Multiple nonparametric regression and model validation for mixed regressors

Dissertation zur Erlangung des Grades eines
Doktors der Wirtschaftswissenschaft

Eingereicht an der Fakultät für Wirtschaftswissenschaften der
Universität Regensburg

Vorgelegt von: Joachim Schnurbus

Berichterstatter:
Prof. Dr. Rolf Tschernig, Universität Regensburg
Prof. Dr. Harry Haupt, Universität Bielefeld

Tag der Disputation: 06.07.2011

Contents

List of Figures

List of Tables

1 Introduction

This dissertation covers four essays on model validation in the context of nonparametric mixed kernel regression. In the first section of this chapter the conceptual framework of the dissertation is detailed, while the second section summarizes the content of the four essays and discusses their contribution relative to the existing literature.

1.1 Motivation and framework

In all of the following a regression framework is considered, where the behavior of a scalar continuous response variable y depends on K explanatory variables $x_1, \ldots, x_K$, the covariates, which are allowed to be mixed, i.e. continuous as well as discrete. In general, the behavior of y (w.r.t. the covariates) can be analyzed by estimating the complete conditional density (or conditional distribution) of this variable, or by estimating a single or several aspects of this density, e.g. the conditional mean or conditional quantiles. This work covers regression analysis based on L_2-norm (least squares regression), L_1-norm (least absolute deviations regression), and weighted versions of the latter (quantile regression).

For least squares regression, I assume that the conditional expectation

$$E(y|x_1, \ldots, x_K) = f(x_1, \ldots, x_K) \tag{1.1}$$

holds. Hence, the underlying regression model is

$$y = f(x_1, \ldots, x_K) + u$$

with an additive relationship between the systematic part of y, covered by the function $f(\cdot)$, and the error term u. From equation (1.1) $E(u|x_1, \ldots, x_K) = 0$ follows. In

a multiple regression framework[1], the conditional expectation of y is estimated, which corresponds to an estimation of the regression function $f(\cdot)$.

An analogous representation for quantile regression is based on assuming

$$Q_\vartheta(y|x_1,\ldots,x_K) = f(x_1,\ldots,x_K)$$

where $Q_\vartheta(.)$ is the conditional ϑ-quantile, with $\vartheta \in]0,1[$, with the special case of least absolute deviations regression resulting for $\vartheta = 0.5$. Specifying the regression function $f(\cdot)$ is an essential step within both frameworks, as all statements that are based on the estimated specification (e.g. interpretations, hypothesis tests, ...) depend on the specification of the regression function. Moreover, the derivation of statistical properties frequently requires the assumption of correctly specified functional form, compare Davidson & MacKinnon (2004, page 87).

This dissertation covers specifications from the parametric, semiparametric, and nonparametric class of specifications. A parametric specification, where the regression function also depends on a finite-dimensional parameter vector $\boldsymbol{\beta}$ requires to assume a certain function class that restricts the functional relationship between covariates and the response variable, that has to be completely specified a priori with all interactions, quadratic/cubic terms (or other representations of possible nonlinearity), etc. The benefits of typical parametric specifications are an easy computability and interpretability[2], and well-known properties[3]. These benefits come at the cost of a rigid assumption on the functional form, which is thus likely to be misspecified, unless extensive subject matter information (which is often not existing) can be provided by economic theory[4].

Nonparametric specifications do not require such rigid assumptions on $f(x_1,\ldots,x_K)$, usually only regularity conditions like a certain degree of differentiability are required. This flexibility comes along with usually high costs in terms of computation, of interpretation, and of the selection of an appropriate nonparametric configuration, compare chapters 3

[1] For this context, y is also denoted as regressand, while the $x_1,\ldots,x_K$ are also denoted as regressors.

[2] The partial effects of parametric specifications can be represented by a single formula that holds for the whole range of observations (global modeling).

[3] Note that there are also very sophisticated possibly highly nonlinear parametric specifications.

[4] Analogous to econometrics, this has to be considered for biometry, social science research, etc.

and 4. Semiparametric specifications are a combination of parametric and nonparametric specifications, and can be appropriate, if there is subject matter information w.r.t. some parts of the regression function, e.g. considering the relationship of regressand and some of the regressors.

Model validation covers specifications of the aforementioned classes of specifications[5]. For the given context, model validation thus corresponds to a validation of the estimated function $\hat{f}(x_1, \ldots, x_K)$, where we explicitly allow for continuous as well as discrete covariates. Usually a validation of the function aims at obtaining the correct specification and estimation of the regression function. For real data sets this aim is unrealistic, even more in the context of mixed regressors, unless a strong subject matter relationship is given by economic theory. Hence, applied work should focus on finding an "adequate" functional form in a sense that it passes the model validation procedure and is also appropriate for the purposes of the research, instead of searching for the correct functional form underlying the unknown data generating process. I also follow this more realistic reasoning and aim at finding an adequate specification of the regression function. The following paragraphs contain some details on the non- and semiparametric regression[6] estimation and the model validation framework.

Nonparametric and semiparametric estimation

According to Györfi et al. (2002, page 18) there are "four paradigms of nonparametric regression, local averaging, local modeling, global modeling . . . , and penalized modeling". The first three paradigms appear in this dissertation, as nonparametric kernel regression belongs to local averaging or local modeling (depending on the type of local estimation), while the B-splines approach (compare Eilers & Marx, 1996) we use for the semiparametric specifications of the quantile regression function, is subsumed under global modeling.

Nonparametric kernel regression methods became popular with the development of the local constant estimator by Nadaraya (1964) and Watson (1964). A kernel function is a weight function, that depends on the covariate value, on the value of the position where

[5] A discussion of further benefits and limitations of these classes of specifications can be found in chapter 5.

[6] The following paragraphs consider the regression estimation. For specific details on quantile regression estimation we refer to the literature.

the function is estimated (local estimation), and on a smoothing parameter, the bandwidth. In this exposition a second order Gaussian kernel function is used for continuous covariates, where observations are weighted according to the standard normal density[7]. There are different kinds of local estimation, where the underlying functional relationship between the response variable and continuous covariates is approximated by a local average (local constant estimator) or a local linear or local polynomial relationship[8].

Considering the asymptotic properties, nonparametric kernel regression estimators are consistent and asymptotically normally distributed under standard regularity conditions, compare Pagan & Ullah (1999)[9], although the rate of convergence of nonparametric estimators slows down with every additional continuous covariate, which is known as the curse of dimensionality, compare Härdle et al. (2004, pages 133f.). An impression of the curse of dimensionality can be obtained by Table 4.2 in Silverman (1986, page 94). Even though nonparametric regression estimators can be asymptotically unbiased (compare Pagan & Ullah, 1999, page 111), in finite samples they are usually biased[10]. Hence, the precision of a nonparametric estimator cannot be judged solely by the variance. Instead, mainly the mean squared error (MSE) of a nonparametric estimator is used, which is equal to the sum of variance and squared bias. As the MSE is only a measure for pointwise precision, usually the integrated (over all covariate values) MSE, the IMSE, is used, for details compare e.g. Härdle (1990, chapter 4).

In my work I follow the recent nonparametric mixed kernel regression approach proposed by Li and Racine (compare Li & Racine, 2004a, 2007; Racine & Li, 2004). Until this approach was developed, there were two ways to include discrete covariates in a multiple nonparametric regression. The first way was to treat discrete covariates as if they were continuous and weight them accordingly by kernel functions for continuous covariates (compare e.g. Anglin & Gencay, 1996). Of course, this made the curse of dimensionality

[7] For an overview on kernel functions and their relative efficiency (in the context of density estimation), compare Silverman (1986, page 43).

[8] Compare Fan & Gijbels (1996) for an extensive overview on local polynomial estimation.

[9] For mixed covariates compare Li & Racine (2007, chapter 4).

[10] Details on statistical properties as well as on bias-reduction techniques, like higher order kernels, can be found in Pagan & Ullah (1999).

even worse. A second way was to split the sample into subsamples corresponding to the different category combination of the discrete covariates. For each subsample, a nonparametric regression on the continuous covariates was conducted. This approach is called frequency approach (Li & Racine, 2007, chapter 3) and is only feasible, if each subsample contains enough observations for the corresponding nonparametric estimation.

In the approach of Li & Racine (2004a), the discrete covariates are also smoothed (as the continuous covariates) with certain kernel functions, that also depend on a smoothing parameter. There are four kernels for discrete covariates, the kernel of Wang & van Ryzin (1981) for ordered covariates (i.e. the covariate has a natural ordering), the kernel of Aitchison & Aitken (1976) for unordered covariates and corresponding kernel functions of Li & Racine (2004a) for both scale levels[11]. Smoothing discrete covariates leads to an additional bias, but the variance is reduced such that the (I)MSE may be improved. An appealing feature of the approach of Li & Racine (2004a) is that the number of discrete covariates does not matter asymptotically. Hence, there is no curse of dimensionality w.r.t. discrete covariates. The case of a nonparametric kernel regression with solely discrete covariates is considered by Ouyang et al. (2009).

Every kernel (for discrete as well as continuous covariates) depends on a bandwidth, which has to be estimated or selected prior to the final nonparametric regression. Different approaches on bandwidth estimation or selection can be found in nearly every textbook covering nonparametric regression[12] (e.g. Härdle, 1990; Fan & Gijbels, 1996; Pagan & Ullah, 1999; Györfi et al., 2002; Li & Racine, 2007), but also in general textbooks like Davidson & MacKinnon (2004). Compare also the work of Yang & Tschernig (1999) or the essay on classical bandwidth selection methods of Loader (1999). For mixed kernels, two data-driven methods are proposed for the regression context, least-squares cross-validation (compare Li & Racine, 2007, chapter 4) and the approach of Hurvich et al. (1998). In terms of (I)MSE, a lower bandwidth usually yields a smaller bias, but increases the variance. For the context of nonparametric quantile regression with mixed covariates, compare Li & Racine (2008).

Semiparametric regression function specifications consist of a parametric part and a

[11] For an overview on scale levels, compare Stevens (1946).

[12] However, the earlier ones only cover the univariate case.

nonparametric part and are intended to alleviate problems related to the curse of dimensionality. The cost are additional assumptions on the functional form and thus potential misspecification. In the dissertation, only partially linear specifications are considered, where the parametric and the nonparametric part are additively connected. Such specifications seem a natural choice, if there is a strong subject matter information for the relationship between the response variable and some of the covariates, while the impact of other covariates is not clear. The nonparametric part of the semiparametric regression function specification is modeled using kernels, following the estimation approach of Robinson (1988)[13], while we use B-splines (compare Eilers & Marx, 1996) for the quantile regression function specification. For a comprehensive treatment of partially linear models compare Härdle et al. (2000). Semiparametric regression methods in general are depicted in Ruppert et al. (2003), while Yatchew (2003) provides an interesting differencing-approach. For applications of semiparametric methods compare e.g. Horowitz & Lee (2002) or the recent work of Koenker (2010) on semiparametric quantile regression. Methods for semiparametric regression in the presence of mixed covariates can be found in part II of Li & Racine (2007) or e.g. in Li & Racine (2010).

In the following analyses only independent data are considered, although the mixed kernel methods can be applied to weakly-dependent data as in Li et al. (2009) or to more general data constellations, compare Li & Racine (2007, part V). Besides the already mentioned nonparametric and semiparametric literature, also the paper of Bierens (1983) and the overview essays of Bierens (1987), Delgado & Robinson (1992), Härdle & Linton (1994), Cai & Li (2009), and the monograph of Wand & Jones (1995) comprehensively illustrate kernel estimation and the underlying theory. For an overview on the work of Li and Racine and coauthors on non- and semiparametric mixed covariate estimation compare Li & Racine (2007) or Racine (2008). Besides their papers on mixed kernel regression estimation, also the following papers consider mixed covariates: Unconditional density estimation for mixed variables is shown in Li & Racine (2003) and extended to also cover irrelevant variables in Ouyang et al. (2006). Hall et al. (2004) consider conditional density estimation. Their framework is extended in Racine et al. (2004) to also allow for

[13] Li (1996b) has shown that the number of covariates in the nonparametric part may not exceed five, to obtain a $\sqrt{n}$-consistent estimator of the parameters in the parametric part.

multiple response variables. Li et al. (2009) provide a framework for the estimation of average treatment effects. Henderson (2010), Li (1996a), Li et al. (2009), and Maasoumi & Racine (2009) propose tests for a mixed kernel or similar framework, while Kiefer & Racine (2009) link kernels and Bayes models. Other approaches of including discrete covariates in a nonparametric estimation are e.g. provided by Ahmad & Cerrito (1994) and Delgado & Mora (1995). Applications of the mixed kernel methods for various areas of economics (besides those that are considered in detail in chapters 2 to 5) are e.g. provided by Maasoumi et al. (2007) and Haupt & Petring (2011) for growth regressions, while Henderson & Millimet (2008) analyze the gravity equation and Gyimah-Brempong & Racine (2010) investigate the relationship between aid and capital investment for developing countries. Chakrabarty et al. (2006) are analyzing household consumption expenditures, while Wilson & Carey (2004) consider returns to scale of hospitals in the United States. Further studies consider the effect of alcohol availability on crime (Gyimah-Brempong & Racine, 2006) or the effect of crop insurances for agricultural risk protection (Racine & Ker, 2006).

