R}^p$ and $\mathbf{T}_i = (T_{1i}, \ldots, T_{qi})^t \in \mathbb{R}^q$ is assumed to be given by

$$Pr_{ik} = Pr(Y_i = k \,|\, \mathbf{X}_i, \mathbf{T}_i) = \frac{\exp(\boldsymbol{\eta}_k)}{\sum_{l=1}^{K} \exp(\boldsymbol{\eta}_l)} = \frac{\exp\left(\mathbf{X}_i{}^t \boldsymbol{\beta}_k + m_k(\mathbf{T}_i)\right)}{\sum_{l=1}^{K} \exp\left(\mathbf{X}_i{}^t \boldsymbol{\beta}_l + m_l(\mathbf{T}_i)\right)}. \tag{4.7}$$

We set for identification $\boldsymbol{\beta}_K = \mathbf{0}$, and $m_K(\cdot) \equiv 0$, i.e. K is the reference mode. Each $m_k(\cdot)$, $k = 1, \ldots, K$, is assumed to be a smooth function with domain $\mathbb{R}^q$ and each $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp})^t$, $k = 1, \ldots, K-1$, denotes an unknown parameter vector.

We have to rearrange the covariates which are presented in the economic formulation of the model from the previous section. In the semiparametric approach we consider individual- and mode-specific covariates and divide these covariates into two distinct groups, the first group will be modeled parametrically and the second group will go into the nonparametric part of our approach.

The first group comprehends the dummy variables of the socio-economic characteristics (w_i) and variables with only few parameter values which should not be modeled nonparametrically. Additionally, it covers the characteristics (z_{ik}) of the mode of transport k as individual i perceives them. These variables are covered in $\mathbf{X}$. Secondly, we have socio-economic characteristics (w_i) which are quasi-continuous and hence are suitable to be modeled nonparametrically, denoted by $\mathbf{T}$.

Supplementary, we consider purely mode-specific covariates, which are independent of the individual perception $\mathbf{Z} = (z_k)$. Note that the nonparametric functions m_k capture any mode-specific effect, hence $\mathbf{X}$ must not contain mode-specific dummies, see further discussion below.

It would be easy to find estimators for the vectors $\boldsymbol{\beta}_k$, if the functions $m_k(\cdot)$ were known, and vice versa. Hence, we follow the ideas of profiled likelihood by Severini and Wong (1992). In the following, $\mathcal{L}(\boldsymbol{\eta}_i(\cdot), y_i)$ denotes the log-likelihood of (4.7) of the ith observation with predictor $\boldsymbol{\eta}_i$.

We use an iterative estimation procedure and regard the functions $m_k(\cdot)$ as nuisance when estimating the finite-dimensional parameters $\boldsymbol{\beta}_k$. Therefore, we compute the profile likelihood which is in this case only a function of $\boldsymbol{\beta}_k$

$$\mathcal{L}_p(\boldsymbol{\beta}_k) = \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k), y_i), \tag{4.8}$$

Alternately, in order to estimate the so-called least favorable curve $m_{k,\boldsymbol{\beta}\cdot}(t)$ at point $\mathbf{t} := (t_1, \ldots, t_q)$ for given $\boldsymbol{\beta}_k$, $k = 1, \ldots, K-1$, take a q-dimensional kernel $K : \mathbb{R}^q \to \mathbb{R}$,

bandwidth matrix $\mathbf{H} \in \mathbb{R}_+^{q\times q}$, and consider the local likelihood

$$\mathcal{L}_s(m_{k,\boldsymbol{\beta}.}(\mathbf{t})) = \sum_{i=1}^{N} (\det \mathbf{H})^{-1} K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big)\mathcal{L}(\boldsymbol{\eta}_i(m_{k,\boldsymbol{\beta}.}(\mathbf{t})), y_i), \tag{4.9}$$

note that the estimate of $m_{k,\boldsymbol{\beta}.}(\cdot)$ will depend on all $\boldsymbol{\beta}_l$, $l = 1,\ldots,K-1$, indicated by the additional index '$\boldsymbol{\beta}.$'.

To meet the requirements of a Newton-Raphson algorithm in the estimation procedure we have to calculate the first two derivatives l'_{ik} resp. l''_{ik} of

$$l_i(\boldsymbol{\eta}_i(\cdot)) := \mathcal{L}(\boldsymbol{\eta}_i(\cdot), y_i),$$

with respect to $\eta_i = \mathbf{x}_i^t\boldsymbol{\beta}_k + m_k(\mathbf{t}_i)$.

To obtain the maximum likelihood, successively from mode 1 to mode K, we have to solve the first order condition (setting the first derivatives equal to zero) and implement the following Newton-Raphson-type algorithm:

1. *Find appropriate starting values* $\boldsymbol{\beta}_k^{(0)}$, $m_k^{(0)}(\cdot)$, $k = 1,\ldots,K-1$ *(e.g. by fitting an appropriate parametric MNL) and set* $j = 0$.

2. *For* $k = 1,2,\ldots,K-1$, *compute*

$$\begin{aligned} \boldsymbol{\beta}_k^{(j+1)} &= \boldsymbol{\beta}_k^{(j)} - \mathcal{B}^{-1}\sum_{i=1}^{N} l'_{ik}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k^{(j)}))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i)) \\ \textit{with} \qquad \mathcal{B} &= \sum_{i=1}^{N} l''_{ik}(\boldsymbol{\eta}_i(\boldsymbol{\beta}_k^{(j)}))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i))(\mathbf{x}_i + m'^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}_i))^t \end{aligned}$$

3. *For* $k = 1,2,\ldots,K-1$, *compute*

$$m^{(j+1)}_{k,\boldsymbol{\beta}.}(\mathbf{t}) = m^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}) - \frac{\sum_{i=1}^{N}(\det \mathbf{H})^{-1}K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big)l'_{ik}\big(\boldsymbol{\eta}_i(m^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}))\big)}{\sum_{i=1}^{N}(\det \mathbf{H})^{-1}K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big)l''_{ik}\big(\boldsymbol{\eta}_i(m^{(j)}_{k,\boldsymbol{\beta}.}(\mathbf{t}))\big)}$$

for all points $\mathbf{t}$ *at which the function* $m_{k,\boldsymbol{\beta}.}(\cdot)$ *is to be estimated.*

4. *Repeat steps* **2.**–**3.** *for* $j = 1,2,\ldots$ *until convergence.*

It is convenient to estimate the functions $m_{k,\boldsymbol{\beta}.}(\cdot)$ in step **3** at the observation points $\mathbf{t}_i$, $i = 1,\ldots,N$, as this guarantees that independent of the bandwidth choice at least for one observation $K\big(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{t}_i)\big)$ is nonzero.

This estimation procedure delivers asymptotically normal, $\sqrt{n}$-consistent and efficient estimators for the vectors $\boldsymbol{\beta}_k$ owing to likelihood estimation. For the m_k one obtains consistent

estimators with statistical properties typical for nonparametric kernel smoothing, see also Rodríguez-Póo et. al. (2003).

Recall that the m_k will automatically capture any mode-specific effect. If there is also a vector of mode characteristics $\mathbf{Z}_k = (Z_{1k}, \ldots, Z_{rk})^t \in \mathbb{R}^r$ available, then one might want to regress them on the m_k. Note that a nonparametric modeling of their influence would not make much sense as the support is discrete consisting of few values. Therefore, the influence of the mode-specific covariate vector should be modeled by a simple linear relation with unknown parameter $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_r)^t$ resulting in an optional fifth step:

5. *After convergence of $\boldsymbol{\beta}_k$ and $m_{k,\boldsymbol{\beta}.}$, perform an additional regression*

$$\sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot m_{k,\boldsymbol{\beta}.} = \gamma_0 + \sum_{j=1}^{r} \gamma_j \cdot \sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot Z_{jk} \,.$$

4.4. Estimation Results

The data from which we draw up the estimation were obtained by means of a written query between September and October (1996). The query was either incorporated in the student registration envelope (University of The Basque Country) or answered at the begining of a lecture (University of Deusto). The basic descriptive statistics are shown in Table 4.1.

Variable	Description	Mean (std)
Sex	dummy, 1 if female	0.5578 (0.496)
Age	dummy, 1 if age under 21	0.4511 (0.498)
University	dummy, 1 if Public University	0.6280 (0.483)
Education	parents‘ level of education	1.753 (0.806)
Income	Value of family's house	19473.8427 (7253.465)
Frequency	minutes between two services	15.2882 (8.208)
Price	price of ticket	136.5730 (80.607)
Trip Time	duration of trip, in minutes	45.3624 (15.664)

Table 4.1.: *Statistics of the explanatory variables: mean and std. deviat. in brackets.*

Some of the variables need additional explanations. The variables sex, age and uni are individual-specific dummy variables. The variable education is given by the parents' level of education and used to describe the human capital that allows parents to obtain a better job and higher wage. These four variables are covered in $\mathbf{X}$.

We include one proxy-variable to approximate the income of the individual. This variable is basically the value of the family's house, given district and specific location. The continuous variable income is put in the nonparametric part $\mathbf{T}$ of our semiparametric approach. Additionally, in the estimation of the nonparametric part, we divide the income by its standard deviation.

The variables price and trip time are individual- as well as mode-specific variables. These two variables depend on the individual characteristics and on the mode of transport which is chosen by the individual. However, we only observe the price and the trip time of the travel mode, which the individual has chosen. Hence, we have to interpret these values as individual-specific covariates and therefore we additionally incorporate these two variables in $\mathbf{X}$.

The variable frequency is a fully mode-specific covariate. The time between two services depends only on the mode of transport. The private car has a frequency equal to zero as the driver can decide to start whenever he wants. The other transport modes have a specific frequency depending on the corresponding traffic system. This mode-specific variable is included in $\mathbf{Z}$.

In this section we want to compare the results achieved through different estimation methods for the transport demand function and try to explain the differences. We have proved several model formulations, on the one hand parametric and on the other hand semiparametric approaches.

4.4.1. Binary logit model

In a first step, we estimate a simple binary logit model for the decision between public and private transport modes. Thus, in the private case, we subsume the options to take the own car or to go by car as a passenger. The public segment combines bus, train, underground and all other options. This model represents the first important decision in the choice of transport mode between being self-determined or being dependent on public supply. The descriptive statistics show that 85.3%(14.7%) of the individuals choose a public (private) transport mode. We consider both cases, the fully parametric model (4.10) and the semiparametric approach (4.11), in which the income variable is modeled nonparametrically.

$$\mathbb{P}(Y = \text{Publ}) = \frac{\exp\left(\mathbf{X}^t\boldsymbol{\beta}_k\right)}{1 + \exp\left(\mathbf{X}^t\boldsymbol{\beta}_k\right)}, \tag{4.10}$$

where $\mathbf{X}^t$ =(1,Age, Sex, Uni, Edu, Price, Time, Inc).

$$\mathbb{P}(Y = \text{Publ}) = \frac{\exp\left(\tilde{\mathbf{X}}^t\boldsymbol{\beta}_k + m_k(\text{Inc})\right)}{1 + \exp\left(\tilde{\mathbf{X}}^t\boldsymbol{\beta}_k + m_k(\text{Inc})\right)}, \tag{4.11}$$

where $\tilde{\mathbf{X}}^t$ =(Age, Sex, Uni, Edu, Price, Time). The intercept is included in the nonparametric part m_k(Inc).

The results of the estimation are given in Table 4.2 and Figure 4.1. As can be seen, the coefficients and standard deviations of all parametric covariates are almost the same in both cases.