Model validation

Model validation for the current setup aims at finding an adequate specification for the (quantile) regression function from the parametric, semiparametric, or nonparametric class of specifications. The validation of the estimated $f(x_1, \ldots, x_K)$ can be broken down into two parts, first, the adequate covariates $x_1, \ldots, x_K$ should be included, second, the adequate functional relationship $f(\cdot)$ for these covariates should be applied. Considering the first part, adequate covariates means that we on the one hand omit no important covariates, which is not under consideration in this dissertation as for the applications we use well-known data sets proposed in the literature, although omitted covariates can be a serious problem in applied work, yielding biased and inconsistent estimates. On the other hand, irrelevant covariates should usually not be included for efficiency reasons. In nonparametric kernel regression and quantile regression, the relevance of discrete covariates is already indicated by the estimated bandwidths approaching their upper bounds[14], compare Li & Racine (2004b), unless the kernel of Wang & van Ryzin

[14] Hall et al. (2007) consider the estimation of nonparametric regression functions in the presence of irrelevant covariates.

(1981) is used (compare chapter 3). In local constant kernel regression, also the relevance of continuous covariates is indicated by the estimated bandwidths (if the bandwidths are chosen "correctly"). Additionally, for a mixed covariate context, there are two tests of significance, the test of Racine (1997) for continuous covariates and the test of Racine et al. (2006) for discrete covariates. To my best knowledge there are no tests so far for the significance of covariates in a mixed kernel quantile regression framework.

For the second part, the validation of the functional relationship, in the literature usually a test for correct specification is conducted. In accordance to the findings of Davidson & MacKinnon (2004, page 690) that "Nonparametric methods can be useful even when we are primarily interested in estimating a parametric model", these tests for correct parametric or semiparametric specification usually rely on nonparametric estimation techniques. Consider for example the tests of Delgado & Stengos (1994), Fan & Li (1996), Horowitz & Spokoiny (2001), Lee (2000), Li & Wang (1998), Stute (1997), Whang & Andrews (1993), and Zheng (1996) for regression functions and the essay on their performance by Miles & Mora (2003), and for a quantile regression framework e.g. the test of Zheng (1998). The test of Hsiao et al. (2007) for correct parametric specification of a regression function is the only test that explicitly allows for a mixed covariate structure of the underlying nonparametric estimation and is thus also applied in this dissertation.

If it is desirable for validating the functional form to directly judge and compare the performance of the different specifications for a given data set, these specification tests are not sufficient. The fit[15] of specifications is also no sufficient basis for such a model comparison (as always), since usually the more flexibility a specification provides, the better the fit that can be obtained. Additionally, the fit of parametric specifications can be improved by including additional covariates, whereas for nonparametric specifications, a very high fit can be obtained by choosing the bandwidths sufficiently small. Hence, to protect against such overfit, frequently a comparison of the prediction performance is applied. We follow this approach and conduct a Monte Carlo simulation with repeated subsampling, i.e. within each replication we randomly assign the sample observations into

[15] In this dissertation, the fit of estimated regression function specifications is measured by the squared correlation between observed and fitted response values.

an estimation and a validation subsample. Next, we estimate all considered specifications using the observations of the estimation subsample and predict for the observations of the validation subsample. For the judgment of the prediction performance for the regression case, we compute the measure ASEP, the average squared error of prediction, which is an estimator of the integrated mean squared error of prediction (IMSEP). We estimated solely the MSEP, if we had conditioned on a certain covariate combination, but as we randomly determine the observations of estimation and prediction subsample with various covariate values (under the assumption that the given sample observations cover the range of covariate values), we estimate the IMSEP. Without resampling it is likely that a bad estimate for the IMSEP of the different specifications is obtained, as this corresponds to an estimation of IMSEP using only one observed ASEP, thus this would not provide a solid basis for the validation (although in the literature, often a prediction performance comparison is conducted without repeated subsampling as e.g. by Anglin & Gencay, 1996; Gencay & Yang, 1996; Bin, 2004). From the prediction performance simulation we obtain an empirical distribution of ASEPs for each considered specification. For the quantile regression framework, the average theta weighted error (ATWE) is applied, as this corresponds to the appropriate loss function. Further details on the validation can be found in the subsequent section and in chapters 2, 3, and 5.

1.2 Overview and contribution

The four essays in chapters 2, 3, 4, and 5 have several common features, that are partially indicated by the dissertation title. All essays consider the nonparametric kernel regression and/or kernel quantile regression for a mixed covariate setting, following the approaches of Li and Racine and several coauthors. Each essay also covers the topic model validation, mainly the validation of regression function specifications of different classes of specifications (parametric versus semiparametric versus nonparametric), but also model validation in the context of finding an appropriate nonparametric configuration, compare chapter 4.

There are two main contributions of this dissertation. First, the essays provide guidelines and tools (compare chapters 3 and 4) to assist choosing an appropriate nonparamet-

ric configuration. The nonparametric configuration covers the type of local estimation, the kernel functions for discrete and continuous covariates, and the method of bandwidth estimation. This is important, as the estimation results depend on the nonparametric configuration. Second, as specifications of the three different classes of specifications are compared, the dissertation provides a framework for showing, which specification is preferable in a given data constellation, and also what are the costs of using a simpler, e.g. linear parametric specification. Additionally, every essay contains an application on a real data set, where the proposed methods and tools are demonstrated. Moreover, all essays provide the necessary computational details (seeds, any setup that might deviate from the default, etc.) to allow for a complete reproduction of the results[16] of the corresponding study in R (R Development Core Team, 2010).

Statistical validation of functional form in multiple regression using R

The essay in chapter 2 considers choice and validation of the functional form of a parametric specification of the regression function. Several validation techniques are introduced that mainly rely on nonparametric mixed kernel regression, thus the nonparametric estimation and interpretation is also discussed. First, we test for correct parametric specification using the test of Hsiao et al. (2007). Second, a Monte Carlo simulation with repeatedly drawn subsamples (drawn without replacement) for a comparison of the prediction performance between parametric and nonparametric specifications, measured by the average squared error of prediction (ASEP), where lower values are preferable. This measure is computed for every specification for each of the 10,000 replications. The empirical ASEP distribution functions of both specifications are compared graphically, as they allow for a stochastic dominance interpretation following Vinod (2008, Section 4.3.6). This is already mentioned in Haupt et al. (2010a). Additionally, a paired t-test for a significant difference between the means of these ASEP-distributions is conducted as well as the number of replications where one specification outperforms the other in terms of ASEP is computed.

Besides the nonparametric validation techniques, graphical tools are introduced to support the model validation process. These tools are based on the R-package `relax` of

[16] For a guideline on reproducible research, compare Koenker & Zeileis (2009).

Wolf (2009). The `slider` function of this package[17] can be used for designing interactive graphs, where the on-screen output can be instantly adjusted by sliders or buttons. We show by an `R`-code example how these graphical tools can be used for data exploration, and give examples for their support of the validation of functional form, e.g. by analyzing the ASEP-distributions.

The proposed techniques and tools are applied on a parametric specification for the wage data set of Hamermesh & Biddle (1994), where the impact of look on earnings for male and female employees is analyzed. A parsimonious (low model complexity) parametric specification is selected from a class of nested specifications using the Schwarz criterion and successfully validated by the proposed techniques. Hence, the parsimonious parametric specification seems to possess an adequate functional form. As for the nested parametric specifications, the information criteria of Schwarz and Akaike do not coincide, we also (successfully) validate the specification selected by the latter criterion.

A final contribution of this essay is an outlook on a quantile regression framework, since the mean regression possibly does not capture the central tendency of the conditional wage for the given data set. An extensive treatise of specification search and validation for a nonlinear quantile regression framework can be found in the fourth essay (chapter 5).

On Nonparametric Estimation of a Hedonic Price Function

The second essay (chapter 3) considers the validation of functional form, again for a parametric specification. In contrast to the previous essay, this essay is a replication study and the parametric specification under consideration was found to be inappropriate in the paper of Anglin & Gencay (1996). Anglin & Gencay (1996) found a semiparametric, partially linear specification, estimated by the approach of Robinson (1988), superior to a parametric log-linear specification for a Canadian housing data set, where the sales price of the houses is analyzed w.r.t. different characteristics (e.g. the number of bedrooms). Parmeter et al. (2007) replicated the analysis of Anglin & Gencay (1996) and suggested a nonparametric mixed kernel specification, as they rejected the hypothesis of correct partially linear specification using the test of Delgado & Gonzalez Manteiga (2001). In

[17] Equivalent functions can also be found in the `aplpack`-package of Wolf (2010).

both papers, the authors argue for the superiority of the semiparametric or nonparametric specification partially by showing their superior in-sample fit, which is no valid criterion for comparing specifications of different specification classes, as the fit is usually always higher if more flexibility is allowed for in a functional specification. We find that both papers do not accurately validate the specifications, as Anglin & Gencay (1996) conduct a prediction performance comparison, but only once non-randomly split the sample, thus either specification could be superior depending on which observations are chosen for estimation or prediction subsample. Parmeter et al. (2007) in turn do not consider the parametric specification, but their test of the partially linear specifications does not allow for mixed covariates (and thus different kernels and bandwidths) for the variables in the nonparametric part of the semiparametric specification.

We reconsider the major specifications of both previous papers and again use our nonparametric model validation techniques introduced in the first essay. The test of Hsiao et al. (2007) is conducted on the parametric specification, but is extended in comparison to the first essay. Now we allow for different nonparametric configurations that vary in the discrete kernel type, the bandwidth estimation method, the type of local estimation, and whether the continuous covariate enters logarithmized or not. For each of the nonparametric configurations, we get two p-values, as we apply IID- as well as wild-bootstrap for determining the null distribution of the test statistic. We find that only the configurations using local linear estimation in combination with the discrete kernels of Aitchison & Aitken (1976) and Wang & van Ryzin (1981) and with bandwidths selected by the method of Hurvich et al. (1998) lead to a rejection of the hypothesis of correct parametric specification on a 5% significance level, all other test configurations (covering 28 out of 32 p-values) do not yield a rejection. This indicates on the one hand that the test of Hsiao et al. (2007) is sensitive w.r.t. the nonparametric configuration, on the other hand that the simple loglinear parametric specification may not be largely misspecified. In the Monte Carlo prediction performance comparison, we include the log-linear specification, a mixed kernel version of the partially linear specification of Anglin & Gencay (1996), and the nonparametric mixed kernel specification of Parmeter et al. (2007), which also corresponds to the configuration with the smallest p-value in the test of Hsiao et al. (2007). We find that the parametric specification is superior to the semi-

and nonparametric specification, and that the semiparametric specification is superior to the nonparametric specification. We also find that the semiparametric specification predicts remarkably poorly for some of the 10,000 replications. Hence, we can successfully validate the parametric specification, indicating that for the given data set neither the specification proposed by Anglin & Gencay (1996) nor that proposed by Parmeter et al. (2007) is adequate. As additional contribution, we carefully analyze the discrete kernels and estimated bandwidths of the nonparametric configuration of Parmeter et al. (2007), where we find out that the discrete kernel of Wang & van Ryzin (1981) has only limited smoothing abilities for the underlying discrete covariate.

This paper is the first work in the context of mixed covariates that shows that also the nonparametric configuration search might need a lot of consideration, as for the test of Hsiao et al. (2007), there are a lot of possible nonparametric configurations that lead to different test results. Furthermore, also the kernel choice for weighting discrete covariates needs attention, which is an important contribution, as usually (based on the experience for kernels for continuous covariates) the kernel choice is not considered as that important for a nonparametric estimation, compare e.g. Silverman (1986, page 43).

Black box bandwidths — How hat matrix analysis illuminates nonparametric mixed kernel regression

The focus of this paper (chapter 4) is on the validation of different nonparametric configurations. For this purpose, the hat matrix $\mathbf{H}$ is analyzed, which is the connection between the vector of observed and of fitted response values ($\mathbf{y}$ and $\widehat{\mathbf{y}}$) as in $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In the nonparametric literature, solely the trace of the hat matrix was under consideration so far, compare e.g. Hurvich et al. (1998).

We investigate the structure of the hat matrices for local linear and local constant kernel regression, where (besides the type of local estimation) the hat matrix depends on the kernel functions, the bandwidths, and the covariate values. Next, we analyze the behavior of the hat matrix for different local estimation types and varying bandwidth values for continuous and discrete covariates, using simulated data. For the hat matrix analysis, several measures are developed, like the percentage of (potential) overfitting observations. All the measures that are relevant for comparing the hat matrix of different nonparametric configurations are summarized in a single plot of the COlumn sums against

the main-DIagonal elements of the hat matrix, denoted CODI-plot. This plot also allows for an identification of observations that are (potentially) very important/unimportant for the nonparametric estimation in general. We reconsider the 16 nonparametric configurations of the Hsiao et al. (2007)-test of the second essay (chapter 3) and analyze, which information is instantly available after nonparametric estimation by fit and estimated bandwidths, where a new measure for comparing discrete kernels is proposed, the percentage of smoothing. Then, we show which additional information is gained by a hat matrix analysis (which also has negligible computational costs) and how this information prevents from selecting a disadvantageous (again in terms of overfit) nonparametric configuration. The analysis of fit, estimated bandwidths, and hat matrices allows for a reduction from 16 to 3 considerable nonparametric configurations, where the corresponding p-values are always higher than 5% for the test of Hsiao et al. (2007) in the second essay.

Apart from the trace, the hat matrix analysis for nonparametric kernel regression is completely new contribution to the literature, and thus also all proposed measures and the CODI-plot.

Cross-validating fit and predictive accuracy of nonlinear quantile regressions

The final paper (chapter 5) focuses on the validation of functional form for a quantile regression framework, thus this work addresses and extends the idea of the outlook in the first essay. We again consider specifications from the three classes of specifications for the quantile regression function that allow for a different amount of nonlinearity. Again, the nonparametric specification is based on mixed kernels, while the nonparametric part in the partially linear specification is modeled with B-Splines, where we also allow for a monotonicity restriction. We use the nonparametric estimation approach of Li & Racine (2008), where the conditional quantiles are estimated from the conditional density of the response variable, while the bandwidths stem from conditional density estimation and are appropriately adjusted.

As the specifications originate from very different specification classes and there is no method of estimating the model complexity of the nonparametric quantile regression specification (like the trace of the hat matrix for nonparametric kernel regression), yet, our analysis is completely based on cross-validation methods which are similar to those

for judging the prediction performance in chapters 2 and 3. In our cross-validation simulations, we randomly (drawn without replacement) split the sample in an estimation and a prediction subsample and compare fit and prediction performance by the appropriate measures for a quantile regression context (compare Gneiting, 2010). A major goal of our analysis is to achieve a specification comparison that is as fair as possible. Hence, we extensively discuss the adjustments that are necessary (as we only use the observations in the estimation subsample for estimating the specifications, and not all sample observations) for comparing the different classes of specifications. Such adjustments consider the number and position of the knots for the B-spline approach or a rescaling of the estimated bandwidths for the mixed kernel approach. Additionally, we carefully discuss the different advantages and limitations underlying the three classes of specifications.

In this paper, we apply the methods on the Boston housing data set, where the median value of houses in different areas is analyzed w.r.t. several covariates, e.g. the percentage of lower status population. We cross-validate three specifications for the conditional first and second quartile and find out that for the conditional median, the nonparametric specification is superior, while the partially linear specification is preferable when the conditional first quartile is under consideration. Hence, we show that the degree of nonlinearity of the quantile regression function that is required for a given data set may vary with the conditional quantile under consideration. Additionally, our cross-validation approach allows for a comparison of the estimated partial effects of all specifications, where we find areas of increased variance for the more flexible specifications.

An extensive Monte Carlo simulation is developed, where the cross-validation approach is repeatedly conducted for randomly drawn observations of 12 DGPs. These DGPs differ in the distribution of the continuous covariate (normal or uniform distribution), the degree of nonlinearity (two different degrees that can also be interpreted as different signal-to-noise ratios), and the error distribution (a heteroscedastic error distribution, a distribution that is heteroscedastic and skewed, and a mix of both). This Monte Carlo simulation gives an impression, which specifications are preferable in terms of fit or prediction performance for what kind of DGP, where we explicitly allow for typical data constellations in a quantile regression framework.

This is the first paper in the context of quantile regression, where the framework for

a fair comparison between such different classes of specifications is derived and applied to real data (the famous Boston housing data) and studied for various nonlinear DGPs in an extensive Monte Carlo simulation.

2 Statistical validation of functional form in multiple regression using R

This essay is joint work with Harry Haupt[1] and Rolf Tschernig[2].

It is published in H. D. Vinod (Ed.), *Advances in Social Science Research Using R*, volume 196 of *Lecture Notes in Statistics* chapter 9, (pages 155-166), Springer, 2010, compare Haupt et al. (2010b).

Link:

http://www.springer.com/statistics/business%2C+economics+%26+finance/book/978-1-4419-1763-8

[1]Department of Business Administration and Economics, Bielefeld University, hhaupt@wiwi.uni-bielefeld.de

[2]Department of Economics, University of Regensburg, rolf.tschernig@wiwi.uni-regensburg.de

3 On nonparametric estimation of a hedonic price function

This essay is joint work with Harry Haupt[1] and Rolf Tschernig[2].

It is published in the *Journal of Applied Econometrics*, compare Haupt et al. (2010a).

Link:

http://onlinelibrary.wiley.com/doi/10.1002/jae.1186/abstract

[1]Department of Business Administration and Economics, Bielefeld University, hhaupt@wiwi.uni-bielefeld.de

[2]Department of Economics, University of Regensburg, rolf.tschernig@wiwi.uni-regensburg.de

4 Black box bandwidths — How hat matrix analysis illuminates nonparametric mixed kernel regression

This essay is under review in *Computational Statistics & Data Analysis*.

Abstract: Nonparametric mixed kernel regression provides a functionally flexible approach for a setting of mixed discrete and continuous covariates (compare Li & Racine, 2007). As every smoothing method it is based on a priori decisions on the nonparametric configuration, among others choosing the approach for bandwidth estimation. Though the estimated bandwidths have an essential influence on the outcomes of nonparametric regression, this influence remains largely unexplored and unexploited – especially in a multiple regression setting.

The aim of this paper is to show how the hat matrix (smoothing matrix) of nonparametric mixed kernel regression can help to select an appropriate nonparametric configuration. Various measures for comparing the hat matrix of different nonparametric configurations can be summarized in CODI-plots. Hat matrix analysis does not depend on the covariate dimension and enables a detailed analysis of the impact of single observations on nonparametric regression estimation. The scope of the nonparametric mixed kernel regression hat matrix analysis is illustrated using artificial data with simulated covariates and also for different nonparametric configurations using real data.

Keywords: CODI-plots, mixed covariates, model selection, overfitting observations, percentage of smoothing.

4.1 Introduction

In a regression framework, the hat matrix $\mathbf{H}$ is the connection between the vectors of observed $\mathbf{y}$ and estimated $\hat{\mathbf{y}}$ values of the response variable,

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}. \tag{4.1}$$

It is of dimension $n \times n$ and the sum of all entries of $\mathbf{H}$ is equal to n. The hat matrix is commonly used for parametric linear specifications, e.g. to compute the leverage effect of observations on the estimated ordinary least squares (OLS) regression hyperplane (e.g. Ruppert et al., 2003), for a general outlier analysis (e.g. Rousseeuw & van Zomeren, 1990), and/or for constructing robust estimators (e.g. Chave & Thomson, 2003). For OLS regression the hat matrix is well-known to be the symmetric and idempotent matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Thus it constitutes a projection matrix which only depends on the regressor matrix $\mathbf{X}$ and the sum of diagonal entries and hence the trace of the hat matrix $\mathbf{H}$ is equal to the rank of $\mathbf{X}$, the model complexity.

Simplifying assumptions such as a linear in parameters regression function play a crucial role as regression interpretations and tests are usually based on the assumption of correctly specified functional form. Unfortunately the choice of a specific functional form commonly cannot be backed up by subject matter information. Hence a promising approach lies in imposing as little structure as possible, for example, by using nonparametric specifications. Q. Li and J. S. Racine extended the nonparametric kernel approach (Li & Racine, 2004a, 2007; Racine & Li, 2004) to allow for simultaneous smoothing of mixed covariates. For the application of nonparametric mixed kernel estimation, however, various decisions on the nonparametric configuration have to be made, such as choosing the type of local estimation and the bandwidth selection method. In a multiple regression setting with mixed covariates, data-driven procedures for bandwidth estimation are required.

Up to now, only the trace of the hat matrix appears in the literature considering nonparametric regressions (see Hurvich et al., 1998), serving as an approximate measure of model complexity, though in contrast to the parametric case the hat matrix lacks the crucial property of idempotency. For the general case of nonparametric kernel regression

with mixed covariates this paper contributes to the literature by showing how further aspects of the hat matrix beyond its trace can be used to evaluate different nonparametric configurations and choose among them in terms of avoiding overfit[1]. The effects of different choices of nonparametric configurations – varying bandwidths for continuous and discrete covariates as well as different local estimation types – on the hat matrix are analyzed using an example with simulated artificial covariates. Different tools for hat matrix analysis are proposed, especially the CODI-plot which summarizes the relevant hat matrix information required for comparing nonparametric configurations. Further, a measure for comparing different discrete kernels is introduced, the percentage of smoothing. Finally, using an example of Canadian housing data I illustrate the gain in information provided by nonparametric hat matrix analysis compared to solely interpreting fit and estimated bandwidths.

Section 4.2 contains the theory of nonparametric kernel regression and hat matrices. In section 4.3, the hat matrix behavior (with respect to varying local estimation type and bandwidths) and the tools for hat matrix analysis (in particular the CODI-plot) are illustrated on an example data set. Section 4.4 covers the comparison of different nonparametric configurations for a Canadian housing data set. Finally, section 4.5 summarizes the results.

4.2 Theory on nonparametric mixed kernel regression and hat matrices

This section covers the theoretical background on nonparametric kernel regression and on hat matrices. I review local linear and local constant kernel estimators[2] in subsection

[1] Overfitting can even occur for a data-driven bandwidth selection, as numerical optimization may fail. Alternative simulation-based comparisons of the prediction performance of different nonparametric configurations (e.g., Haupt et al., 2010a,b; Henderson & Millimet, 2008) rely on resampling (and the corresponding assumptions about the data generating process) and have quite considerable computational costs depending on the covariate dimension.

[2] This paper intends to facilitate applied nonparametric analysis for a setting of mixed covariates. Hence, I restrict the attention to local estimation methods that are already implemented, for example in the

4.2.1. Subsection 4.2.2 depicts the weighting of observations by kernel functions, while subsection 4.2.3 shows the structure of hat matrices for nonparametric kernel regression.

4.2.1 Local kernel regression

The local linear estimator for the conditional mean at position $\mathbf{x}_0$, which is the $(K \times 1)$-vector of (discrete and/or continuous) covariate values, is

$$\hat{\alpha}(\mathbf{x}_0) = (1, 0, \ldots, 0) \left(\mathbf{X}'_{0,(C)} \mathbf{W}(\mathbf{x}_0) \mathbf{X}_{0,(C)}\right)^{-1} \mathbf{X}'_{0,(C)} \mathbf{W}(\mathbf{x}_0) \mathbf{y} \tag{4.2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X}_{0,(C)} = \begin{pmatrix} 1 & x_{1,1} - x_{0,1} & \ldots & x_{1,C} - x_{0,C} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i,1} - x_{0,1} & \ldots & x_{i,C} - x_{0,C} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} - x_{0,1} & \ldots & x_{n,C} - x_{0,C} \end{pmatrix},$$

$$\mathbf{W}(\mathbf{x}_0) = \begin{pmatrix} W(\mathbf{x}_0, \mathbf{x}_1, \mathbf{h}) & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & W(\mathbf{x}_0, \mathbf{x}_i, \mathbf{h}) & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & W(\mathbf{x}_0, \mathbf{x}_n, \mathbf{h}) \end{pmatrix}.$$

The vector of covariate values of the ith observation is $\mathbf{x}_i$, while $W(.)$ is the weighting function. The latter depends on the bandwidth vector $\mathbf{h}$ as detailed in subsection 4.2.2. Besides the estimated mean regression effect, the local linear estimator yields also estimated first partial derivatives for all C continuous covariates ($C \leq K$).

An alternative local kernel estimator, the local constant estimator, is obtained if the matrix $\mathbf{X}_{0,(C)}$ is reduced to the first column, a column vector of ones. Thus we get

$$\hat{\alpha}(\mathbf{x}_0) = \sum_{i=1}^{n} \frac{W(\mathbf{x}_0, \mathbf{x}_i, \mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_0, \mathbf{x}_l, \mathbf{h})} y_i, \tag{4.3}$$

where i is replaced by l in the denominator.

---

np-package of Hayfield & Racine (2011) for R (R Development Core Team, 2010).

4.2.2 Weighting of observations

Next, choice of weighting function $W(\mathbf{x}_0, \mathbf{x}_i, \mathbf{h})$, and the determination of the smoothing parameters contained in $\mathbf{h}$ have to be considered. Considering the first point, I choose the following structure of $W(.)$ for a multiple regression framework with K covariates:

$$W(\mathbf{x}_0, \mathbf{x}_i, \mathbf{h}) = \prod_{k=1}^{K} w_k(x_{0,k}, x_{i,k}, h_k),$$

where $w_k(.)$ are the weighting (kernel) functions of the (continuous and discrete) covariates. As $W(.)$ is the product of all kernel functions, it is called generalized product kernel. The kernel functions differ with respect to the scale level of the underlying covariate.

In this paper, for continuous covariates a second order Gaussian kernel[3] is used as weighting function, thus the observations are weighted according to the standard normal density. The bandwidth h_k with $h_k \in]0, \infty[$ is also called smoothing parameter, as higher values of h_k yield a smoother (i.e. less curvature) estimated function. An extremely large h_k causes a linear relationship between covariate and response variable in a local linear regression, while it causes irrelevance of the underlying covariate in a local constant regression.