Variable	Parametric logit	Semiparametric logit
Constant	0.164 (0.383)	... (...)
Age	-0.525^{***} (0.154)	-0.531^{***} (0.153)
Sex	-0.785^{***} (0.153)	-0.775^{***} (0.153)
University	-0.292 . (0.161)	-0.290 . (0.161)
Education	0.181 . (0.094)	0.177 . (0.095)
Income-V.H	0.398 *** (0.103)	... (...)
Price	0.018 (0.094)	0.027 (0.095)
Time	-0.058^{***} (0.006)	-0.059^{***} (0.006)
N	1780	1780
LogL	-634.45	-603.59

Table 4.2.: *Estimates of a binary model. Standard deviation in brackets. Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1*

The signs of the effects are as expected. Price is an individual and mode-specific covariate, but we have to interpret the price of a transport mode as the price as individual i perceives it. From Table 4.2 we see that the price has only very little effect in this case. Time is also individual- and mode-specific and we have to interpret it in the same way as price. But, because of sparse data we have set all values of time > 60 to 60 without substantial modifications in the interpretation. Hence, the trips we consider lasts at most 60 minutes. We can see, that longer trip time has a significant negative effect on choosing public transport modes.

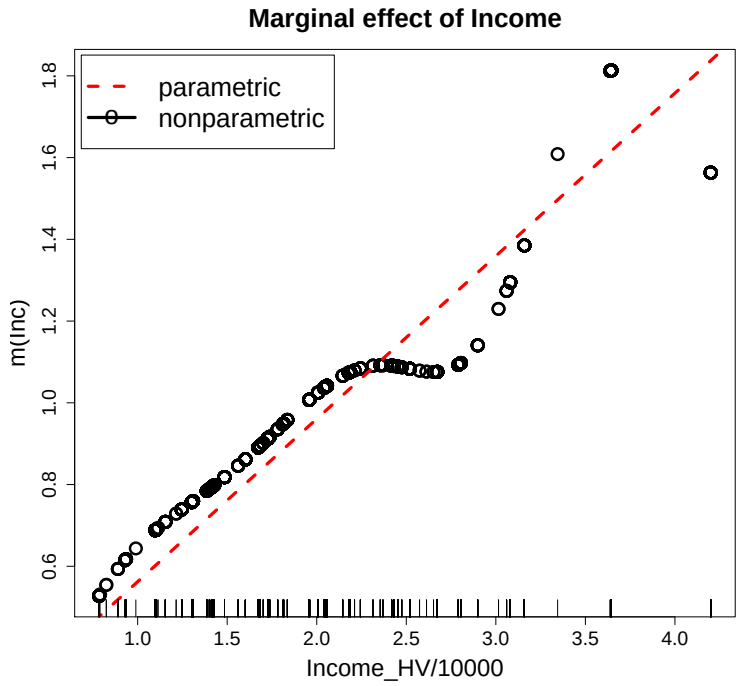

Figure 4.1.: *Marginal effect of Income m_{Inc} for a semiparametric binary logit model*

The effect of income is positive and highly significant in the parametric model. Additionally, the nonparametric effect on income captures the intercept of the parametric model. In principle, we can adhere to the parametric interpretation, if we look at Figure 4.1. The plot of the marginal effect of income in the semiparametric case shows a clearly upward trend for low income groups and only a slight negative effect for higher income which is mainly due to sparse data in this region.

Therefore, the main conclusion of this subsection is that if we estimate a binary model, parametric and semiparametric approaches produce the same results.

4.4.2. Multinomial logit model

In a second step, we consider a more comprehensive model. We estimate the choice between the different transport modes with aid of a multinomial logit model. The multinomial logit model is the easiest model for unordered discrete choices. In our example, we can think of individuals which are faced with a discrete set of alternatives and we can assume that the different transport modes do not have a natural ordering. Hence, our multinomial logit model has six possible outcomes (own car, car passenger, bus, train, underground and others). The percentages for the different transport modes are given in Table 4.3.

Transport modes	Car	Pass	Bus	Train	Under-ground	Others
Percentage	10.8	3.9	29.3	14.0	15.7	26.3

Table 4.3.: *Percentages for the different transport modes.*

In order to guarantee identifiability of the model we have to define a reference category – here: 'Others'. All estimated coefficients have to be interpreted as higher or lower chances of choosing the specific mode compared to the reference category. Again, we consider both cases, the fully parametric model (4.12) and the semiparametric approach (4.13), in which the income variable is modeled nonparametrically.

$$\mathbb{P}(Y = k) = \frac{\exp\left(\mathbf{X}^t \boldsymbol{\beta}_k\right)}{\sum_{l=1}^{6} \exp\left(\mathbf{X}^t \boldsymbol{\beta}_l\right)}, \tag{4.12}$$

where $\mathbf{X}^t$ =(1,Age, Sex, University, Education, Price, Time, Income).

$$\mathbb{P}(Y = k) = \frac{\exp\left(\tilde{\mathbf{X}}^t \boldsymbol{\beta}_k + m_k(\text{Inc})\right)}{\sum_{l=1}^{6} \exp\left(\tilde{\mathbf{X}}^t \boldsymbol{\beta}_l + m_l(\text{Inc})\right)}, \tag{4.13}$$

where $\tilde{\mathbf{X}}^t$ =(Age, Sex, University, Education, Price, Time). The intercept is included in the nonparametric part m_k(Inc).

The results for the estimated coefficients of the multinomial logit models are presented in Table 4.4 and 4.5. Due to the estimation procedure we get a coefficient for every mode except the reference for every individual-specific variable. These coefficients has to be interpreted as chances compared to the reference. In the parametric model it is possible to include either an intercept for every mode or one or more mode-specific covariates. In the nonparametric model we can use a two-step procedure. First estimate the whole model only with individual-specific covariates and afterwards decompose the nonparametric estimates into mode-specific components and a remaining part which is not avowed by other covariates. The informations which we can gather from the multinomial logit model are much more detailed than in the binary case.

First, we will present the parametric coefficients with mode-specific intercepts in Table 4.4. Again we have the effect of the individual-specific variables age, sex, university and education. We see that the coefficients are quite different for the distinct modes. The variables price and time must again be interpreted as the price (time) of a transport mode as individual i perceives it. We see that all considered modes have an negative coefficient of price as well as time compared to the reference category.

	Private		Public		
Variable	Own Car	Car Passenger	Bus	Train	Underground
Constant	3.257 *** (0.522)	0.618 (0.747)	4.958 *** (0.452)	0.095 (0.507)	2.108 *** (0.432)
Age	−0.453 * (0.203)	−0.198 (0.280)	0.286 . (0.151)	0.145 (0.179)	0.189 (0.159)
Sex	−0.838 *** (0.202	−0.566 * (0.284)	0.111 (0.154)	0.286 (0.184)	−0.250 (0.162)
University	−0.054 (0.217)	−0.487 (0.306)	0.135 (0.171)	0.663 ** (0.205)	−0.598 *** (0.177)
Education	0.067 (0.125)	0.194 (0.177)	−0.254 * (0.103)	0.160 (0.114)	−0.259 * (0.110)
Income-V.H	0.513 *** (0.145)	0.864 *** (0.190)	−0.059 (0.133)	1.003 *** (0.132)	−0.264 . (0.135)
Price	−0.551 *** (0.138)	−0.932 *** (0.245)	−2.606 *** (0.181)	−1.459 *** (0.173)	−0.099 (0.089)
Time	−0.089 *** (0.008)	−0.061 *** (0.011)	−0.029 *** (0.006)	−0.037 *** (0.007)	−0.022 *** (0.006)

Table 4.4.: *Coefficients of the parametric multinomial logit model. Standard deviation in brackets. (n =1780, LogL = -2 432) Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1*

The last model which we consider in this paper is the semiparametric multinomial logit model. The coefficients for the parametric part of the semiparametric approach are presented in Table 4.5. In this case, we model the influence of the income variable through the nonparametric functions m_k for each mode k.

	Private		Public		
Variable	Own Car	Car Passenger	Bus	Train	Underground
Constant	... (...)	... (...)	... (...)	... (...)	... (...)
Age	−0.451 * (0.176)	−0.207 . (0.262)	0.287 * (0.121)	0.119 . (0.150)	0.183 . (0.140)
Sex	−0.847 *** (0.175)	−0.527 * (0.267)	0.111 . (0.123)	0.288 * (0.155)	−0.265 * (0.142)
University	−0.060 (0.184)	−0.520 * (0.285)	0.134 . (0.135)	0.651 *** (0.171)	−0.608 *** (0.153)
Education	0.047 (0.107)	0.190 . (0.171)	−0.275 *** (0.082)	0.117 . (0.095)	−0.274 ** (0.097)
Income-V.H	... (...)	... (...)	... (...)	... (...)	... (...)
Price	−0.513 *** (0.126)	−0.939 *** (0.243)	−2.583 *** (0.168)	−1.404 *** (0.159)	−0.093 . (0.084)
Time	−0.088 *** (0.007)	−0.065 *** (0.011)	−0.028 *** (0.005)	−0.036 *** (0.006)	−0.022 *** (0.005)

Table 4.5.: *Coefficients for the semiparametric multinomial logit model. Standard deviation in brackets. (n =1780, LogL = -2 158.49) Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1*

The smoothing parameter for the nonparametric part was chosen on a grid of bandwidths between 0.2 and 1. For the presentations of the results, we selected $h = 0.4$ without changing the interpretation of the results. We can recognize almost the same values for the coefficients and standard deviations of the parametric covariates (age, sex, uni, education, price and time) as in the fully parametric model with mode-specific intercepts.

4.4.3. Marginal effects

However, in order to interpret the results of a multinomial logit model, it is more plausible to look at the marginal effects of the covariates. Especially, if one wants to sell the results to people with less statistical background. As our main objective is to give recommendations for public policy makers, we try to present results to confirm their decisions. The marginal effects of the characteristics are derived by differentiating the probability with respect to the corresponding covariate (Greene, 2006).

$$\delta_{ik} = \frac{\partial Pr_{ik}}{\partial x_i} = Pr_{ik} \left[\beta_k - \sum_{l=1}^{K} Pr_{il} \beta_l \right], \tag{4.14}$$

The marginal effects can be calculated from the parameter estimates, whereby every subvector of β enters every marginal effect. Table 4.6 shows these effects for the parametric approach. To interpret the marginal effects we consider for example, own car and sex and say: the parametric marginal effect is -0.072. Thus, if the variable sex changes from 0 (male) to 1 (female), the probability of choosing alternative 'own car' decreases by 7.2%. By means of the marginal effects we can give the following conclusions.

The effects of price for the alternatives bus and train are negative. We can conclude that especially these alternatives suffer from rising prices. The effects of time are small and very similar in absolute values, so that we should not overinterpret the differences. The effect of income is very diverse. On the one hand we have positive effects for the private options and the train, on the other hand we have negative effects for bus and underground. The semiparametric approach will allow more insights in these effects.

The results for the marginal effects of the semiparametric approach are given in Table 4.7. The nonparametric modeling allows us to give more meaningful interpretations of the marginal effects of the covariate income. Due to the graphical presentation of the marginal effects of income, we can on the one hand sustain the significance for the private transport modes as well as the alternative train. On the other hand, we can get an indication why

	Private		Public			
Variable	Own Car	Car Passenger	Bus	Train	Underground	Others
Constant	0.108 (0.134)	−0.059 (0.055)	0.648 (0.328)	−0.285 (0.197)	0.022 (0.153)	−0.434 (0.186)
Age	−0.050 (0.039)	−0.009 (0.007)	0.049 (0.027)	0.011 (0.018)	0.018 (0.016)	−0.019 (0.018)
Sex	−0.072 (0.056	−0.017 (0.014)	0.044 (0.031)	0.050 (0.041)	−0.025 (0.018)	0.019 (0.025)
University	−0.008 (0.017)	−0.021 (0.023)	0.025 (0.028)	0.080 (0.049)	−0.086 (0.035)	0.011 (0.038)
Education	0.012 (0.010)	0.009 (0.006)	−0.043 (0.023)	0.029 (0.016)	−0.027 (0.016)	0.020 (0.017)
Income-V.H	0.025 (0.025)	0.021 (0.015)	−0.062 (0.048)	0.098 (0.053)	−0.057 (0.024)	−0.025 (0.033)
Price	0.057 (0.068)	0.007 (0.017)	−0.348 (0.177)	−0.033 (0.059)	0.121 (0.075)	0.196 (0.097)
Time	−0.005 (0.004)	−0.001 (0.001)	0.001 (0.002)	0.001 (0.002)	0.001 (0.002)	0.006 (0.002)

Table 4.6.: *Marginal effect of the parametric multinomial logit model. Standard deviation in brackets.*

the effects of income on bus and underground were insignificant in the parametric model. Both effects show an almost linear relationship for lower income groups but an conspicuous bump in the middle-class and a substantial valley for well-off people. Thus, the group of rich people do not favor the alternatives bus and underground.