In contrast, there are only a few different kernel functions for discrete covariates and their choice needs additional care compared to the case of continuous covariates, see for example Haupt et al. (2010a). For discrete covariates, the weighting functions also differ for ordered (the variable has a natural ordering) and unordered covariates. The kernel function of Li & Racine (2004a) for unordered discrete covariates is

$$w_k(x_{0,k}, x_{i,k}, h_k) = \begin{cases} 1 & \text{for} \quad x_{i,k} = x_{0,k}, \\ h_k & \text{for} \quad x_{i,k} \neq x_{0,k}, \end{cases}$$

where for estimation at position $x_{0,k}$ all observations that have the same category of the kth covariate as $x_{0,k}$ are weighted by 1. Observations with other category values for $x_{i,k}$ are weighted by h_k with $h_k \in [0, 1]$. A value of $h_k = 1$ implies complete smoothing, that

[3] There are a lot of different kernel functions for continuous covariates, but the continuous kernel choice does not heavily affect the relative efficiency of kernel estimators, compare Table 3.1 in Silverman (1986, page 43) for density estimation, or the statement of Fan & Gijbels (1996, page 76) for kernel regression.

is x_k is irrelevant, since there is no distinction between the categories of the underlying discrete covariate. In contrast, a value of $h_k = 0$ means that there is no smoothing of the respective discrete covariate at all, as only the observations of the same category as $x_{0,k}$ are used for estimation. This corresponds to the so-called frequency approach (compare Li & Racine, 2007, chapter 3). The kernel for ordered discrete covariates is

$$w(x_{0,k}, x_{i,k}, h_k) = h_k^{|x_{i,k} - x_{0,k}|},$$

where the weighting depends on the distance of the category of the observations to the category of the position of interest. The values $h_k = 0$ and $h_k = 1$ have an equivalent interpretation to the case of unordered discrete covariates.

Alternative kernel functions used in the configurations of section 4.4 for discrete covariates are the kernel function of Aitchison & Aitken (1976) and the kernel of Wang & van Ryzin (1981). The kernel of Aitchison and Aitken for unordered covariates is

$$w_k(x_{0,k}, x_{i,k}, h_k) = \begin{cases} 1 - h_k & \text{for} \quad x_{i,k} = x_{0,k}, \\ \frac{h_k}{q-1} & \text{for} \quad x_{i,k} \neq x_{0,k}, \end{cases}$$

where q is the number of categories of the discrete covariate. Contrary to the kernels of Li and Racine, the bandwidth for the Aitchison and Aitken kernel can take values in $[0, (q-1)/q]$ instead of $[0, 1]$. The kernel of Wang and van Ryzin for ordered covariates weights observations according to

$$w_k(x_{0,k}, x_{i,k}, h_k) = \begin{cases} 1 - h_k & \text{for} \quad x_{i,k} = x_{0,k}, \\ \frac{1}{2}(1 - h_k) h_k^{|x_{i,k} - x_{0,k}|} & \text{for} \quad x_{i,k} \neq x_{0,k}. \end{cases}$$

As is shown in Haupt et al. (2010a), the Wang and van Ryzin kernel has only limited smoothing abilities and may lead to inefficient estimation.

The second point to consider before turning to the regression function estimation is the determination of the smoothing parameters. In a multiple mixed covariate setting there are no rules of thumb or the like for determining $\mathbf{h} = (h_1, \ldots, h_K)'$, but there exist two data-driven approaches that are able to deal with mixed covariates, least-squares cross-validation (compare Li & Racine, 2007, chapter 4) and the approach of Hurvich et al. (1998) based on a modified Akaike information criterion.

4.2.3 Hat matrices for nonparametric kernel regression

In this subsection, the structure of the hat matrices for local constant and local linear estimation is presented. Equation (4.1) reveals that a hat matrix states the relationship between observed and fitted response. A typical element of the hat matrix will be denoted as H_{ji}, for row j and column i, where H_{ji} represents the weight of the ith observation for computing the fitted value of the jth observation. Diagonal entry H_{ii} represents the weight of an observation i for the computation of the respective fitted value. The entries of row j show, how much weight the response value of each of the n observations has for the computation of the fitted value $\hat{y}_j$, equivalently, the entries of column i show the weight of y_i for the computation of all n fitted response values.

For the hat matrices of the nonparametric local constant and local linear approach, based on equations (4.3) and (4.2), we have to consider that the regression hyperplane is estimated at every observational point in the K-dimensional covariate space according to

$$\hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_j \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} \hat{\alpha}(\mathbf{x}_1) \\ \vdots \\ \hat{\alpha}(\mathbf{x}_j) \\ \vdots \\ \hat{\alpha}(\mathbf{x}_n) \end{pmatrix}.$$

The hat matrix for a local linear estimation follows from equation (4.2) where all entries in $\mathbf{x}_0$ are replaced by the entries of $\mathbf{x}_j$ for the representation of row j of the hat matrix,

$$\mathbf{H}^{LL} = \begin{pmatrix} (1,0,\ldots,0)\left(\mathbf{X}'_{1,(C)}\mathbf{W}(\mathbf{x}_1)\mathbf{X}_{1,(C)}\right)^{-1}\mathbf{X}'_{1,(C)}\mathbf{W}(\mathbf{x}_1) \\ \vdots \\ (1,0,\ldots,0)\left(\mathbf{X}'_{j,(C)}\mathbf{W}(\mathbf{x}_j)\mathbf{X}_{j,(C)}\right)^{-1}\mathbf{X}'_{j,(C)}\mathbf{W}(\mathbf{x}_j) \\ \vdots \\ (1,0,\ldots,0)\left(\mathbf{X}'_{n,(C)}\mathbf{W}(\mathbf{x}_n)\mathbf{X}_{n,(C)}\right)^{-1}\mathbf{X}'_{n,(C)}\mathbf{W}(\mathbf{x}_n) \end{pmatrix}.$$

In applied data analysis, it may occur that $\mathbf{X}'_{j,(c)}\mathbf{W}(\mathbf{x}_j)\mathbf{X}_{j,(c)}$ is singular for a row j. If this happens, then I suggest to replace the corresponding row by the unit vector with jth element equal to 1, as the closer the term gets to singularity, the more equal the structure of row j gets to this unit vector.

The hat matrix for a local constant estimation, obtained from equation 4.3, is

$$\mathbf{H}^{LC} = \begin{pmatrix} \frac{W(\mathbf{x}_1,\mathbf{x}_1,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_1,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_1,\mathbf{x}_i,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_1,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_1,\mathbf{x}_n,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_1,\mathbf{x}_l,\mathbf{h})} \\ \vdots & \ddots & & & \vdots \\ \frac{W(\mathbf{x}_j,\mathbf{x}_1,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_j,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_j,\mathbf{x}_i,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_j,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_j,\mathbf{x}_n,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_j,\mathbf{x}_l,\mathbf{h})} \\ \vdots & & & \ddots & \vdots \\ \frac{W(\mathbf{x}_n,\mathbf{x}_1,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_n,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_n,\mathbf{x}_i,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_n,\mathbf{x}_l,\mathbf{h})} & \cdots & \frac{W(\mathbf{x}_n,\mathbf{x}_n,\mathbf{h})}{\sum_{l=1}^{n} W(\mathbf{x}_n,\mathbf{x}_l,\mathbf{h})} \end{pmatrix}.$$

For $\mathbf{H}^{LL}$ and $\mathbf{H}^{LC}$ it holds that the sum of the row elements is equal to 1, i.e. a total weight of 100% is distributed for the computation of the corresponding fitted value.

In contrast to OLS the hat matrices for a nonparametric regression are usually neither symmetric[4] nor idempotent. Hence, the column sums are not necessarily equal to 1. This fact allows for a detailed analysis of the structure of each observation, as observations with column sum smaller (larger) than 1 are relatively unimportant (important) for the computation of all fitted values and thus likely for the nonparametric estimation in general (for the given data structure).

4.3 Behavior and properties of hat matrices for nonparametric kernel regression

In the previous subsection we have seen that the hat matrix for nonparametric kernel regression consists of four building blocks, the type of local estimation (constant or linear), the kernel functions (for continuous and discrete covariates), the smoothing parameters, and the covariate values. In this section we analyze the behavior of hat matrices for local constant versus local linear estimation and varying bandwidths of continuous (subsections 4.3.1 and 4.3.2) and discrete covariates (subsection 4.3.3). For this purpose the tools for hat matrix analysis are introduced by way of example.

I generate artificial data[5] for $n = 100$ observations of two covariates, a continuous

[4] This can be easily seen for $\mathbf{H}^{LC}$, as the numerators are symmetric, but the denominators are not, unless all denominators are equal. Here, one can also easily see that the sum of the row entries is always equal to 1, as the sum of numerators for all fractions in a row equals the corresponding denominator.

[5] Whenever a seed is advisable for reproducing data and estimations, a seed of 42 is used.

covariate x_c which is distributed according to the standard normal density and a binary covariate x_d with equal probabilities of entry for $x_d = 0$ and $x_d = 1$. Concerning the nonparametric configuration, x_c is weighted by a second order Gaussian kernel and x_d by the (unordered) kernel of Li and Racine. In the following, three different bandwidths for weighting each of the covariates are considered. I use a moderate bandwidth of $h_c = 0.4$, as well as a relatively large (small) bandwidth value of 0.4 times (divided by) 20, thus $h_c = 8$ ($h_c = 0.02$). The moderate bandwidth is close to that selected by the normal reference rule-of-thumb[6] for kernel density estimation, compare Li & Racine (2007, page 14). For h_d the values 0, 0.5, and 1, are used.

4.3.1 Effect of different bandwidths for continuous covariates

First, the behavior of $\mathbf{H}^{LC}$ and $\mathbf{H}^{LL}$ for the three values of h_c and a value of $h_d = 1$ (irrelevant x_d) is analyzed.

Table 4.1: Summary statistics of the main-diagonal elements of the hat matrix for local constant kernel regression and various bandwidths of the continuous covariate.

h_c	Minimum	1.Quartile	2.Quartile	3.Quartile	Maximum	Sum
0.02	0.152	0.354	0.479	0.630	1.000	52.120
0.4	0.027	0.028	0.032	0.051	0.406	5.425
8	0.010	0.010	0.010	0.010	0.011	1.017

Table 4.1 contains Tukey's five and the sum of the main-diagonal elements for $\mathbf{H}^{LC}$. We can see that the main-diagonal entries take values in $[0.01, 1]$. Generally, for local constant and local linear estimators it holds that the diagonal entries take values in $[1/n, 1]$, where a value of 1 is obtained, if only the diagonal element is used in row j, to

[6] Clearly, this normal reference rule-of-thumb is designed for kernel density estimation. Here, for analyzing the hat matrix, we are in a regression context. Selecting bandwidths for a regression context requires to take care for the underlying relationship between covariates and explanatory variable. As I only want to analyze the behavior of hat matrices for nonparametric kernel regression, nothing is assumed on the underlying relationship between covariates and explanatory variable. Hence, $h_c = 0.4$ is by no means an optimal bandwidth in some sense, it is only used as a moderate bandwidth for the covariate x_c.

compute the jth fitted value. A value of $1/n$ is obtained if all entries in a row are equal, thus the corresponding fitted value is simply the arithmetic mean of all observations. From Table 4.1 we can see that the latter case is (almost) given for all rows of $\mathbf{H}^{LC}$ for a relatively large bandwidth of $h_c = 8$, while smaller bandwidths cause larger main-diagonal entries. This is in line with the increased model complexity that is obtained by a smaller bandwidth. As the main-diagonal elements of $\mathbf{H}^{LC}$ take values in $[1/n, 1]$, $tr(\mathbf{H}^{LC}) \in [1, n]$. Here, for the local constant estimator and $h_c = 8$ the trace is close to 1, implying that all fitted values are (almost) equal to each other and to the mean of $\mathbf{y}$. This is equivalent to OLS, when only an intercept is included and no covariates. For $h_c = 0.02$ we have a rather large model complexity of 52.12, which corresponds to more than half of the observations.

While the last column of Table 4.1 shows $tr(\mathbf{H}^{LC})$, the corresponding trace for the local linear estimator is higher (due to higher main-diagonal entries on average) with $tr(\mathbf{H}^{LL})$ of 62.027, 7.122, and 2.033 for the small, moderate, and large h_c, respectively. Hence, for the same bandwidths the local linear estimator yields a somewhat increased model complexity which is in line with the additional first derivatives that are estimated for the continuous covariates. The trace of 2.033 is close to 2 which is the lower bound of $tr(\mathbf{H}^{LL})$ for the given case with only one continuous covariate. This case is equivalent to an OLS framework with an intercept and a linear inclusion of x_c, where we thus have to estimate two parameters. In general it holds that $tr(\mathbf{H}^{LL}) \in [1 + C, n]$.

The off-diagonal elements of $\mathbf{H}^{LL}$ can take negative values. These negative weights are in absolute values smaller than 0.05 for all considered h_c-cases. Table 4.2 depicts the structure of the off-diagonal elements of $\mathbf{H}^{LL}$ for five weight categories. The first weight

Table 4.2: Percentage of the off-diagonal elements of the hat matrix for local linear kernel regression in five weight-categories and various bandwidths of the continuous covariate.

h_c	$H_{ji} \leq -\frac{1}{n^2}$	$-\frac{1}{n^2} < H_{ji} < 0$	$H_{ji} = 0$	$0 < H_{ji} < \frac{1}{n^2}$	$H_{ji} \geq \frac{1}{n^2}$
0.02	0.81	16.94	59.26	19.10	3.89
0.4	4.77	25.09	0.00	12.43	57.71
8	10.52	0.20	0.00	0.09	89.19

category shows the percentage of off-diagonal elements with a substantial negative weight (a weight of less than $-n^{-2}$), while the fifth category shows the percentage of entries with a substantial positive weight. The three categories in-between show the percentage of almost-zero weights. Of course, the borders of these categories are somewhat arbitrary, but this Table is only intended to show the weight structure and the implications do not heavily change for different borders. We see that with increasing h_c the percentage of substantial positive and negative weights increases. The 59.26 percent of zero entries is due to the singularity in some rows of the hat matrix where the corresponding row is manually replaced by the unit vector (compare subsection 4.2.3).

The structure of the rows of the hat matrix for local constant and local linear estimation is depicted in Figure 4.1. The upper/center/lower panels correspond to small/moderate/large bandwidths for x_c, while the left panels show $\mathbf{H}^{LC}$ and the right panels $\mathbf{H}^{LL}$. Each grey line shows the weight structure of one row of a hat matrix, while the thick red line exemplifies the weight structure of the row that corresponds to the tenth-smallest x_c-value. The black points show the main-diagonal elements. We can see a U-shaped structure of the main-diagonal elements for the case of a moderate bandwidth. This structure also appears for larger bandwidths (though not visible for the local constant case because of the chosen ordinate-scale). By reducing the bandwidth substantially, this structure is lost. From the right panels we can also see the negative off-diagonal elements in $\mathbf{H}^{LL}$ which are mainly visible in the lower right panel, where the rows of $\mathbf{H}^{LL}$ possess a linear (in x_c) structure which is equivalent to that of the corresponding OLS framework. From this linear structure we can also see that in the local linear framework, the main-diagonal entry is not necessarily the largest row entry.

The negative off-diagonal elements are an important issue for the analysis of hat matrices for local linear estimation. Negative weights cause that more than 100% of positive weight can be distributed to the elements of a single row. Even though the sum of the row elements in $\mathbf{H}^{LL}$ is always equal to 1, the sums of row elements for the hat matrix in absolute values are larger than 1 whenever a row contains negative elements. For the local linear estimator we analyze the ordinary hat matrix $\mathbf{H}^{LL}$, but also the hat matrix with elements in absolute values, denoted as $\mathbf{H}^{LL,abs}$. For $h_c = 0.02$, $h_c = 0.4$, and $h_c = 8$ we obtain means for the rows of $\mathbf{H}^{LL,abs}$ of 1.009, 1.010, and 1.167, as well as

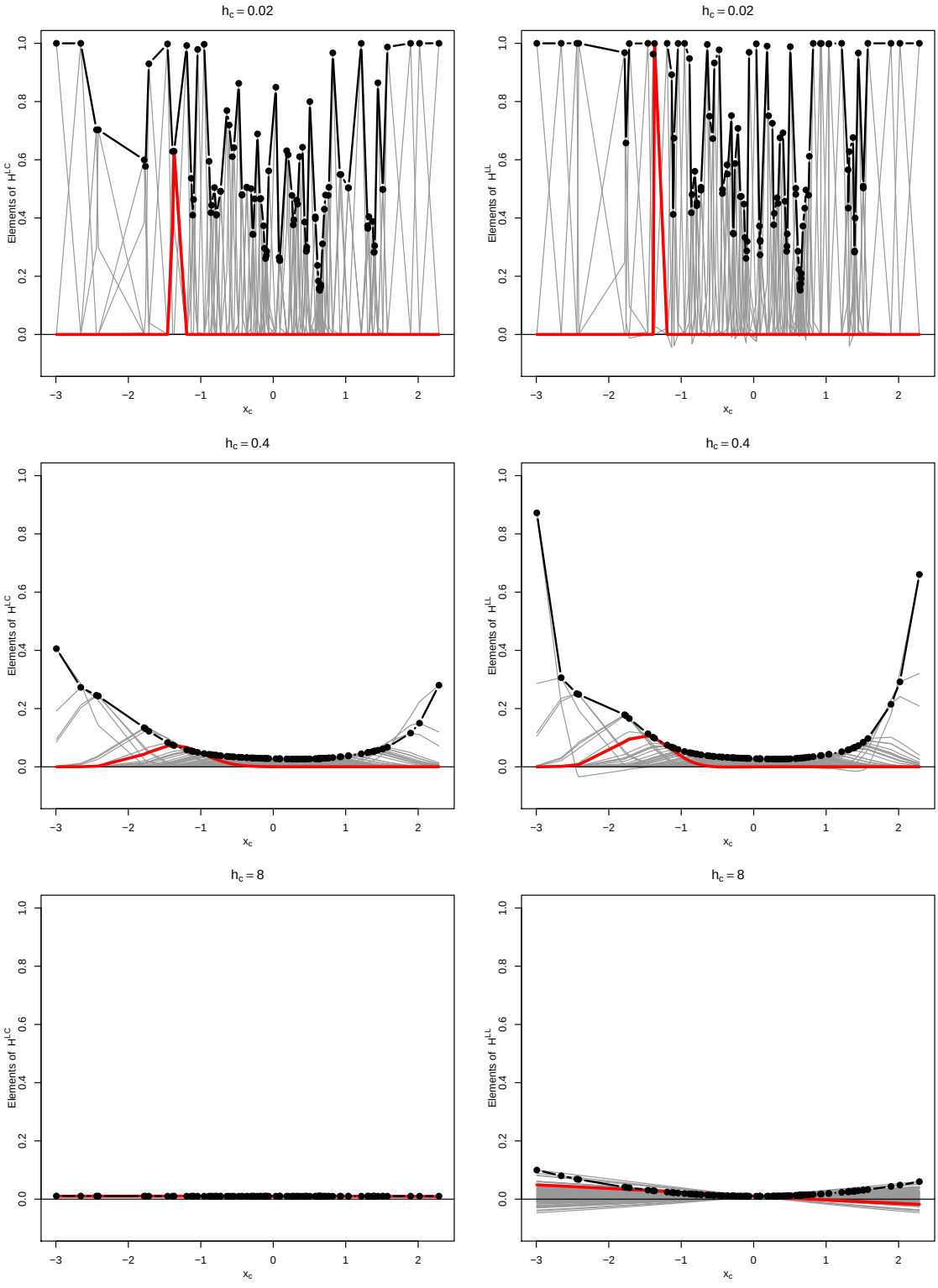

Figure 4.1: Elements of the hat matrix for local constant (left panels) and local linear (right panels) kernel regression and various bandwidths of the continuous covariate.

maximum values of 1.091, 1.331, and 2.344. Hence, the weight that is distributed within a single row raises with increasing bandwidth. This can already be expected from the grey lines in the right panels of Figure 4.1, as especially the grey lines in the lower right panel correspond to a lot of negative hat matrix entries.

4.3.2 Identification of (potential) overfitting observations and CODI-plot

Next, I introduce the CODI-plot that relates the sum of the hat matrix COlumns to the main-DIagonal elements. This graphical tool allows to illustrate the complete relevant (for a comparison of different nonparametric configurations) structure and figures for analyzing the hat matrix of a nonparametric configuration in a single graph and can be used irrespective of the number of covariates.

The left panels of Figure 4.2 show the CODI-plots for $\mathbf{H}^{LC}$, the right panels for $\mathbf{H}^{LL,abs}$. The abscissa scale is identical for all six panels, the ordinate scale is adjusted for the lower panels that show the results for $h_c = 8$.

For local constant estimation a diagonal entry of larger than 0.5 means that for the computation of the jth fitted value, the weight of the "own" observation is higher than the sum of all other weights in a row. In the following I denote observations with diagonal values of larger than 0.5 as "potential overfitting observations", those with main-diagonal values of larger than 0.8 as "overfitting observations". Originally, the term overfit means that an estimation achieves a (relatively) very good fit at the given sample of observations, but is e.g. not able to reasonably predict for other observations of the same data generating process.

The horizontal dashed-dotted purple (dashed red) line is at an ordinate value of 0.5 (0.8). The corresponding purple (red) numbers show the percentage of main-diagonal entries that exceed the purple (red) line, thus the percentage of potential overfitting observations (overfitting observations). The diagonal black line in the panels shows a natural restriction as the column sum cannot be smaller than the corresponding main-diagonal entry (by using $\mathbf{H}^{LL,abs}$ instead of $\mathbf{H}^{LL}$). Hence, observations close to this diagonal line are only important for the estimation at the "own" covariate-position and not for other covariate positions.

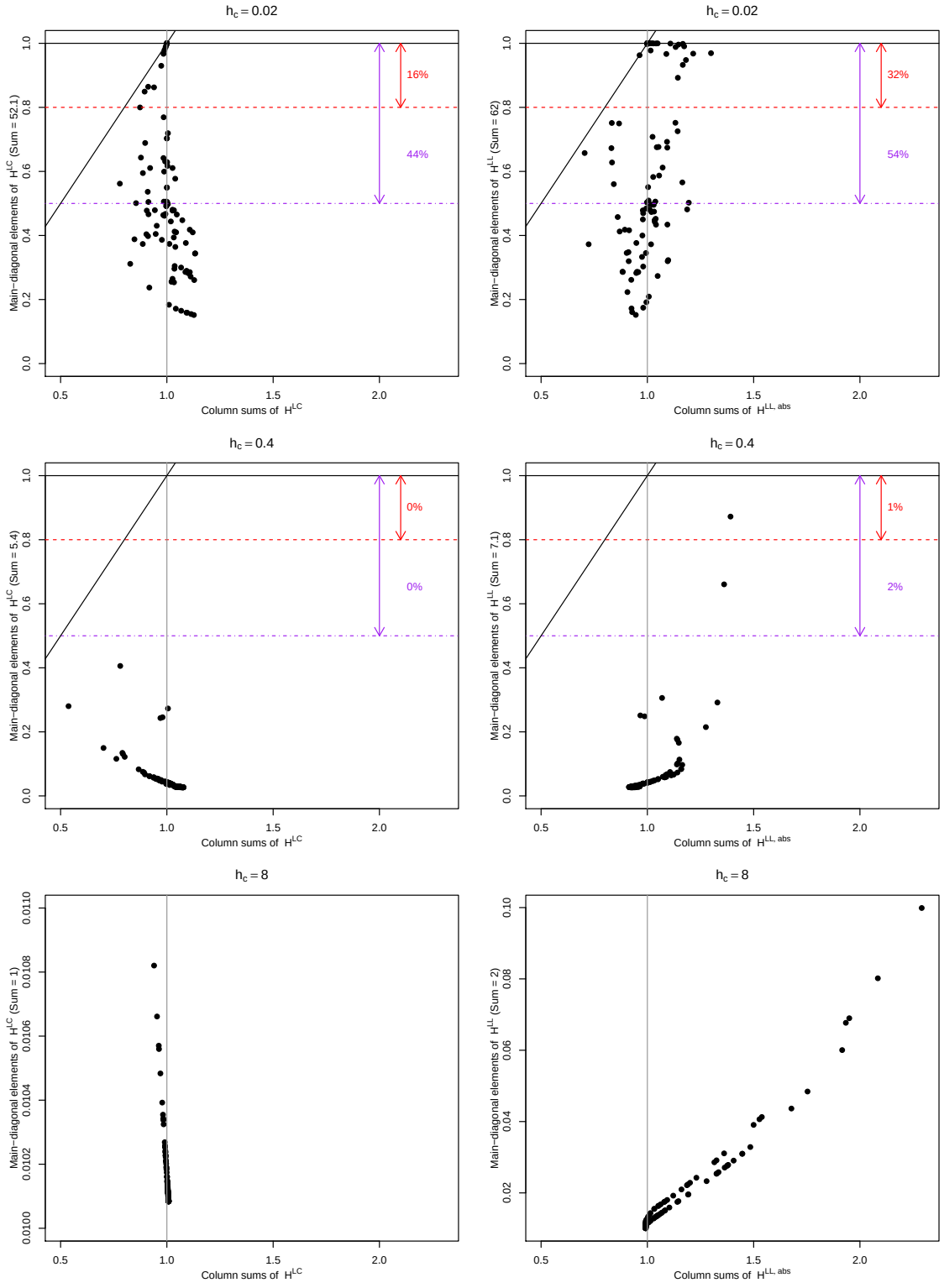

Figure 4.2: CODI-plots: Scatterplots of column sums against main-diagonal elements for the hat matrix of local constant (left panels) and local linear (right panels where all hat matrix elements are in absolute values) kernel regression and various bandwidths of the continuous covariate.