	Private		Public			
Variable	Own Car	Car Passenger	Bus	Train	Underground	Others
Constant	... (...)	... (...)	... (...)	... (...)	... (...)	... (...)
Age	−0.047 (0.046)	−0.008 (0.008)	0.087 (0.060)	0.018 (0.014)	0.030 (0.018)	0.003 (0.002)
Sex	−0.069 (0.067)	−0.012 (0.014)	0.093 (0.064)	0.068 (0.052)	−0.009 (0.005)	0.055 (0.041)
University	0.001 (0.001)	−0.018 (0.019)	0.059 (0.040)	0.099 (0.075)	−0.084 (0.050)	0.018 (0.013)
Education	0.009 (0.008)	0.009 (0.009)	−0.071 (0.049)	0.021 (0.016)	−0.037 (0.022)	0.009 (0.006)
Income-V.H	... (...)	... (...)	... (...)	... (...)	... (...)	... (...)
Price	0.044 (0.043)	−0.001 (0.001)	−0.486 (0.332)	−0.067 (0.051)	0.129 (0.077)	0.246 (0.181)
Time	−0.005 (0.007)	−0.001 (0.001)	0.003 (0.002)	0.001 (0.001)	0.003 (0.002)	0.011 (0.008)

Table 4.7.: *Marginal effects of the semiparametric multinomial logit model. Standard deviation in brackets.*

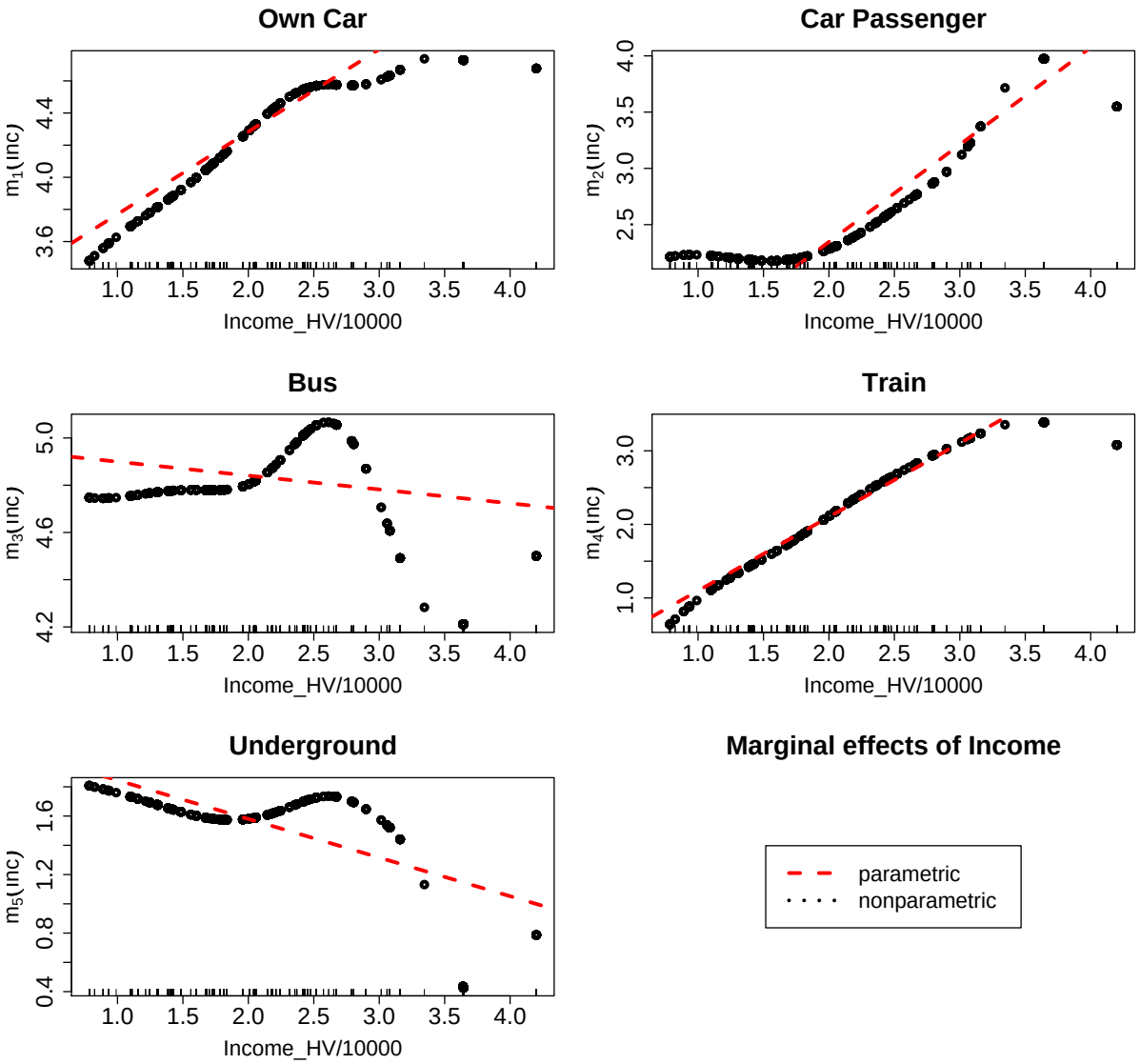

Figure 4.2.: *Marginal effect of Income* $m_k(Inc)$ *with* $k = 1, ..., 5$ *for a semiparametric multinomial logit model.*

For the purpose of identifiability, intercepts are not explicitly given in the proposed semiparametric model – the functions m_k already account for any intercept effects. Such intercept effects describe the unexplained heterogeneity over modes. Thus, it sometimes may be of interest to find an indicator that can substitute and in this way explain part of the m_k. In order to instrument this, we consider the mode-specific variable frequency (= minutes between services), given in Table 4.8.

After the model given in (4.13) has been fitted, we perform the following additional linear

Transport modes	Car	Pass	Bus	Train	Under-ground	Others
minutes between two services	0	5	15	7	20	25

Table 4.8.: *Values of frequency for the different transport modes.*

regression to determine the influence of the mode-specific variable frequency:

$$\sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot m_k(\text{Inc}) = \gamma_0 + \gamma_1 \cdot \sum_{k=1}^{K} \mathbb{1}_{\{y_i=k\}} \cdot \text{Freq}_k \tag{4.15}$$

The estimated coefficient of frequency is given by $\hat{\gamma_1} = -0.096$ (with standard error 0.004, thus highly significant). The estimated intercept, $\hat{\gamma_0} = 3.121(0.066)$, is highly significant, too. This reflects the non-explained 'mode' (intercept) effect. The interpretation of this effect is the same as in the parametric model with the mode-specific effect frequency. The time between two services has a negative effect on taking the corresponding transport mode.

4.5. Conclusions

The main aim of this study is to analyze the influence of socio-economic as well as mode- and individual-specific covariates on the transport demand function for university students in the Bilbao area. The prior objective is methodological, with the view of comparing the results achieved through different estimation methods (parametric and semiparametric). All of this, in terms of checking if it changes the recommendations for public policy makers.

If we compare the results of the parametric and semiparametric approach, we can conclude that:

- In the binary logit model, the estimates are very similar and the signs of the coefficients are as expected.
- For the multinomial logit model, the results between the parametric and semiparametric approach are distinct.
- For the income variable as well as price, the marginal effects are very different for underground and bus depending on the estimation method.
- Contrary, for the socio-economic covariates the marginal effects in both approaches are almost the same.

Therefore, we point to the differences between both approaches for the multinomial model. We model the level of utility that individuals (i) can achieve using a specific mode of transport (car, train, bus, underground) in a multinomial model. Thereby, we can identify the effects in the characteristics given by various covariates on the choice of mode of transport. We outline the enhancement based on a semiparametric approach, which could be further advanced by including more continuous covariates.

We analyze the formulated hypothesis, about the importance of the methodological approach, in order to obtain results to guide decisions of public policy makers. All of this, with the aim of capturing public transport passengers and improve the transport supply. If the estimated marginal effects, associated with an explanatory variable are very different, depending on the use of parametric or semiparametric approach, the transport policy recommendations would change in each case. For example, in the semiparametric approach, the sensitivity of transport demand to changes in price is much higher than in the parametric approach, and this affects the potential efficacy of this theoretical tool to replace private car uses with public transport modes. In our case, the semiparametric approach suggests that reducing the price of public transport would be more efficient than estimated by a parametric approach.

Further research projects could go along with this project. It is possible to extend the model to a nested logit approach with nonparametric modeled covariates or gather informations for a more comprehensive data base with different price, time and frequency values for each individual for every mode. Then, an individual and mode-specific effect could be identified and not only an individual perception of the different characteristics.

4.6. References

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

Bilbao Ubillos, J., Fernández-Sainz, A. (2004). The influence of quality and price on the demand for urban transport: the case of university students. *Transportation Research Part A,* **38**: 607–614.

Choo, S. and Mokhtarian, L. (2004). What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. *Transportation Research Part A: Policy and practice*, **38-3**: 201–222.

Domencich, T. and McFadden, D. (1975). *Urban Travel Demand: A Behavioral Analysis*. North Holland, Amsterdam.

Greene, W.H. (2008). Econometric analysis. *Pearson international (ed.), Prentice Hall.*

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004b). Nonparametric and Semiparametric Models. *Springer Series in Statistics, Berlin.*

Hensher, D.A. (2001). The valuation of commuter travel time savings for car drivers: evaluating alternative model specifications. *Transportation,* **28**: 101–118.

Huang, J.-C. and Nychka, D. (2000). A Nonparametric Multiple Choice Model within the Random Utility Framework. *Journal of Econometrics,* **2**: 207–225.

Kneib, T., Baumgartner, B. and Steiner, W.J. (2007). Semiparametric Multinomial Logit Models for Analysing Consumer Choice Behaviour. *Advances in Statistical Analysis*, **91/3**: 225–244.

Krinsky, I and Robb, A. (1986). On approximating the statistical properties of elasticities. *Review of Economics and Statistics,* **68**: 715–719.

Maddala, G.S. (1983). Limited-dependent and qualitative variables in econometrics. *Cambridge University Press.*

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In *Frontiers in Econometrics*, Zarembka, P. (ed.): 105–142.

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlquist et. al. (eds), *Spatial Interaction Theory and Residential Location*, 75–96. Amsterdam: North Holland.

McFadden, D. (1981). Econometric Models of probabilistic Choice. In C. Manski

and D. McFadden (eds), *Structural Analysis of Discrete Data: with Econometric Applications*, Cambridge, Mass.: M.I.T. Press.

Müller, M. (2001). Estimation and testing in generalized partial linear models - A comparative study. *Statistics and Computing*, **11/4**: 299–309.

Nguyen, K.P. (1999). Demand, supply, and pricing in urban road transport: the case of Ho Chi Minh City, Vietnam. *Research in Transportation economics,* **5**: 107–154.

Rodríguez-Póo, J. M., Sperlich, S. and Vieu, P. (2003). Semiparametric Estimation of Weak and Strong Separable Models. *Econometric Theory*, **19/6**: 1008–1039.

Severini, T.A. and Wong, W.H. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of Statistics*, **20/4**: 1768–1802.

Tutz, G. and Scholz, T. (2004). Semiparametric Modeling of Multicategorical Data. *Journal of Statistical Computation and Simulation*, **74/3**: 183–200.