From the panels we see that the overfitting observations mainly occur for the case of a small bandwidth. Also, the local linear estimator has higher percentages which is in line with the increased model complexity for this estimator (if the same bandwidth h_c is used for both estimators). The panels for $\mathbf{H}^{LL,abs}$ show a somewhat different structure than those for $\mathbf{H}^{LC}$ as for a local linear estimation the largest column sum values appear for the overfitting observations[7]. Nevertheless, the percentage of (potential) overfitting observations as well as whether the points are closely located to the black diagonal line is important for judging a nonparametric configuration, as the estimated regression function will possess a high variance in the vicinity of such observations. Additionally, we can see from the CODI-plots which observations are very important (unimportant) for the computation of all fitted values, whenever the corresponding abscissa value, the column sum, is much larger (smaller) than 1.

In summary, a CODI-plot shows the structure of the hat matrix for a nonparametric estimation in a single graph. On the abscissa, the columns sums are plotted (for absolute valued $\mathbf{H}$-entries whenever local linear estimation is used), the ordinate-axis refers to the main-diagonal elements. We can see the trace of the hat matrix (compare the ordinate-label) and the percentages of (potential) overfitting observations (if relevant) from the CODI-plots. Also, a diagonal line with intercept 0 and slope 1 is added, that the observations cannot exceed.

4.3.3 Effect of different bandwidths for discrete covariates

Finally, the effect of different bandwidths for discrete covariates on the hat matrices for nonparametric kernel regression is analyzed. In Table 4.3 we have the trace of the hat matrices for the corresponding bandwidth combinations. We already know the entries in the columns with $h_d = 1$ as they are equivalent to the results of subsection 4.3.1. Starting from these columns we see that a lower bandwidth for x_d leads to a higher trace of the hat matrix which is analogous to the effect of lower bandwidths for a continuous covariate. For local constant estimation, $h_c = 8$, and $h_d = 0$, we obtain a trace of close

[7] This is partially due to the negative weighting of observations in local linear estimation, partially due to the very simple structure of the example with only one continuous covariate. In a multivariate framework, these structures are in general not visible.

Table 4.3: Sum of the main-diagonal elements of the hat matrix for both types of local estimation and various bandwidths of both covariates.

	local constant			local linear		
$h_c \setminus h_d$	0	0.5	1	0	0.5	1
0.02	69.408	58.020	52.120	80.565	67.509	62.027
0.4	10.671	7.070	5.425	13.209	9.038	7.122
8	2.035	1.338	1.017	4.063	2.692	2.033

to 2 which corresponds to a case where the continuous covariate is irrelevant and we have a separate estimation for both categories of x_d. This means that the estimated fitted response value for an observation with $x_d = 0$ is the mean of all observations that have $x_d = 0$ (analogously for $x_d = 1$). For a local linear estimation, $h_c = 8$, and $h_d = 0$ we get a trace of close to 4. This is equivalent to OLS estimation when we separately estimate intercept and slope (for x_c) for both x_d-categories, implying 4 estimated parameters.

In the subsequent paragraphs, the effect of different h_d on the rows and columns of the nonparametric hat matrices is demonstrated. For a tight argumentation, I restrict the analysis to the case of a moderate bandwidth for the continuous covariate ($h_c = 0.4$), but the implications remain qualitatively unchanged for other h_c-values.

The effect of different discrete bandwidths (for a moderate h_c) on a row of $\mathbf{H}^{LC}$ (for local linear estimation we obtain analogous results) can be seen from Figure 4.3 where again the row of the observation with the tenth-smallest x_c-value is exemplified. The points/triangles in the left (right) panel show the row entries of $\mathbf{H}^{LC}$ for $h_d = 0$ ($h_d = 0.5$). In both panels, the red line shows the case of an irrelevant discrete covariate (i.e. $h_d = 1$). The black points (grey triangles) show the weight of the observations with the same (other) x_d-value as the observation with the tenth-smallest x_c. In the left panel we can see that all grey triangles have an ordinate value of zero, as we have a separate (frequency approach) estimation for the observations of both categories of x_d. The points/triangles in the right panel are located much closer to the red line than in the left panel and the grey triangles do not all have zero weight. Here, we see some smoothing of the discrete covariate x_d, i.e. the observations of the opposite x_d-category are also used for estimation, but with only half the weight of an observation of the correct

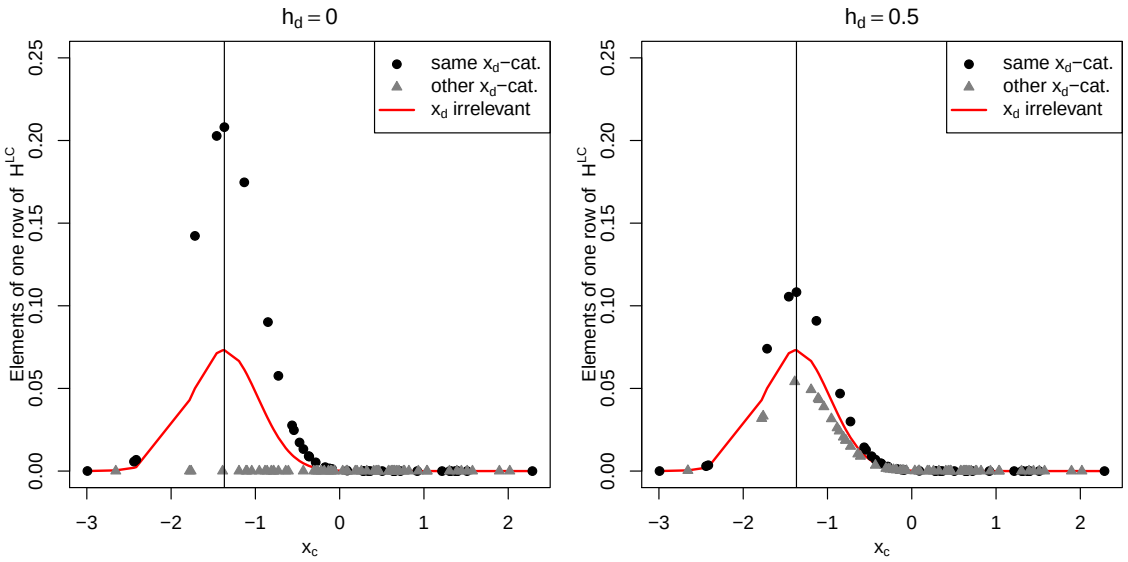

Figure 4.3: One row of the hat matrix for local constant kernel regression for a moderate bandwidth of the continuous covariate and various bandwidths of the discrete covariate.

category.

Figure 4.4 shows the CODI-plots, introduced in subsection 4.3.2, for different h_d-values. The left (right) panels show the hat matrix results for local constant (linear) estimation, while the upper (lower) panels show the results for $h_d = 0$ ($h_d = 0.5$). The results for $h_d = 1$ are already shown in the center panels of Figure 4.2. By a comparison of the upper panels to the lower panels we can see that the structure basically remains unchanged by varying h_d, a h_d-value lower than 1 leads to a shift of some of the observations mainly towards higher ordinate-values. Hence, as would be expected, the percentage of (potential) overfitting observations is going to rise with decreasing h_d.

4.4 Canadian housing example

In section 4.3 I presented the effects of bandwidths and local estimation type on the nonparametric hat matrix and suggested various tools for hat matrix analysis. In this

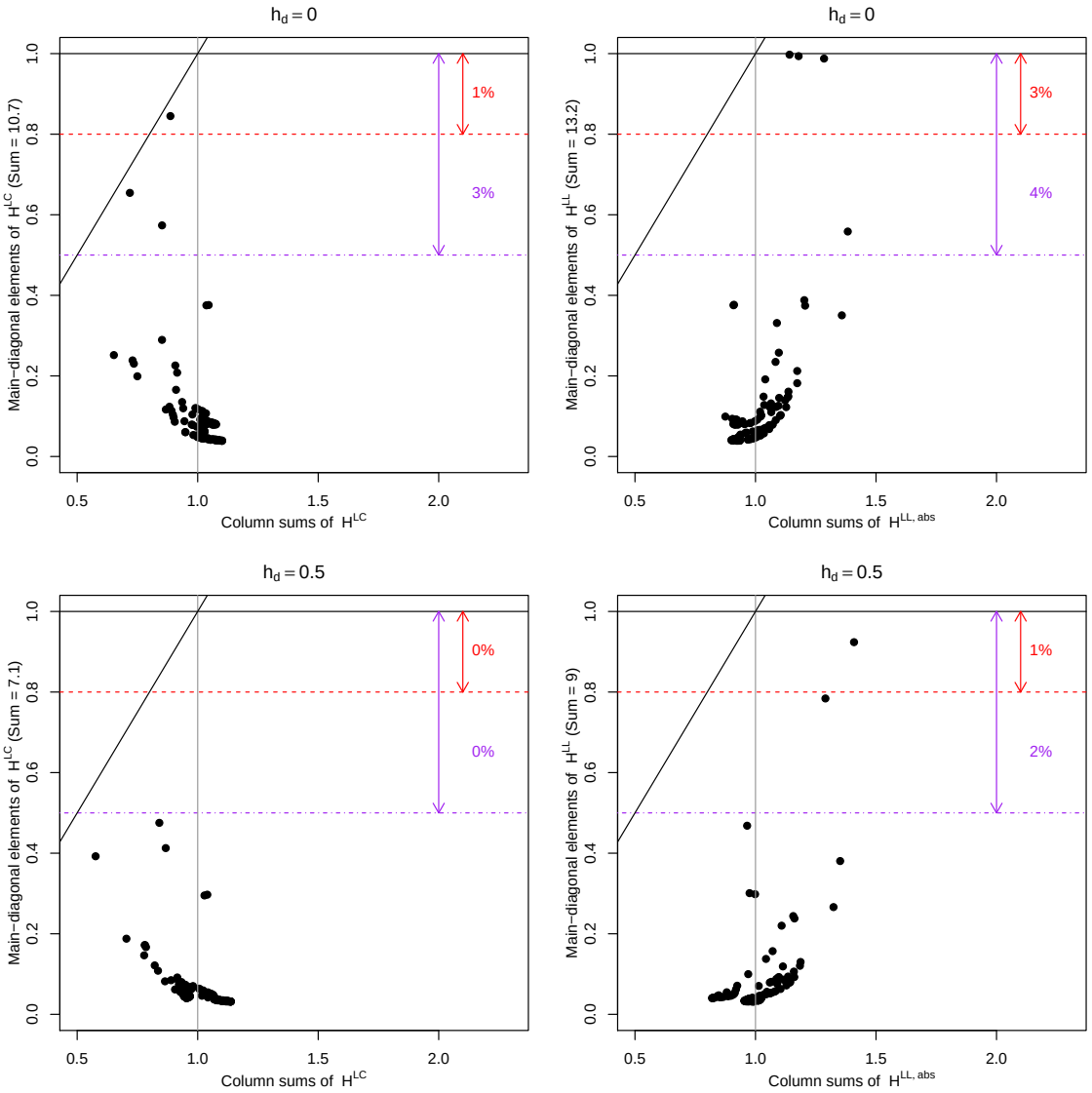

Figure 4.4: CODI-plots: Scatterplots of column sums against main-diagonal elements for the hat matrix of local constant (left panels) and local linear (right panels where all hat matrix elements are in absolute values) kernel regression, a moderate bandwidth for the continuous covariate, and various bandwidths of the discrete covariate.

section, these tools are applied to the Canadian housing data set[8] of Anglin & Gencay (1996) for analyzing the nonparametric configurations found in Table I. of Haupt et al. (2010a). Various p-values for the test of Hsiao et al. (2007) for parametric misspecification are displayed, ranging from 0.0226 to 0.2607, indicating that for a significance level of 5% the different employed nonparametric configurations lead to different outcomes of the test. Clearly the problem is, that we do not know which of the 16 configurations is preferable.

Table 4.4: Nonparametric configurations for Canadian housing data set.

config.	Discr. Kernels	**h**-selection	Estimation type	x_c
1	AA-WvR	LSCV	LC	lot
2	AA-WvR	LSCV	LC	$lnlot$
3	AA-WvR	LSCV	LL	lot
4	AA-WvR	LSCV	LL	$lnlot$
5	AA-WvR	AIC_c	LC	lot
6	AA-WvR	AIC_c	LC	$lnlot$
7	AA-WvR	AIC_c	LL	lot
8	AA-WvR	AIC_c	LL	$lnlot$
9	LR-LR	LSCV	LC	lot
10	LR-LR	LSCV	LC	$lnlot$
11	LR-LR	LSCV	LL	lot
12	LR-LR	LSCV	LL	$lnlot$
13	LR-LR	AIC_c	LC	lot
14	LR-LR	AIC_c	LC	$lnlot$
15	LR-LR	AIC_c	LL	lot
16	LR-LR	AIC_c	LL	$lnlot$

Table I. of Haupt et al. (2010a) contains 16 configurations for the nonparametric kernel regression summarized in Table 4.4 that differ with respect to the type of local

[8] The data set consists of $n = 546$ observations for the response variable $lnsell$, i.e. the logarithmized sale price of a house. There are six binary covariates ($ca, drv, ffin, ghw, rec, reg$), four ordered discrete covariates ($bdms, fb, gar, sty$), and as continuous covariate the lot size (untransformed as lot or logarithmized as $lnlot$). The discrete covariates allow for 24,576 category combinations.

estimation (LC: local constant versus LL: local linear), the discrete kernel type (AA-WvR: kernels of Aitchison & Aitken (1976) for unordered and of Wang & van Ryzin (1981) for ordered discrete covariates, versus LR-LR: kernels of Li and Racine for both discrete variable types), the data-driven bandwidth selection procedure (LSCV: least-squares cross-validation, versus AIC_c: approach of Hurvich et al. (1998)), and finally whether the continuous covariate is logarithmized or not.

4.4.1 Analysis of fit and estimated bandwidths

Fit, estimated bandwidths, and the hat matrix are the pieces of information, that are directly available after the nonparametric estimation. As a first step, fit[9] and estimated bandwidths of all nonparametric configurations are analyzed, compare Table 4.5. Concerning the fit, we can see that the configurations with bandwidths selected by least-squares cross-validation always have a better fit than the corresponding configurations with AIC_c-selected bandwidths. The discrete kernel type does not matter much in terms of fit (though it is better for the configurations with AA-WvR), apart from configuration 4 (versus 12). Whether the continuous variable enters in logs or levels does almost not influence the fit of the local linear configurations, but the fit of the local constant configurations always clearly increases whenever the lot size is logarithmized.

The bandwidth of the continuous covariate is always larger when bandwidth selection is conducted by AIC_c (indicating a smaller model complexity). The standard deviation of lot ($lnlot$) is equal to 2168.2 (0.4), hence for most of the 16 cases the obtained bandwidth when lot is used, is of a similar size relative to the standard deviation, as when $lnlot$ is used. For the cases 8 and 16 (7 and 15) we get a very large (relatively large) bandwidth indicating a linear (almost linear) relationship between $lnlot$ and $lnsell$ for these cases[10].

An interpretation of the bandwidths for the discrete covariates is quite demanding as the discrete kernels have different properties (compare subsection 4.2.2), e.g. different upper bounds for the bandwidths. Therefore, discrete bandwidths are not directly analyzed, but

[9] Fit is measured as PR^2, the squared correlation between observed and fitted response values.

[10] This is already commented in Haupt et al. (2010a).

Table 4.5: Estimated bandwidths and PR^2 for Canadian housing data set.

config.	ca	drv	$ffin$	ghw	rec	reg	$bdms$	fb	gar	sty	$(ln)lot$	PR^2
1	0.05	0.02	0.25	0.02	0.04	0.13	0.33	0.00	1.00	0.27	943.20	0.74
2	0.07	0.00	0.23	0.02	0.05	0.11	0.21	0.00	0.90	0.35	0.15	0.87
3	0.12	0.00	0.16	0.02	0.21	0.17	0.10	0.02	0.71	0.19	1415.86	0.84
4	0.09	0.00	0.24	0.05	0.04	0.11	0.16	0.00	0.61	0.31	0.19	0.89
5	0.17	0.12	0.15	0.04	0.50	0.30	0.62	0.36	0.92	0.62	1471.98	0.67
6	0.18	0.13	0.14	0.06	0.50	0.27	0.66	0.37	0.82	0.56	0.30	0.78
7	0.11	0.13	0.13	0.03	0.50	0.27	0.56	0.26	1.00	0.51	14054.90	0.76
8	0.14	0.12	0.13	0.03	0.50	0.23	0.62	0.27	1.00	0.51	1021539.38	0.76
9	0.05	0.02	0.34	0.02	0.05	0.14	0.19	0.00	0.67	0.20	957.08	0.73
10	0.07	0.00	0.35	0.02	0.06	0.12	0.12	0.00	0.61	0.23	0.16	0.86
11	0.15	0.00	0.17	0.03	0.35	0.19	0.07	0.01	0.45	0.10	1354.37	0.84
12	0.10	0.00	0.29	0.01	0.22	0.22	0.09	0.03	0.79	0.08	0.47	0.83
13	0.20	0.13	0.17	0.04	1.00	0.32	0.35	0.20	0.62	0.47	1420.17	0.66
14	0.22	0.15	0.15	0.06	1.00	0.28	0.38	0.21	0.54	0.42	0.29	0.77
15	0.12	0.14	0.14	0.02	1.00	0.20	0.33	0.13	0.78	0.40	9139.81	0.75
16	0.16	0.12	0.14	0.03	1.00	0.19	0.35	0.14	0.69	0.38	1334167.64	0.75

by using an alternative measure, the "percentage of smoothing"

$$pos_k = \frac{w_k(x_{0,k}, x_{i,k} \neq x_{0,k}, h_k)}{w_k(x_{0,k}, x_{i,k} = x_{0,k}, h_k)} \cdot 100.$$

Hence pos_k is the weight of an observation of the "wrong" category relative to the weight of an observation of the "correct" category for a certain bandwidth h_k. For ordered discrete covariates, I define "wrong" category as $|x_{i,k} - x_{0,k}| = 1$. The percentages of smoothing for the discrete covariates are contained in Table 4.6. The largest differences

Table 4.6: Percentage of smoothing of the discrete covariates for Canadian housing data set.

config.	ca	drv	$ffin$	ghw	rec	reg	$bdms$	fb	gar	sty
1	5.3	2.3	33.2	2.2	4.5	15.0	16.3	0.0	50.0	13.3
2	7.2	0.4	30.5	2.3	5.6	12.1	10.7	0.1	45.0	17.7
3	13.6	0.3	19.6	2.1	26.6	20.3	5.0	1.1	35.5	9.5
4	10.4	0.0	31.5	5.0	3.9	12.7	8.1	0.1	30.4	15.4
5	20.5	13.1	17.9	4.4	100.0	43.0	31.1	18.0	46.0	31.0
6	22.6	15.6	15.8	6.2	100.0	36.3	33.1	18.3	40.8	27.8
7	12.8	15.0	15.2	3.2	100.0	36.3	28.2	12.9	50.0	25.5
8	16.4	13.7	14.9	3.1	100.0	29.1	30.8	13.3	50.0	25.3
9	5.2	2.5	33.7	2.4	4.8	13.9	19.5	0.0	67.3	19.7
10	7.1	0.4	35.4	2.4	5.6	12.1	12.4	0.2	60.6	23.0
11	15.3	0.3	17.1	2.8	35.3	18.7	7.4	1.3	44.9	9.7
12	9.9	0.1	29.1	1.3	21.6	21.8	8.9	3.4	79.5	8.0
13	20.2	12.5	17.2	3.8	100.0	31.6	35.4	20.3	62.0	47.2
14	22.3	14.7	15.2	5.5	100.0	27.7	37.9	21.0	54.0	41.6
15	12.2	14.1	14.4	2.5	100.0	19.8	33.4	13.1	77.6	40.5
16	15.7	12.1	13.6	2.9	100.0	19.4	35.3	14.4	68.8	38.2

occur for the different bandwidth selection methods, where for example the covariate rec is irrelevant whenever the bandwidth is selected via AIC_c. Variation in the discrete kernel type causes only small differences for the binary covariates (again except for the cases 4 and 12). For ordered covariates there are large differences for gar, which gets a smoothing of up to nearly 80% when the corresponding Li and Racine kernel is used.

Haupt et al. (2010a) have shown that the kernel of Wang & van Ryzin (1981) has limited smoothing abilities. It is not possible to get a percentage of smoothing of more than 50% using this kernel, thus covariates also cannot be smoothed out. The transformation of the continuous covariate does not seem to have a large impact on the discrete covariates, especially the smoothing for the covariates drv, ghw, and fb is relatively stable. Local linear estimation instead of local constant estimation leads to some different bandwidths, but overall the local estimation type does not seem to be as influential as the bandwidth selection approach for this data set.

After inspecting the estimated bandwidths and the fact that for some of the cases the upper bound of the bandwidth for the Wang and van Ryzin kernel is reached for the covariate gar, the approaches that use the kernels of Li and Racine seem preferable[11]. Hence, by a simultaneous inspection of fit and estimated bandwidths, one would take one of configurations 9-16, probably one of 9-12 due to the better fit.

4.4.2 Hat matrix analysis

Next, I show the additional/new information obtained by a thorough hat matrix analysis. Smaller values of estimated bandwidths indicate a rising model complexity (compare section 4.3). However, in a multiple mixed kernel framework, the vector of estimated bandwidths cannot serve as an estimator for the model complexity as it is not clear how the vector information should be transferred into a scalar measure of model complexity. The hat matrix allows for estimating the model complexity by its trace. Additionally, as it contains the weight of each observation for the computation of every fitted value, it allows for the detection of (potential) overfitting observations and also indicates, whether observations are more/less important for the computation of all fitted values (visible in CODI-plot).

Table 4.7 shows the trace of the hat matrices (also in percentage of observations) and the percentage of potential overfitting observations and overfitting observations. We can

[11] Even for the configurations with $pos < 50$, the upper bound of the Wang and van Ryzin-kernel could have affected the numerical optimization within the bandwidth selection process. Hence, in a regression context, the ordered kernel of Li and Racine should be used instead, as it allows for a smoothing between 0 and 100 percent and comes without costs.

Table 4.7: Trace of the hat matrix (absolute as well as in percentage of observations) and percentage of (potential) overfitting observations for Canadian housing data set.

config.	$tr(\mathbf{H})$	$100 \cdot tr(\mathbf{H})/n$	$\%\{H_{ii} > 0.5\}$	$\%\{H_{ii} > 0.8\}$
1	216.4	39.6	34.6	15.6
2	243.2	44.5	39.6	18.9
3	240.0	44.0	40.3	19.8
4	271.5	49.7	45.8	26.0
5	89.8	16.5	7.1	2.0
6	90.7	16.6	6.0	1.5
7	88.4	16.2	7.3	1.6
8	88.5	16.2	5.9	1.8
9	202.8	37.1	31.1	13.6
10	227.5	41.7	35.0	15.9
11	232.7	42.6	38.1	18.5
12	180.8	33.1	25.6	11.9
13	84.0	15.4	6.4	1.8
14	84.7	15.5	4.8	1.5
15	80.0	14.7	5.3	1.6
16	82.1	15.0	5.5	1.8

see that the configurations where the bandwidths are selected via LSCV have a much higher trace and thus model complexity. Also the percentages of (potential) overfitting observations are clearly higher. As these results indicate a very high variability of the estimated regression functions for the LSCV configurations, the configurations 13-16 now seem preferable, since they also use the kernels of Li and Racine (although configuration 13 could be excluded, as it has a much lower fit, but nearly the same model complexity and even slightly more potential overfitting observations).

Figure 4.5 (4.6) shows the CODI-plots for configurations 9-12 (13-16). We observe that the distribution is widespread for the local linear configurations 11, 12, 15, and 16, which is mainly due to the negative weights. The results from Table 4.7 are also visible in Figures 4.5 and 4.6, e.g. a lot of (potential) overfitting observations for the configurations 9-12. In all eight cases (9-16) there are observations that have a rather low impact on the estimation as their column sum is only about 0.5. For the local linear configurations 15 and 16 we can see observations that have a very large impact on the estimation (column sums of about 3). Column sums of more than 2 are only obtained for configurations 7, 8, 15, and 16. These are the only configurations where the continuous covariate enters (almost) linearly. We have already seen in section 4.3 that very large bandwidths for a continuous covariate can lead to a lot of negative row entries for a hat matrix in a local linear framework, and thus to a large column sum of $\mathbf{H}^{LL,abs}$ (compare lower right panel in Figure 4.2).

The results of the hat matrix analysis show that the most reliable (no overfit) nonparametric configurations seem to be those of configurations 14, 15, and 16 (and not those of configurations 9-12). The decision, which of the corresponding p-values of configurations 14-16 is adequate (p-values between 0.0526 and 0.1328) is left open as this would require to analyze the different bootstrapping procedures of Table I. of Haupt et al. (2010a) as well as the residuals of the tested parametric specification.