5. Summary and Outlook

The main conclusions of the three projects are already given in the end of each chapter. Thus, opinions expressed in this chapter reflect the view of the author and a critical examination of the findings of this thesis.

The development of nonparametric methods was very fast and widespread in the recent past. A lot of theoretical considerations, e.g. about optimal bandwidth choice, asymptotics or convergence of nonparametric methods, as well as issues of applications in miscellaneous research areas have been published.

This thesis can be divided into two main parts. First, a review about a rather theoretical aspect of nonparametric modeling. Many authors pay a lot of attention on the choice of an optimal bandwidth. They have proposed different approaches to calculate optimal bandwidth considering various error measures. It is worth comparing all these bandwidth selection methods and to give an update of all existing method in a research report. Such an exhaustive report as presented in this thesis has not been written for a long time and hence an update about the state of the art was necessary. This essay contributes to the ongoing discussion and provides a basis for further discussion about optimal bandwidth choice.

However, although a lot of measures exist, an overall optimal choice cannot be given. Due to this fact, practitioners with some experience in nonparametric applications choose the corresponding bandwidth using well implemented standard routines or just by eye. This method is quite promising and in most cases adequate, especially in applied research.

In the second part of this thesis a semiparametric model for categorical data is proposed. Modeling categorical data is quite common in a lot of applications in economic and political science. However, most authors restrict their research to a parametric model or univariate approaches in nonparametric modeling. The model proposed in this thesis is an extension of a generalized partial linear model for the multinomial logit case. The model formulation is based on a likelihood approach whereby it benefits from the well-known mathematical properties. The model is very flexible and captures nonlinearities as well as interactions between the covariates. Three-dimensional plots of the probabilities allow meaningful interpretations, even for people with less statistical background.

As the model allows considering individual- and mode-specific variables, it covers a wide range of data problems. In this thesis, we consider two different applications. Firstly, the problem of voter profiling in a multi-party system like Germany. Identifying voters is a very heplful tool for political parties. In this context, the semiparametric model overcomes many deficiencies of purely descriptive statistics or parametric multinomial logit models which are very common in the realm of party affiliation. The results in the third chapter underline the need for more sophisticated modeling than the conventional MNL can offer.

The second application is concerned with demand of transport, in which individuals are faced with the choice between several modes of transport. As the aim of the first paper was to give insights about the data structure of voter profiling, this project must be seen in a more methodological context. The main goal was to show different approaches in modeling urban transport demand and to compare the results. Although, the data in this case are not entirely adequate, we see again that the semiparametric approach can give more insights. A valuable task would be to collect new, more comprehensive data to get results which are well interpretable and useful for drawing up sensible suggestions for policy makers.

Both applications point out the results of diverse real data problems in which the semiparametric approach is very suitable. Categorical data problems can be found in various research areas, like medicine, social science or economic science. Hence, a lot of further applications can be found, not least applications about consumer choice behaviour in marketing as well as decisions about residential or commercial locations for founding a business.

This wide field of applications underlines the relevance of the multinomial logit model. The deficiencies of previous approaches, in which parametric predictors as well as purely additive structures were proposed, accentuate the need of semiparametric modeling. The framework of generalized partial linear models offers well-known mathematical properties. Therefore, further research projects could be initiated, for which this thesis can serve as a suitable starting point.

Supplementary References

Büning, H. and Trenkler, G. (1994). Nichtparametrische statistische Methoden, *de Gruyter, Berlin.*

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). Nonparametric and Semiparametric Models, *Springer Series in Statistics, Berlin.*

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). Progress in data-based bandwidth selection for kernel density estimation, *Computational Statistics* **11**: 337-381.

Kotz, S. and Johnson, N. (1982). Encyclopedia of statistical sciences, *John Wiley and Sons, Toronto.*

Müller, M. (2001). Estimation and testing in gerneralized partial linear models - A comparative study. *Statistics and Computing*, **11/4**: 299–309.

Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and its Applications* **9 No. 1**: 141–142.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, **A 135**: 370–384.

Parzen, E. (1962). On estimation of a probability density function and mode, *Annals of Mathematical Statistics* **33 No. 3**: 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27 No. 3**: 832–837.

Upton, G.J.G. and Cook, I. (2008). Oxford dictionary of statistics, *Oxford University Press, Oxford.*

Watson, G.S. (1964). Smooth Regression Analysis, *Sankhyā: The Indian Journal of Statistics, Series A* **26 No. 4**: 359–372.

A. Appendix

This appendix gives the **R** code used to calculate the results presented in the three projects of this thesis. A short introduction is given at the beginning of the code for each chapter. The main parts of the code are marked with illustrative headings.

A.1. Programming code – Bandwidth Selection Methods for Kernel Density Estimation

This section gives the **R** code for analyzing the optimal bandwidth selection methods based on the calculation from the FORTRAN code. The FORTRAN code calculates the values for various bandwidths and the corresponding ise-values (intergrated squared errors). The implemented **R** code loads the results and calculates various measures which are presented in detail in the paper.

R Code

```
## Load the data
#setwd("...Daten/n=100/")  ##choose corresponding folder
#Load Fortran results
h.orig<-read.table("band10aP.dat")
xyd.orig<-read.table("band10P.dat")
ise.orig<-read.table("ise10P.dat")

names=c("Simple Normal Distribution","Two Normal Distributions",
"Three Normal Distributions","Three Gamma Distribution",
"Two Gamma Distributions","Simple Gamma Distributions")
```

```
## Sort and rearrange the data

n<-250 ##Anzahl Wiederholungen
Nm<-6     ##Anzahl betrachteter Modelle
m<-1      ##Auswahl der Dichtefunktion (1,6)
p<-9 ##Speicherplatz Spalte mit ISE-Werten
h<-h.orig[m,]
rows<-seq((m-1)*n+1,m*n)
xydo<-xyd.orig[rows,]
ise<-ise.orig[rows,]
xyd<-round(xydo,4)

##calculate the measures m1-m4
l<-dim(xyd)[2]
my<-mean(xyd)
sd<-sd(xyd)
s1.orig<-rbind(my,sd)
a<-1:(l/2)
b<-(l/2+1):l
s1<-s1.orig[,a]
s2<-s1.orig[,b]

##calculate the measure m5-m7
Avih<-rep(NA,(p+3))
L2ih<-rep(NA,(p+3))
L1ih<-rep(NA,(p+3))
for(i in 1:(p+3)){
Avih[i]<-mean(xyd[,i]-xyd[,p])
L2ih[i]<-mean((xyd[,i]-xyd[,p])^2)
L1ih[i]<-mean(abs(xyd[,i]-xyd[,p]))}

##calculates the measures m8-m10

Avise<-rep(NA,(p+3))
L2ise<-rep(NA,(p+3))
L1ise<-rep(NA,(p+3))
```

```
for(i in 1:(p+3)){
Avise[i]<-mean(xyd[,i+(p+3)]-xyd[,l/2+p])
L2ise[i]<-mean((xyd[,i+(p+3)]-xyd[,l/2+p])^2)
L1ise[i]<-mean(abs(xyd[,i+(p+3)]-xyd[,l/2+p]))}

##data output

aus1 <- rbind(s1, s2, Avih, L2ih, L1ih, Avise, L2ise, L1ise)

for (i in 1:10){
for (j in 1:12){
if(abs(aus1[i,j])>0.0001) aus1[i,j]<-round(aus1[i,j], digits = 4)
else aus1[i,j]<-round(aus1[i,j], digits = 10)}}

meas<-c("Criteria","CV","MCV","OSCV(L)","OSCV(R)",
"Stab","refPI","SBG","SBE","ISE","Mix 2/1","Mix 1/2", "Mix 1/1")
meth<-c("$m_1$","$m_2$","$m_3$","$m_4$",
"$m_5$","$m_6$","$m_7$","$m_8$","$m_9$","$m_{10}$")

aus1 <- cbind(meth,aus1)
aus1 <- rbind(meas,aus1)

##save the output in "out10.txt"
#write(t(aus1),ncolumns=13,append=TRUE, sep="&", file = ".../out10.txt")
```

A.2. Programming code – Semiparametric voter profiling in a multi-party system

This section gives the **R** code for analyzing voter profiles in a multi-party system like Germany. The main part of the program is the implementation of the iterative estimation procedure by means of the function `multgplm()`. The results at the end of each part of the code are saved in an **R.data** file. Therefore, all subsections can be directly used in **R** once the corresponding folder is chosen.

R Code – Data generation; SOEP-Data

This subsection shows the **R** code for reading in the data from the GSOEP survey [1] and preparing the data for the usage in the function `multgplm()`.

```
memory.limit(4095)
#####Import SOEP-Data
#setwd("...Daten/soep_Daten")   ##choose corresponding folder

library(foreign)  ##read foreign data types
  read.dta("wp.dta")->wp
  read.dta("tp.dta")->tp
  read.dta("wh.dta")->wh
soep_data<-merge(wp,wh,by="hhnrakt")
soep_data<-merge(soep_data,tp,by="persnr")

##Res. before reunification
soep2<-soep_data[,c("tp121")]
#Party, HH-net-income, sex, year of birth
soep<-soep_data[,c("wp12001","wh5101","wp12401","wp12402")]
##soep[,4] year of birth -> age
soep[,4]<-2007-soep[,4]
#combine the different items  ## n=18075
soep<-cbind(soep,rep(0,length(soep[,1])),
rep(0,length(soep[,1])),soep2)
which(soep[,1]!="Does not apply")->ind ; soep<-soep[ind,]
```

[1] SOEP Wave Report 1-2008, DIW Berlin, 2008

```
##9.288 do not support any party -> out
which(soep[,1]=="CDU [Christian Democratic Union]")->ind
soep[ind,6]<-5
which(soep[,1]=="CSU [Christian Social Union]")->ind ; soep[ind,6]<-5
which(soep[,1]=="SPD [Social Democratic Party of Germany]")->ind
soep[ind,6]<-1
which(soep[,1]=="Alliance '90/Greens")->ind ; soep[ind,6]<-2
which(soep[,1]=="Left Party.PDS")->ind  ; soep[ind,6]<-3
which(soep[,1]=="FDP [Free Democratic Party]")->ind ; soep[ind,6]<-4
which(soep[,6]!=0)->ind  ; soep<-soep[ind,]
#376 support other parties -> out <5%
which(soep[,7]!="Abroad")->ind  ; soep<-soep[ind,]
##227 live in foreign countries before reunification -> out
which(soep[,2]>0)->ind ; soep<-soep[ind,]
##383 made no declaration about their income ->out
which(soep[,2]>400)->ind ; soep<-soep[ind,]
##32 person earn less than 400 Euro -> out
which(soep[,2]<15000)->ind ; soep<-soep[ind,]
##32 person earn more than 15000 Euro -> out
which(soep[,2]<10000)->ind ; soep<-soep[ind,]
##another 77 person earn more than 10000 Euro -> out
which(soep[,4]<90)->ind ; soep<-soep[ind,]
##41 person older than 90 -> out
which(soep[,4]<80)->ind ; soep<-soep[ind,]
##384 person older than 80 -> out
which(soep[,4]<70)->ind ; soep<-soep[ind,]
##1026 person older than 70 -> out
which(soep[,4]<65)->ind ; soep<-soep[ind,]
##866 person older than 65 -> out
which(soep[,3]=="Female")->ind ; soep[ind,5]<-1
soep<-cbind(soep,rep(0,length(soep[,1])))
which(soep[,7]=="GDR (including East Berlin)")->ind ; soep[ind,8]<-1
n<-soep.len<-dim(soep)[1]      ##n=5343 Rest
sample(1:soep.len,size=n,replace=FALSE)->ind ; data<-soep[ind,]
names(data)<-c("Party","Income","Sex","Age",
"Sex_D","Mode","East","East_D")
#save.image("Soep_HH5343.Rdata")
```

R Code – Estimation procedure

This subsection gives the **R** code of the estimation procedure of the semiparametric multinomial logit model. The code is divided into several functions. The description, usage, arguments and the output of the corresponding functions is given.

R function: `logl()`

Description

Calculates the Log-Likelihood as well as the first and the second derivative w.r.t. the predictor η for the corresponding mode k.

Usage

```
logl(theta,x,y,m,k)
```

Arguments

theta	$p \times k$ matrix, coefficients of the parametric part of the semiparametric multinomial logit model or suitable starting values
x	$n \times p$ matrix, covariates which are supposed to be model parametrically
y	$n \times 1$ matrix, dependent variable
m	$n \times k$ matrix, values of the nonparametric functions of the semiparametric multinomial logit model or suitable starting values
k	integer, the corresponding mode

Output

ll	$n \times 3$-matrix, values for the likelihood as well as the first and the second derivative of the semiparametric multinomial logit model for the corresponding mode k.

Details

```
logl<-function(theta,x,y,m,k)
{Ind<-ifelse(y==k,1,0)
 eta<-x%*%theta+m ;etanum<-exp(eta[,k])
 denom<-apply(exp(eta),1,sum)
 li<-Ind*log(etanum/denom)
 li.<-(Ind-etanum/denom)
 li..<-(-(etanum*denom-etanum^2)/(denom^2))
 ll<-cbind(li,li.,li..); return(ll)}
```

R function: `kern()`

DESCRIPTION

Calculates the kernel function for the estimation of the nonparametric functions in the semiparametric multinomial logit model.

USAGE

```
kern(T,h,Kern="gaussian",n)
```

ARGUMENTS

T	$n \times q$ matrix, covariates which are supposed to be model nonparametrically
h	scalar or $1 \times q$ matrix, bandwidth
Kern	name of the kernel function (currently only "gaussian" and "quartic" are supported)
n	integer, length of the data

OUTPUT

kern	$n \times n$-matrix, kernel function

DETAILS

```
kern<-function(T,h,Kern="gaussian",n)
{
u<-matrix(rep(0,n*n),nrow=n)
for (i in 1:n){u[i,]<-(T[i]-T)/h}
if(missing(Kern)) Kern<-"gaussian"
if(Kern=="quartic"){
kern<-ifelse((abs(u)<=1),15/16*(1-u^2)^2,0)}
if(Kern=="gaussian"){
kern<-1/sqrt(2*pi)*exp(-1/2*u^2)}
return(kern)}
```

R function: `multgplm()`

DESCRIPTION

Fits a semiparametric multinomial logit model to the data with aid of a smoothed and profile likelihood. The estimation is based on an iterative Newton-Raphson algorithm.

USAGE

```
multgplm(y=y,x=x,T=T,b0=b0,m0=m0,max.iter=20,modes=modes,
h=h,Kern="gaussian",...)
```

ARGUMENTS

y	$n \times 1$ matrix, dependent variable
x	$n \times p$ matrix, covariates which are supposed to be model parametrically
T	$n \times q$ matrix, covariates which are supposed to be model nonparametrically
b0	$p \times k$ matrix, suitable starting values for the coefficients of the parametric part of the model
m	$n \times k$ matrix, suitable starting values for the nonparametric functions of the model
max.iter	integer, maximum number of iterations
h	scalar or $1 \times q$ matrix, bandwidth
Kern	name of the kernel function (currently only "gaussian" and "quartic" are supported)

OUTPUT

b	$p \times k$ matrix, estimated coefficients of the parametric part of the semiparametric multinomial logit model
m	$n \times k$ matrix, estimated values of the nonparametric functions of the semiparametric multinomial logit model
Time	integer, duration of the estimation procedure
se_b	standard errors of the estimated coefficients b
t_b	t-values of the estimated coefficients b
LogL	Log-Likelihood for the final estimates
y	$n \times 1$ matrix, dependent variable
x	$n \times p$ matrix, covariates which are supposed to be model parametrically
T	$n \times q$ matrix, covariates which are supposed to be model nonparametrically
sd1	integer, standard-deviation of the first nonparametrically modeled covariate
sd2	integer, standard-deviation of the second nonparametrically modeled covariate
h	scalar or $1 \times q$ matrix, bandwidth

Details

```
multgplm<-function(y=y,x=x,T=T,
b0=b0,m0=m0,max.iter=20,modes=modes,h=h,Kern="gaussian",...){
#Preparing the data
y <- as.matrix(y)
n<-dim(y)[1]
x<-as.matrix(x)
betadim<-dim(x)[2]
T <- as.matrix(T)
Tdim<-dim(T)[2]
##Ordering of the data (for graphical purposes)
or1 <- order(T[, 1])
if (Tdim==2) or2 <- order(T[, 2])
##correction of the standard-errors (for estimation)
sd1<-sd(T[,1])
if (Tdim==2) sd2<-sd(T[,2])
T.1<-T[,1]/sd1
if (Tdim==2) T.2<-T[,2]/sd2
if (Tdim==2) T<-cbind(T.1,T.2)
##check for missing values and setting to defaults
if(missing(Kern)) Kern<-"gaussian"
if(missing(h)) h<-h.select(T,y)
if(length(h)==1) h<-rep(h,2)
if(missing(modes)) modes<-nlevels(factor(y))
if(modes!=nlevels(factor(y))){stop("number of modes incorrect")}
if(missing(max.iter)) max.iter<-20

##Setting of suitable starting values
b<-matrix(rep(0,modes*betadim),ncol=modes)
if(!missing(b0)) b[,1:(modes-1)]<-b0 ;  print(b)
m<-matrix(rep(0,n*modes),ncol=modes)
if(!missing(m0)) {
for (k in 1:(modes-1)){m[,k]<-m0[1,k]*T[,1]+m0[2,k]*T[,2]}}
##Calculation of the kernel function (using kern())
Kh1<-kern(T[,1],h[1],Kern,n)
if (Tdim==2) Kh2<-kern(T[,2],h[2],Kern,n)
Kh<-Kh1*Kh2
```

```
##Preparing of empty vectors and matrices
li.mat<-matrix(rep(0,n^2),nrow=n)
li..mat<-matrix(rep(0,n^2),nrow=n)
xtilde<-matrix(rep(0,n*dim(b)[1]),nrow=n)
Bsummand<-array(rep(0,n*betadim^2),c(betadim,betadim,n))
B_k<-rep(0,betadim*betadim*(modes-1))
dim(B_k)=c(betadim,betadim,(modes-1))
mnew<-rep(0,n)
biterations<-rep(0,(modes-1)*max.iter*betadim)
dim(biterations)=c(betadim,modes-1,max.iter)
LogL<-c(rep(0,(modes-1)))

##########################################
## Start iterative estimation procedure ##
##########################################

start<-Sys.time()
it<-0
while (it<max.iter)
{
it<-it+1
for (k in 1:(modes-1))
 {
ll<-logl(b,x,y,m,k)
li.<-ll[,2]      # li. for mode k
li..<-ll[,3]     # li.. for mode k

# Programming of the profile Likelihood as ixj matrix
for (j in 1:n) # li.mat for mode k
{li.mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,2]}
for (j in 1:n) # li..mat for mode k
{li..mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,3]}

for (p in 1:n)
{za<-t(apply((li..mat[,p]*Kh[,p])*x,2,sum))
ne<-sum(li..mat[,p]*Kh[,p])
xtilde[p,]<-x[p,]-za/ne}
```

```
B=matrix(rep(0,betadim^2),nrow=betadim) ##updating b (step 2)
for (p in 1:n) {
Bsummand[,,p]<-as.vector(li..)[p]*(xtilde[p,]%*%t(xtilde[p,]))
B<-B+Bsummand[,,p]}
B_k[,,k]<-B
bnew<-b[,k]-solve(B)%*%apply(as.vector(li.)*xtilde,2,sum)

for (p in 1:n) ##updating m (step 3)
{mza<-sum(li.mat[,p]*Kh[,p])
mne<-sum(li..mat[,p]*Kh[,p])
mnew[p]<-m[p,k]-mza/mne}

biterations[,k,it]<-b[,k]<-bnew
m[,k]<-mnew
}
print(it)
print(b)

#Abort criterion
diff<-matrix(rep(1,betadim*(modes-1)),nrow=betadim)
if(it>1) {for(i in 1:betadim){ for(j in 1:(modes-1)){
diff[i,j]<-abs(biterations[i,j,it]-biterations[i,j,(it-1)]) }}}
print(sum(diff))
if(sum(diff)<0.01) it<-max.iter}

##Output values
se_b<-matrix(rep(0,betadim*(modes-1)),nrow=betadim)
for(z in 1:(modes-1)){
se_b[,z]<-diag(sqrt(solve(-B_k[,,z])))}##St.fehler für b_k
t_b<-b[1:(modes-1)]/se_b           ##t-values für b_k
LogL[k]<-sum(ll[,1])
end<-Sys.time()
Time<-end-start
return(list(b=b,m=m,Time=Time,se_b=se_b,t_b=t_b,
LogL=LogL,y=y,x=x,T=T,sd1=sd1,sd2=sd2,h=h))}
```

```
##########################
## Running the function ##
##########################

#do not run:
#setwd(".../Daten/soep_Daten/")    ##choose corresponding folder
#load("Soep_HH5343.Rdata")   #read in the data
#mult_mod<-multgplm(data[,6],x=cbind(data[,5],data[,8]),
#T=cbind(log(data[,2]),data[,4]),h=0.6,max.iter=10)
#attach(mult_mod)
#save.image("Soep_estHH5343.Rdata")  ##save estimation results
```

R Code – Graphical presentation

This subsection gives the **R** code for producing the three-dimensional plots and the plots of the additive decomposition presented in the paper.

Bivariate three-dimensional plots

The three-dimensional plots are generated by using the `rgl`-package proposed by Daniel Adler and Duncan Murdoch (2009).

```
###3D-Graphics
rm(list=ls())
memory.limit(4092)
#setwd(".../Daten/soep_Daten/")    ##choose corresponding folder
load("Soep_estHH5343.Rdata") ##load estimation results
attach(mult_mod)
library(rgl)
f<-female<-1;o<-ost<-0 #Sex and East? 1-woman,0-man resp. 1-east,0-west

##Calculate the probabilities
b_female<-round(b[1,],5)
b_ost<-round(b[2,],5)
len_m<-length(y)
eta<-matrix(rep(NA,len_m*5),nrow=len_m)
for (k in 1:len_m)
{eta[k,]<-(b_female*f+b_ost*o+(m[k,]))}
Prob<-matrix(rep(NA,len_m*5),nrow=len_m)
for (k in 1:len_m)
{Pro<-exp(eta[k,])/(sum(exp(eta[k,])))
Prob[k,]<-Pro}

##Package for bivariate Interpolation and Smooth Surface Fitting
##for Irregularly Distributed Data Points (Akima et.al., 2009)
library(akima)
min_inc<-log(400); max_inc<-log(10000)   ##limits for income
min_age<-21; max_age<-65                ##limits for age
z1<-akima.z1 <- interp(T[,1], T[,2], Prob[,1],
                  xo=seq(min_inc/sd1,max_inc/sd1, length = 202),
```

```
                    yo=seq(min_age/sd2, max_age/sd2, length = 202),
                    linear = FALSE, extrap = TRUE,duplicate="strip")
z2<-akima.z2 <- interp(T[,1], T[,2], Prob[,2],
                    xo=seq(min_inc/sd1,max_inc/sd1, length = 202),
                    yo=seq(min_age/sd2, max_age/sd2, length = 202),
                    linear = FALSE, extrap = TRUE,duplicate="strip")

z3<-akima.z3 <- interp(T[,1], T[,2], Prob[,3],
                    xo=seq(min_inc/sd1,max_inc/sd1, length = 202),
                    yo=seq(min_age/sd2, max_age/sd2, length = 202),
                    linear = FALSE, extrap = TRUE,duplicate="strip")
z4<-akima.z4 <- interp(T[,1], T[,2], Prob[,4],
                    xo=seq(min_inc/sd1,max_inc/sd1, length = 202),
                    yo=seq(min_age/sd2, max_age/sd2, length = 202),
                    linear = FALSE, extrap = TRUE,duplicate="strip")
z5<-akima.z5 <- interp(T[,1], T[,2], Prob[,5],
                    xo=seq(min_inc/sd1,max_inc/sd1, length = 202),
                    yo=seq(min_age/sd2, max_age/sd2, length = 202),
                    linear = FALSE, extrap = TRUE,duplicate="strip")

#setwd(".../images/")     ##choose corresponding folder
open3d();bg3d("white")
surface3d( (akima.z1$x*sd1),akima.z1$y*sd2,akima.z1$z,color="red",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z1$x*sd1)[0:25*8+1],(akima.z1$y*sd2)[0:25*8+1],
akima.z1$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z1$x*sd1)[0:25*8+1],(akima.z1$y*sd2)[0:25*8+1],
akima.z1$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1);axes3d()
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z1$z)-0.001,size=1)
#rgl.snapshot("SPD_Frau_West_Paper.ps")
open3d();bg3d("white")
```

```
surface3d( (akima.z2$x*sd1),akima.z2$y*sd2,akima.z2$z,color="green",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z2$x*sd1)[0:25*8+1],(akima.z2$y*sd2)[0:25*8+1],
akima.z2$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z2$x*sd1)[0:25*8+1],(akima.z2$y*sd2)[0:25*8+1],
akima.z2$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1);axes3d()
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z2$z)-0.001,size=1)
#rgl.snapshot("Grüne_Frau_West_Paper.ps")

open3d();bg3d("white")
surface3d( (akima.z3$x*sd1),akima.z3$y*sd2,akima.z3$z,color="violet",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z3$x*sd1)[0:25*8+1],(akima.z3$y*sd2)[0:25*8+1],
akima.z3$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z3$x*sd1)[0:25*8+1],(akima.z3$y*sd2)[0:25*8+1],
akima.z3$z[0:25*8+1,0:25*8+1], alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1);axes3d()
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z3$z)-0.001,size=1)
#rgl.snapshot("Linke_Frau_West_Paper.ps")

open3d();bg3d("white")
surface3d( (akima.z4$x*sd1),akima.z4$y*sd2,akima.z4$z,color="yellow3",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z4$x*sd1)[0:25*8+1],(akima.z4$y*sd2)[0:25*8+1],
akima.z4$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z4$x*sd1)[0:25*8+1],(akima.z4$y*sd2)[0:25*8+1],
```

```
akima.z4$z[0:25*8+1,0:25*8+1], alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1);axes3d()
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z4$z)-0.001,size=1)
#rgl.snapshot("FDP_Frau_West_Paper.ps")

open3d();bg3d("white")
surface3d( (akima.z5$x*sd1),akima.z5$y*sd2,akima.z5$z,color="black",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z5$x*sd1)[0:25*8+1],(akima.z5$y*sd2)[0:25*8+1],
akima.z5$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z5$x*sd1)[0:25*8+1],(akima.z5$y*sd2)[0:25*8+1],
akima.z5$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1);axes3d()
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z5$z)-0.001,size=1)
#rgl.snapshot("CDU_Frau_West_Paper.ps")
```

Additive three-dimensional plots

The three-dimensional plots are generated by using the `rgl`-package proposed by Daniel Adler and Duncan Murdoch (2009).

The additive decomposition is done with the `mgcv::gam`-package from Simon Wood (2006)

```
rm(list=ls())
#setwd(".../Daten/soep_Daten/")  ##choose corresponding folder
load("Soep_estHH5343.Rdata")
attach(mult_mod)
library(mgcv)    #load mgcv-package
log_income<-log(T[,1])
age<-T[,2]
or1<-order(log_income)
or2<-order(age)
n<-length(y)

##Die CDU/CSU
ym<-m[,5]
gam(ym~s(log_income)+s(age))->mod5
pred5<-predict(mod5,type="terms")

##Die SPD
ym<-m[,1]
gam(ym~s(log_income)+s(age))->mod1
pred1<-predict(mod1,type="terms")

##Die Grünen
ym<-m[,2]
gam(ym~s(log_income)+s(age))->mod2
pred2<-predict(mod2,type="terms")

##Die Linke
ym<-m[,3]
gam(ym~s(log_income)+s(age))->mod3
pred3<-predict(mod3,type="terms")
```

```
##Die FDP
ym<-m[,4]
gam(ym~s(log_income)+s(age))->mod4
pred4<-predict(mod4,type="terms")

##Calculation of additive probabilities
f<-female<-1
o<-ost<-0
b_female<-round(b[1,],5)
b_ost<-round(b[2,],5)
eta<-matrix(rep(NA,n*5),nrow=n)
for (k in 1:n)
{eta[k,1]<-(b_female[1]*female+b_ost[1]*ost+
pred1[k,1]+pred1[k,2]+mod1$coef[1])}
for (k in 1:n)
{eta[k,2]<-(b_female[2]*female+b_ost[2]*ost+
pred2[k,1]+pred2[k,2]+mod2$coef[1])}
for (k in 1:n)
{eta[k,3]<-(b_female[3]*female+b_ost[3]*ost+
pred3[k,1]+pred3[k,2]+mod3$coef[1])}
for (k in 1:n)
{eta[k,4]<-(b_female[4]*female+b_ost[4]*ost+
pred4[k,1]+pred4[k,2]+mod4$coef[1])}
for (k in 1:n)
{eta[k,5]<-(b_female[5]*female+b_ost[5]*ost+
pred5[k,1]+pred5[k,2]+mod5$coef[1])}

Prob<-matrix(rep(NA,n*5),nrow=n)
for (k in 1:n)
{Pro<-exp(eta[k,])/(sum(exp(eta[k,])))
Prob[k,]<-Pro}

##Package for bivariate Interpolation and Smooth Surface Fitting
##for Irregularly Distributed Data Points (Akima et.al., 2009)
library(akima)
min_inc<-log(400); max_inc<-log(10000)   ##limits for income
min_age<-21; max_age<-65                ##limits for age
```

```
akima.z1 <- interp(T[,1], T[,2], Prob[,1],
                   xo=seq(5.991/sd1,9/sd1, length = 202),
                   yo=seq(min(T[,2]), max(T[,2]), length = 202),
                   linear = FALSE, extrap = TRUE,duplicate="strip")

akima.z2 <- interp(T[,1], T[,2], Prob[,2],
                   xo=seq(5.991/sd1,9/sd1, length = 202),
                   yo=seq(min(T[,2]), max(T[,2]), length = 202),
                   linear = FALSE, extrap = TRUE,duplicate="strip")

akima.z3 <- interp(T[,1], T[,2],Prob[,3],
                   xo=seq(5.991/sd1,9/sd1, length = 202),
                   yo=seq(min(T[,2]), max(T[,2]), length = 202),
                   linear = FALSE, extrap = TRUE,duplicate="strip")


akima.z4 <- interp(T[,1], T[,2], Prob[,4],
                   xo=seq(5.991/sd1,9/sd1, length = 202),
                   yo=seq(min(T[,2]), max(T[,2]), length = 202),
                   linear = FALSE, extrap = TRUE,duplicate="strip")

akima.z5 <- interp(T[,1], T[,2], Prob[,5],
                   xo=seq(5.991/sd1,9/sd1, length = 202),
                  yo=seq(min(T[,2]), max(T[,2]), length = 202),
                   linear = FALSE, extrap = TRUE,duplicate="strip")

ind<-which(is.na(akima.z1$z))
z1<-akima.z1$z[-ind]
ind<-which(is.na(akima.z2$z))
z2<-akima.z2$z[-ind]
ind<-which(is.na(akima.z3$z))
z3<-akima.z3$z[-ind]
ind<-which(is.na(akima.z4$z))
z4<-akima.z4$z[-ind]
ind<-which(is.na(akima.z5$z))
z5<-akima.z5$z[-ind]
```

```
library(rgl)     #load package
#setwd(".../images/")     ##choose corresponding folder
open3d();bg3d("white")
par3d(windowRect=c(20,50,900,900))#

surface3d( (akima.z1$x*sd1),akima.z1$y*sd2,akima.z1$z,color="red",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z1$x*sd1)[0:25*8+1],(akima.z1$y*sd2)[0:25*8+1],
akima.z1$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z1$x*sd1)[0:25*8+1],(akima.z1$y*sd2)[0:25*8+1],
akima.z1$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1)
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z1)-0.001,size=1)
axes3d(labels=FALSE,tick=FALSE,xlab="",ylab="",zlab="",nticks=0)
 axis3d(c("x++"),cex=1.3)
 axis3d(c("y-+"),cex=1.3)
 axis3d(c("z++"),cex=1.3)
#rgl.snapshot("SPD_Frau_West_Add.ps")

open3d();bg3d("white")
surface3d( (akima.z2$x*sd1),akima.z2$y*sd2,akima.z2$z,color="green",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z2$x*sd1)[0:25*8+1],(akima.z2$y*sd2)[0:25*8+1],
akima.z2$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z2$x*sd1)[0:25*8+1],(akima.z2$y*sd2)[0:25*8+1],
akima.z2$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1)
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
```

```
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z2)-0.001,size=1)
axes3d(labels=FALSE,tick=FALSE,xlab="",ylab="",zlab="",nticks=0)
 axis3d(c("x-+"),cex=1.3)
 axis3d(c("y+-"),cex=1.3)
 axis3d(c("z+-"),cex=1.3)
#rgl.snapshot("Grüne_Frau_West_Add.ps")

open3d();bg3d("white")
surface3d( (akima.z3$x*sd1),akima.z3$y*sd2,akima.z3$z,color="violet",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z3$x*sd1)[0:25*8+1],(akima.z3$y*sd2)[0:25*8+1],
akima.z3$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z3$x*sd1)[0:25*8+1],(akima.z3$y*sd2)[0:25*8+1],
akima.z3$z[0:25*8+1,0:25*8+1], alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
lines3d(x=max(T[,1]*sd1),y=max(T[,2]*sd2),z=seq(0,0.5,len=100))
aspect3d(1,1,1)
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z3)-0.01,size=1)
axes3d(labels=FALSE,tick=FALSE,xlab="",ylab="",zlab="",nticks=0)
 axis3d(c("x++"),cex=1.3)
 axis3d(c("y+-"),cex=1.3)
 axis3d(c("z++"),cex=1.3)
#rgl.snapshot("Linke_Mann_Ost_Add.ps")

open3d();bg3d("white")
surface3d( (akima.z4$x*sd1),akima.z4$y*sd2,akima.z4$z,color="yellow3",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z4$x*sd1)[0:25*8+1],(akima.z4$y*sd2)[0:25*8+1],
akima.z4$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z4$x*sd1)[0:25*8+1],(akima.z4$y*sd2)[0:25*8+1],
akima.z4$z[0:25*8+1,0:25*8+1], alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
```

```
aspect3d(1,1,1)
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z4)-0.001,size=1)
axes3d(labels=FALSE,tick=FALSE,xlab="",ylab="",zlab="",nticks=0)
 axis3d(c("x-+"),cex=1.3)
 axis3d(c("y--"),cex=1.3)
 axis3d(c("z--"),cex=1.3)
#rgl.snapshot("FDP_Frau_West_Add.ps")

open3d();bg3d("white")
surface3d( (akima.z5$x*sd1),akima.z5$y*sd2,akima.z5$z,color="black",
alpha=c(0.3),front="fill",lit=F)
surface3d( (akima.z5$x*sd1)[0:25*8+1],(akima.z5$y*sd2)[0:25*8+1],
akima.z5$z[0:25*8+1,0:25*8+1],
alpha=c(0.3),front="line",back="cull",lwd=1.5)
surface3d( (akima.z5$x*sd1)[0:25*8+1],(akima.z5$y*sd2)[0:25*8+1],
akima.z5$z[0:25*8+1,0:25*8+1],
alpha=c(0.7),front="line",back="line")
view3d(userMatrix = rotationMatrix(70*pi/190, -1.2,-0.4,-0.95))
aspect3d(1,1,1)
ind<-which(T[,1]*sd1>=min_inc & T[,1]*sd1<=max_inc &
 T[,2]*sd2>=min_age & T[,2]*sd2<=max_age)
points3d((T[,1]*sd1)[ind],(T[,2]*sd2)[ind],min(z5)-0.001,size=1)
axes3d(labels=FALSE,tick=FALSE,xlab="",ylab="",zlab="",nticks=0)
 axis3d(c("x++"),cex=1.3)
 axis3d(c("y--"),cex=1.3)
 axis3d(c("z-+"),cex=1.3)
#rgl.snapshot("CDU_Frau_West_Add.ps")
```

Additive decomposition

The additive decomposition is done with the `mgcv::gam`-package from Simon Wood (2006)

```
##graphics for additive decomposition using mgcv::gam()
library(mgcv)     ##load package
rm(list=ls())
#setwd("...")     ##choose corresponding folder
load("Soep_estHH5343.Rdata")   #read the data
attach(mult_mod)   #attach model values (multinomial logit)

##limits for graphical purposes
min_inc<-log(400); max_inc<-log(10000)
min_age<-20; max_age<-65

#one picture for all (special R-layout)
nf <- layout(matrix(c(1:14), 7, 2, byrow=TRUE))  #14 pictures (7*2)
#1*2 for picture titles
layout.show(nf)
par(mar=c(0, 5, 1, 2), mgp=c(3, 1, 0))
x<-seq(min_inc,max_inc,len=100);y<-0+0*x
plot(x,y,xaxt="n",col="white",col.axis="white",
col.lab="white",bty="n",yaxt="n")
text((min_inc+max_inc)/2,-0.2,"log(income)",cex=2,lwd=2)
x<-seq(min_age,max_age,len=100);y<-0+0*x
plot(x,y,xaxt="n",col="white",col.axis="white",
col.lab="white",bty="n",yaxt="n")
text((min_age+max_age)/2,-0.2,"age",cex=2,lwd=2)
par(mar=c(0, 5, 0, 2), mgp=c(3, 1, 0))

#now 5*2 picture for the corresponding parties
##Die CDU/CSU
y<-m[,5]             #choose aditive decomposable variable
log_income<-T[,1]*sd1   #unscaled income
age<-T[,2]*sd2               #unscaled age
gam(y~s(log_income)+s(age))->mod5 #additive decomposition
plot(mod5,xlim=c(log(400),log(10000)),xpd=FALSE,pages=0,col=1,
shift= mod5$coef[1],xlab="",ylab="",xaxt="n",rug=FALSE,se=FALSE,lwd=3,
```

```
scale=0,cex.axis=1.3,cex.lab=1.6,select=1,yaxt="n")
axis(2,cex.axis=1.2)
plot(mod5,xlim=c(20,60),xpd=FALSE,pages=0,col=1,
shift= mod5$coef[1],xlab="",ylab="",xaxt="n",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=2,yaxt="n")
axis(2,cex.axis=1.2)

##Die SPD
y<-m[,1]
gam(y~s(log_income)+s(age))->mod1
plot(mod1,xlim=c(log(400),log(10000)),xpd=FALSE,pages=0,col="red",
shift= mod1$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=1,xaxt="n",yaxt="n")
axis(4,cex.axis=1.2)
plot(mod1,xlim=c(20,60),xpd=FALSE,pages=0,col="red",
shift= mod1$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=2,xaxt="n",yaxt="n")
axis(4,cex.axis=1.2)

##Die Grünen
y<-m[,2]
gam(y~s(log_income)+s(age))->mod2
plot(mod2,xlim=c(log(400),log(10000)),xpd=FALSE,pages=0,col="green",
shift= mod2$coef[1],xlab="",ylab=expression(m[k]),rug=FALSE,se=FALSE,
lwd=3,scale=0,cex.axis=1.3,cex.lab=1.5,select=1,xaxt="n",yaxt="n")
axis(2,cex.axis=1.2)
plot(mod2,xlim=c(20,60),xpd=FALSE,pages=0,col="green",
shift= mod2$coef[1],xlab="",ylab=expression(m[k]),rug=FALSE,se=FALSE,
lwd=3,scale=0,cex.axis=1.3,cex.lab=1.5,select=2,xaxt="n",yaxt="n")
axis(2,cex.axis=1.2)

##Die Linke
y<-m[,3]
gam(y~s(log_income)+s(age))->mod3
plot(mod3,xlim=c(log(400),log(10000)),xpd=FALSE,pages=0,col="violet",
shift= mod3$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,scale=0,
cex.axis=1.3,cex.lab=1.6,select=1,xaxt="n",yaxt="n")
```

```
 axis(4,cex.axis=1.2)
plot(mod3,xlim=c(20,60),xpd=FALSE,pages=0,col="violet",
shift= mod3$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=2,xaxt="n",yaxt="n")
axis(4,cex.axis=1.2)

##Die FDP
y<-m[,4]
gam(y~s(log_income)+s(age))->mod4
plot(mod4,xlim=c(log(400),log(10000)),xpd=FALSE,pages=0,col="yellow3",
shift= mod4$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=1,xaxt="n",yaxt="n")
axis(2,cex.axis=1.2)  #,xlim=c(log(400),log(10000))
plot(mod4,xlim=c(20,60),xpd=FALSE,pages=0,col="yellow3",
shift= mod4$coef[1],xlab="",ylab="",rug=FALSE,se=FALSE,lwd=3,
scale=0,cex.axis=1.3,cex.lab=1.6,select=2,xaxt="n",yaxt="n")
axis(2,cex.axis=1.2)


#1*2 for picture subtitles
x<-seq(min_inc,max_inc,len=100); y<-0+0*x
plot(x,y,xaxt="n",col="white",col.axis="white",
col.lab="white",bty="n",yaxt="n")
axis(1,xlab="log_income",at=seq(round(min_inc),trunc(max_inc),by=1),
labels=seq(round(min_inc),trunc(max_inc),by=1),pos=1,outer=TRUE)

x<-seq(min_age,max_age,len=100); y<-0+0*x
plot(x,y,xaxt="n",col="white",col.axis="white",
col.lab="white",bty="n",yaxt="n")
axis(1,xlab="age",at=seq(min_age,max_age,by=5),
labels=seq(min_age,max_age,by=5),pos=1,xpd=TRUE)
```

A.3. Programming code – A Semiparametric Model of Urban Transport Demand

R Code – Data generation; Bilbao-Data

This subsection shows the **R** code for reading in the data from the query filled in by students at the University in Bilbao. Also the preparation of the data for the usage in the function `multgplm()` is done.

```
##Bilbao Data
rm(list=ls())

library(VGAM) ; library(KernSmooth) ; library(sm);
library(mlogit);library(nnet)
memory.limit(2047)

set.seed(1)

#setwd(".../Daten/")     ##choose corresponding folder
library(foreign)
data<-read.csv("Bilbao.csv",sep=";",header=TRUE)
#EDAD DUM1 MVIAJ EDAD1 SEXO
names(data)<-c("Age","Priv_D","Mode","Age_D","Sex_D",
#UPV ESTP RENTA (Einkommen) FREC TIEMPO PRECIO AED (years of education)
"Uni_D","Income_Edu","Income_HV","Freq","Duration","Price","Edu_Parents")
attach(data)
head(data)
levels(as.factor(Price))
levels(as.factor(Duration))
levels(as.factor(Freq))
Income_Edu<-ifelse(Income_Edu==0,1,Income_Edu)
levels(as.factor(Income_Edu))
#save.image("Bilbao_Data.RData")
```

R Code – Estimation procedure

This subsection gives the **R** code of the estimation procedure of the semiparametric multinomial logit model in the case of the urban transport problem in Bilbao. The code is basically the same as in the second project. However, there are some slightly changes because of the slightly different data structure as well as sparse data. Hence, some additional iterations are used in order to assure convergence of the estimates. The presented code is directly applicable in **R**.

Details

```
##Fits a semiparametric multinomial logit model to the data with aid of a
##smoothed and profile likelihood. The estimation is based on an
##iterative Newton-Raphson algorithm.  ##Bilbao-Data
library(mlogit); library(MASS); library(sm)
multgplm<-function(y=y,x=x,T=T,
b0=b0,m0=m0,max.iter=20,modes=modes,h=h,Kern="gaussian",...){

#Preparing the data
y <- as.matrix(y)
n<-dim(y)[1]
x<-as.matrix(x)
betadim<-dim(x)[2]
T <- as.matrix(T)
Tdim<-dim(T)[2]

##Ordering of the data (for graphical purposes)
or1 <- order(T[, 1])
if (Tdim==2) or2 <- order(T[, 2])

##correction of the standard-errors (for estimation)
sd1<-sd(T[,1])
if (Tdim==2) sd2<-sd(T[,2])
T.1<-T[,1]/sd1
if (Tdim==2) T.2<-T[,2]/sd2
if (Tdim==2) T<-cbind(T.1,T.2)

##check for missing values and setting to defaults
if(missing(Kern)) Kern<-"gaussian"
```

```
if(missing(h)) h<-h.select(T,y)
if(length(h)==1) h<-rep(h,2)
if(missing(modes)) modes<-nlevels(factor(y))
if(modes!=nlevels(factor(y))){stop("number of modes incorrect")}
if(missing(max.iter)) max.iter<-20

##Setting of suitable starting values
b<-matrix(rep(0,modes*betadim),ncol=modes)
if(!missing(b0)) b[,1:(modes-1)]<-b0 ;  print(b)
m<-matrix(rep(0,n*modes),ncol=modes)
if(!missing(m0)) {
for (k in 1:(modes-1)){m[,k]<-m0[1,k]*T[,1]+m0[2,k]*T[,2]}}

##Calculation of the kernel function (using kern())
Kh<-Kh1<-kern(T[,1],h[1],Kern,n)
if (Tdim==2) {Kh2<-kern(T[,2],h[2],Kern,n)
Kh<-Kh1*Kh2}

##Preparing of empty vectors and matrices
li.mat<-matrix(rep(0,n^2),nrow=n)
li..mat<-matrix(rep(0,n^2),nrow=n)
xtilde<-matrix(rep(0,n*dim(b)[1]),nrow=n)
Bsummand<-array(rep(0,n*betadim^2),c(betadim,betadim,n))
B_k<-rep(0,betadim*betadim*(modes-1))
dim(B_k)=c(betadim,betadim,(modes-1))
mnew<-rep(0,n)
biterations<-rep(0,(modes-1)*max.iter*betadim)
dim(biterations)=c(betadim,modes-1,max.iter)
LogL<-c(rep(0,(modes-1)))
B=matrix(rep(0,betadim^2),nrow=betadim)

##########################################
## Start des iterativen Schätzprozesses ##
##########################################

start<-Sys.time()
```

```
#####################################################################
# additional iteration to stabilize the estimation procedure      #
# outer iteration over modes, inner iteration for the same mode   #
#####################################################################

for (k in 1:(modes-1))
{
for (it in 1:2)
{
ll<-logl(b,x,y,m,k)
li.<-ll[,2]      # li. for mode k
li..<-ll[,3]     # li.. for mode k

# Programming of the profile Likelihood as ixj matrix
for (j in 1:n) # li.mat for mode k
{li.mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,2]}
for (j in 1:n) # li..mat for mode k
{li..mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,3]}

for (p in 1:n)
{za<-t(apply((li..mat[,p]*Kh[,p])*x,2,sum))
ne<-sum(li..mat[,p]*Kh[,p])
xtilde[p,]<-x[p,]-za/ne}

for (p in 1:n) { ##updating b (step 2)
Bsummand[,,p]<-as.vector(li..)[p]*(xtilde[p,]%*%t(xtilde[p,]))
B<-B+Bsummand[,,p]}
B_k[,,k]<-B

##using ginv() instead of solve()
bnew<-b[,k]-ginv(B)%*%apply(as.vector(li.)*xtilde,2,sum)
for (p in 1:n)                      ##updating m (step 3)
{mza<-sum(li.mat[,p]*Kh[,p])
mne<-sum(li..mat[,p]*Kh[,p])
mnew[p]<-m[p,k]-mza/mne}

biterations[,k,it]<-b[,k]<-bnew
```

```
m[,k]<-mnew
LogL[k]<-sum(ll[,1])
print(b)}}

########################################
# usual iteration process over all modes #
########################################
it<-0
while (it<max.iter)
{
it<-it+1
for (k in 1:(modes-1))
 {
ll<-logl(b,x,y,m,k)
li.<-ll[,2]      # li. for mode k
li..<-ll[,3]     # li.. for mode k

# Programming of the profile Likelihood as ixj matrix
for (j in 1:n) # li.mat for mode k
{li.mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,2]}
for (j in 1:n) # li..mat for mode k
{li..mat[,j]<-logl(b,x,y,matrix(c(rep(m[j,],n)),nrow=n,byrow=TRUE),k)[,3]}

for (p in 1:n)
{za<-t(apply((li..mat[,p]*Kh[,p])*x,2,sum))
ne<-sum(li..mat[,p]*Kh[,p])
xtilde[p,]<-x[p,]-za/ne}
B=matrix(rep(0,betadim^2),nrow=betadim)
for (p in 1:n) { ##updating b (step 2)
Bsummand[,,p]<-as.vector(li..)[p]*(xtilde[p,]%*%t(xtilde[p,]))
B<-B+Bsummand[,,p]}
B_k[,,k]<-B
##using ginv() instead of solve()
bnew<-b[,k]-ginv(B)%*%apply(as.vector(li.)*xtilde,2,sum)
for (p in 1:n)            ##updating m (step 3)
{mza<-sum(li.mat[,p]*Kh[,p])
mne<-sum(li..mat[,p]*Kh[,p])
```

```
mnew[p]<-m[p,k]-mza/mne}
biterations[,k,it]<-b[,k]<-bnew
m[,k]<-mnew
LogL[k]<-sum(ll[,1])}
print(it);print(b)

#Abort criterion
diff<-matrix(rep(1,betadim*(modes-1)),nrow=betadim)
if(it>1) {for(i in 1:betadim){ for(j in 1:(modes-1)){
diff[i,j]<-abs(biterations[i,j,it]-biterations[i,j,(it-1)]) }}}
print(sum(diff))
if(sum(diff)<0.01) it<-max.iter}
##Output values
se_b<-matrix(rep(0,betadim*(modes-1)),nrow=betadim)
for(z in 1:(modes-1)){
se_b[,z]<-diag(sqrt(solve(-B_k[,,z])))}##St.fehler für b_k
t_b<-b[1:(modes-1)]/se_b           ##t-values für b_k
LogL[k]<-sum(ll[,1])
end<-Sys.time()
Time<-end-start
return(list(b=b,m=m,Time=Time,se_b=se_b,t_b=t_b,
LogL=LogL,y=y,x=x,T=T,sd1=sd1,sd2=sd2,h=h))}

#Marginal effects
  marg<-array(rep(0,1780*dim(x)[2]*nlevels(as.factor(y))),
  dim=c(dim(x)[2],nlevels(as.factor(y)),1780))
  for(j in 1:dim(x)[2]){
  for(k in 1:nlevels(as.factor(y))){
eta<-x%*%b+m
etanum<-exp(eta[,k])
denom<-apply(exp(eta),1,sum)
marg[j,k,]<-etanum/denom*(b[j,k]-mean(b[j,]))
  mean_marg<-apply(marg,c(1,2),mean)
  sd_marg<-apply(marg,c(1,2),sd)}}
  round(mean_marg,3)
  round(sd_marg,3)
```

```
  library(VGAM)
    mult_mod1<-vglm(formula = Mode~Age_D+Sex_D+Uni_D+
    I(Income_HV/10000)+Income_Edu+I(Price/100)+Duration,
    family = multinomial(refLevel=6))
 marg_eff<-margeff(mult_mod1)
 mean_marg_eff<-apply(margeff(mult_mod1),c(1,2),mean)
 round(mean_marg_eff,3)
 sd_marg_eff<-apply(margeff(mult_mod1),c(1,2),sd)
 round(sd_marg_eff,3)

#setwd("...")     ##choose corresponding folder
load("Bilbao_Data.Rdata")   #load data
attach(data)
mult_mod<-multgplm(y=data$Mode+1,x=cbind(data$Age_D,data$
Sex_D,data$Uni_D,data$Income_Edu,data$Price/100,data$Duration)
,T=cbind(data$Income_HV),max.iter=100)
attach(mult_mod)
#save.image("Bilbao_Est_Paper.Rdata") ##save results
```

R Code – Graphical presentation

This subsection gives the **R** code for producing the plots presented in the paper.

```
#graphical presentation of the nonparametric functions
#setwd(".../Daten/")     ##choose corresponding folder
load("Bilbao_Data.Rdata")   #Load data
n<-length(y)
Freq.1<-0;Freq.2<-5;Freq.3<-15
Freq.4<-7;Freq.5<-20;Freq.6<-25
Z<-matrix(rep(c(Freq.1,Freq.2,Freq.3,Freq.4,Freq.5,Freq.6),n),ncol=6,byrow=TRUE)
colnames(Z)<-c("Freq.1","Freq.2","Freq.3","Freq.4","Freq.5","Freq.6")
multdata<-as.data.frame(cbind(data,Z)) ;
multdata$Mode<-as.factor(multdata$Mode+1)
mult_data<-mlogit.data(multdata,shape="wide",choice="Mode",varying=13:18)
mult_data$Income_Edu<-ifelse(mult_data$Income_Edu==0,1,mult_data$Income_Edu)
mult_data$Income_HV<-mult_data$Income_HV/10000
mult_data$Price<-mult_data$Price/100
mult_mod_para<-mlogit(Mode~1|Age_D+Sex_D+Uni_D+Income_Edu
+Price+Duration+Income_HV ,data=mult_data,reflevel=6)
#mult_mod<-mlogit(Mode~Freq-1|Age_D+Sex_D+Uni_D+Income_Edu
+Price+Duration+Income_HV ,data=mult_data,reflevel=6)
#summary(mult_mod_para)
modes<-nlevels(as.factor(y))
betadim<-dim(x)[2]
coef(mult_mod_para)[1:((modes-1)*(betadim+2))]
b0<-coef(mult_mod_para)[(modes):((modes-1)*(betadim+1))]
b0<-matrix(b0,nrow=6,byrow=TRUE)
m0<-coef(mult_mod_para)[((modes-1)*(betadim+1)+1):((modes-1)*(betadim+1)+(modes-1))]
mo<-matrix(m0,nrow=1,byrow=TRUE)
a0<-coef(mult_mod_para)[1:(modes-1)]

#setwd(".../Daten/")     ##choose corresponding folder
load("Bilbao_Est_Paper.Rdata")
attach(mult_mod)
par(mfrow=c(3,2))
or1<-order(T)
plot(T[or1]/10000*sd1,mult_mod$m[or1,1],type="p",lwd=3,
```

```
main=("Own Car"),xlab="Income_HV/10000",ylab=expression(m[1](Inc)),
cex.main=2,cex.axis=1.8,cex.lab=1.7)
abline(a0[1],m0[1],lty=2,lwd=3,col=2)
rug(T[or1]/10000*sd1)
plot(T[or1]/10000*sd1,mult_mod$m[or1,2],type="p",lwd=3,
main=("Car Passenger"),xlab="Income_HV/10000",ylab=expression(m[2](Inc)),
cex.main=2,cex.axis=1.8,cex.lab=1.7)
abline(a0[2],m0[2],lty=2,lwd=3,col=2)
rug(T[or1]/10000*sd1)
plot(T[or1]/10000*sd1,mult_mod$m[or1,3],type="p",lwd=3,
main=("Bus"),xlab="Income_HV/10000",ylab=expression(m[3](Inc)),
cex.main=2,cex.axis=1.8,cex.lab=1.7)
abline(a0[3],m0[3],lty=2,lwd=3,col=2)
rug(T[or1]/10000*sd1)
plot(T[or1]/10000*sd1,mult_mod$m[or1,4],type="p",lwd=3,
main=("Train"),xlab="Income_HV/10000",ylab=expression(m[4](Inc)),
cex.main=2,cex.axis=1.8,cex.lab=1.7)
abline(a0[4],m0[4],lty=2,lwd=3,col=2)
rug(T[or1]/10000*sd1)
plot(T[or1]/10000*sd1,mult_mod$m[or1,5],type="p",lwd=3,
main=("Underground"),xlab="Income_HV/10000",ylab=expression(m[5](Inc)),
cex.main=2,cex.axis=1.8,cex.lab=1.7)
abline(a0[5],m0[5],lty=2,lwd=3,col=2)
rug(T[or1]/10000*sd1)
plot(0,0,ylab="",xlab="",col="white",axes=FALSE,
main=("Marginal effects of Income"),cex.main=2)
legend("center",c("parametric","nonparametric"),
col=c(2,1),lty=c(2,3),inset=0.01,lwd=3,cex=1.7)
```

Curriculum Vitae

Personal information

Name	Nils-Bastian Heidenreich
Date of birth	15 April 1980
Place of birth	Wittmund, Deutschland
Nationality	German

Education and Employment

April 2008 – April 2011	Research associate Institute for Statistics and Econometrics University of Göttingen
April 2007 – March 2008	Research assistant Institute for Statistics and Econometrics University of Göttingen
April 2002 – March 2007	Diploma degree (Dipl.-Hdl.) in Business Administration and Education University of Göttingen Diploma thesis: Bewertung von Cashflows mit zeitlich variierenden erwarteteten Renditen mit Hilfe von Vektorautoregressiven Modellen Advisor: Prof. Walter Zucchini
Juli 1999	General qualification for university entrance (Abitur) Mariengymnasium Jever

Scientific work

Working Papers

- Heidenreich, N., Schindler, A. and Sperlich, S., (2010)
  Bandwidth Selection Methods for Kernel Density Estimation - A Review of Performance.
  Available at http://ssrn.com/abstract=1726428.
- Heidenreich, N.-B., Langrock, R., and Sperlich, S., (2011)
  Semiparametric voter profiling in a multi-party system – new insights via flexible modeling.
- Bilbao-Ubillos, J., Fernández-Sainz, A., Heidenreich, N.-B., and Sperlich, S., (2011)
  A Semiparametric Model of Urban Transport Demand.

Talks

- Bandwidth Selection for Kernel Density Estimation - What is a reasonable choice?
  YSM 2009 in Glasgow, April 2009
- Kerndichteschätzung - Bestimmung der optimalen Bandweite
  DFG-SNF Research Group FOR916 in Bern, September 2009
- Semiparametric voter profiling in a multi-party system
  University of the Basque Country in Bilbao, Januar 2011

Memberships

- DFG / SNF FOR916: "Statistical Regularisation and Qualitative Constraints: Inference, Algorithms, Asymptotics and Applications" (April 2008 - March 2011)

Additional Information

- Languages: German (native), English (fluent)
- Software: R, LateX, MS Office, HTML, Fortran

Göttingen, den 18.05. 2011 Nils-Bastian Heidenreich

Versicherung an Eides Statt (§14, Promotionsordung)

Ich, Nils-Bastian Heidenreich, versichere an Eides Statt, dass ich die eingereichte Dissertation "Applications of nonparametric methods in economic and political science" selbstständig verfasst habe. Anderer als der von mir angegebenen Hilfsmittel und Schriften habe ich mich nicht bedient. Alle wörtlich oder sinngemäß den Schriften anderer Autorinnen und/oder Autoren entnommenen Stellen habe ich kenntlich gemacht.

Göttingen, den 18.05. 2011 Nils-Bastian Heidenreich