Hence, a thorough analysis of fit, estimated bandwidths, and the hat matrix (all the information which is available after the nonparametric estimation) allow for a reduction of the set of 16 configurations to a set of three configurations. The next step might be a simulation-based prediction performance comparison (compare Haupt et al., 2010a,b; Henderson & Millimet, 2008), that usually has high computational costs, which are clearly

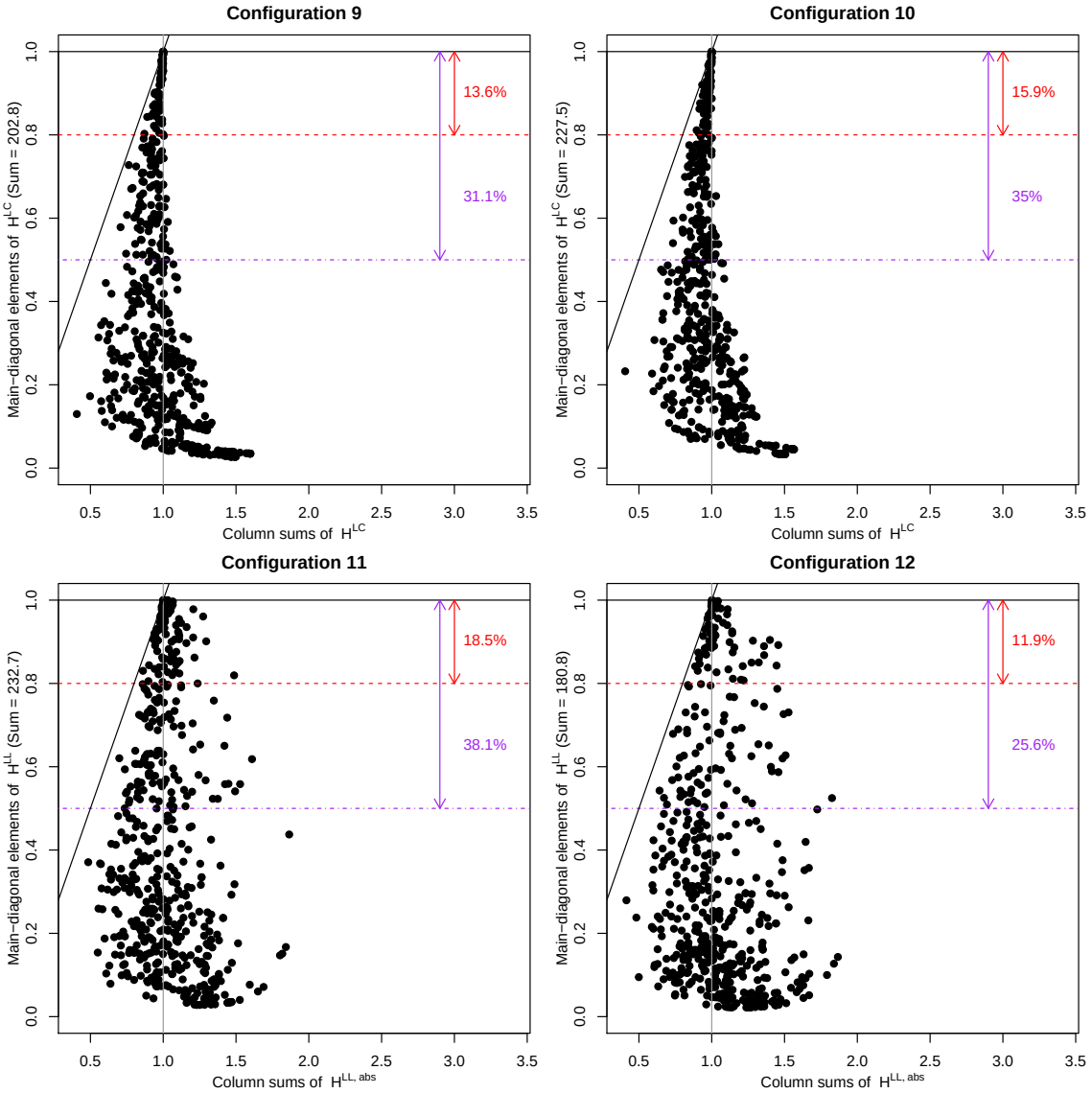

Figure 4.5: CODI-plots: Scatterplots of column sums (computed for absolute values of the hat matrix elements for local linear configurations) against main-diagonal elements for the hat matrix of configurations 9-12 for the Canadian housing data set.

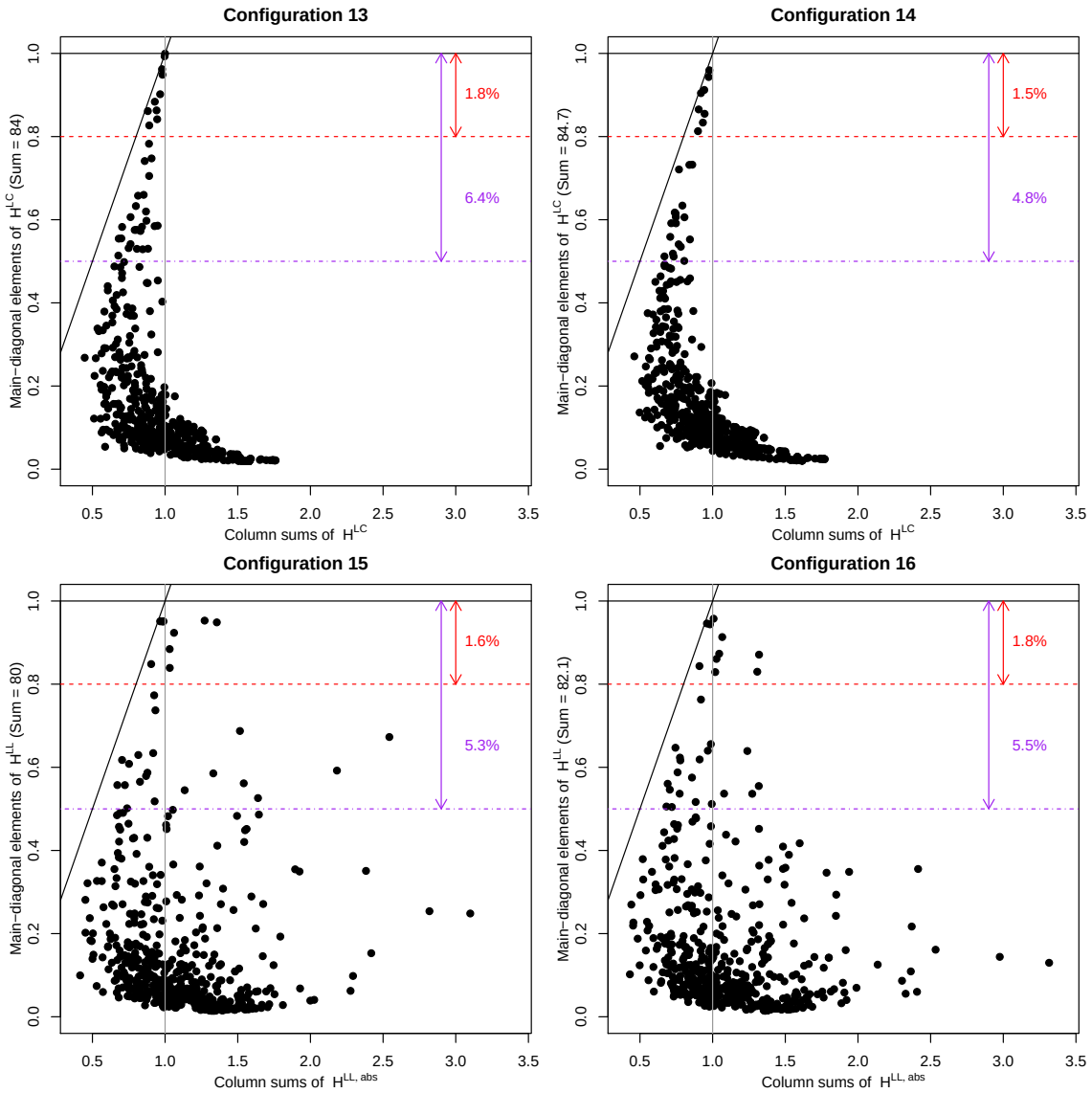

Figure 4.6: CODI-plots: Scatterplots of column sums (computed for absolute values of the hat matrix elements for local linear configurations) against main-diagonal elements for the hat matrix of configurations 13-16 for the Canadian housing data set.

lower if only three nonparametric configurations have to be considered (instead of 16).

4.5 Conclusion

In this work I analyze the structure of hat matrices for multiple nonparametric kernel regression in a setting of mixed covariates. In contrast to existing work this analysis uses information beyond the trace of this matrix. As a visualization tool the CODI-plot is introduced, summarizing and condensing all relevant information of a hat matrix for selecting a nonparametric configuration. As further statistics aiding this selection process I discuss the percentage of overfitting observations and the percentage of smoothing, respectively. A data set of simulated covariates is used to analyze the impact of varying bandwidths and local estimation and explain the interpretation of the CODI-plot. As empirical illustration I show for a well-known Canadian housing data set how the proposed tools for hat matrix analysis can be used to improve the selection of nonparametric configurations beyond the insights from the existing literature.

5 Cross-validating fit and predictive accuracy of nonlinear quantile regressions

This essay is joint work with Harry Haupt[1] and Kathrin Kagerer[2].

It is published in the *Journal of Applied Statistics*, compare Haupt et al. (2011).

Link:

http://www.tandfonline.com/doi/full/10.1080/02664763.2011.573542#.U0fAcYbwqKA

[1]Department of Business Administration and Economics, Bielefeld University, hhaupt@wiwi.uni-bielefeld.de

[2]Department of Economics, University of Regensburg, kathrin.kagerer@wiwi.uni-regensburg.de

Bibliography

Ahmad, I. A. & Cerrito, P. B. (1994). Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference*, *41*, 349 – 364.

Aitchison, J. & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, *63*(3), 413 – 420.

Anglin, P. M. & Gencay, R. (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *11*, 633 – 648.

Bierens, H. J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, *78*(383), 699 – 707.

Bierens, H. J. (1987). Kernel estimators of regression functions. In F. Truman (Ed.), *Advances in Econometrics: Fifth World Congress* (pp. 99 – 144). Cambridge University Press.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, *13*, 68 – 84.

Cai, Z. & Li, Q. (2009). Some recent developments on nonparametric econometrics. *Advances in Econometrics*, *25*, 495 – 549.

Chakrabarty, M., Schmalenbach, A., & Racine, J. S. (2006). On the distributional effects of income in an aggregate consumption relation. *Canadian Journal of Economics*, *39*(4), 1221 – 1243.

Chave, A. D. & Thomson, D. J. (2003). A bounded influence regression estimator based on the statistics of the hat matrix. *Applied Statistics*, *52*(3), 307 – 322.

Davidson, R. & MacKinnon, J. G. (2004). *Econometric theory and methods*. Oxford University Press.

Delgado, M. A. & Gonzalez Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics*, *29*(5), 1469 – 1507.

Delgado, M. A. & Mora, J. (1995). Nonparametric and semiparametric estimation with discrete regressors. *Econometrica*, *63*(6), 1477 – 1484.

Delgado, M. A. & Robinson, P. M. (1992). Nonparametric and semiparametric methods for economic research. *Journal of Economic Surveys*, *6*(3), 201 – 249.

Delgado, M. A. & Stengos, T. (1994). Semiparametric specification testing of non-nested econometric models. *Review of Economic Studies*, *61*, 291 – 303.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–102.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall.

Fan, Y. & Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, *64*(4), 865 – 890.

Gencay, R. & Yang, X. (1996). A forecast comparison of residential housing prices by parametric versus semiparametric conditional mean estimators. *Economics Letters*, *52*, 129 – 135.

Gneiting, T. (2010). Making and evaluating point forecasts. `http://arxiv.org/PS_cache/arxiv/pdf/0912/0912.0902v2.pdf` (08/June/2010).

Gyimah-Brempong, K. & Racine, J. S. (2006). Alcohol availability and crime: a robust approach. *Applied Economics*, *38*, 1293 – 1307.

Gyimah-Brempong, K. & Racine, J. S. (2010). Aid and economic development: A robust approach. *Journal of International Trade and Economic Development*, *19*, 319 – 349.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer.

Hall, P., Li, Q., & Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics*, *89*(4), 784 – 789.

Hall, P., Racine, J. S., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, *99*(468), 1015–1026.

Hamermesh, D. S. & Biddle, J. E. (1994). Beauty and the labor market. *American Economic Review*, *84*, 1174 – 1194.

Haupt, H., Kagerer, K., & Schnurbus, J. (2011). Cross-validating fit and predictive accuracy of nonlinear quantile regressions. *Journal of Applied Statistics*, *38*, 2939–2954.

Haupt, H. & Petring, V. (2011). Assessing parametric misspecification and heterogeneity in growth regression. *Applied Economics Letters*, *18*, 389 – 394.

Haupt, H., Schnurbus, J., & Tschernig, R. (2010a). On nonparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *5*, 894 – 901.

Haupt, H., Schnurbus, J., & Tschernig, R. (2010b). Statistical validation of functional form in multiple regression using R. In H. D. Vinod (Ed.), *Advances in Social Science Research Using R*, volume 196 of *Lecture Notes in Statistics* chapter 9, (pp. 155–166). Springer.

Hayfield, T. & Racine, J. S. (2011). *np: Nonparametric kernel smoothing methods for mixed data types*. R package version 0.40-4.

Henderson, D. J. (2010). A test for multimodality of regression derivatives with application to nonparametric growth regressions. *Journal of Applied Econometrics*, *25*(3), 458 – 480.

Henderson, D. J. & Millimet, D. J. (2008). Is gravity linear? *Journal of Applied Econometrics*, *23*, 137 – 172.

Horowitz, J. L. & Lee, S. (2002). Semiparametric methods in applied econometrics: do the models fit the data? *Statistical Modelling*, *2*, 3 – 22.

Horowitz, J. L. & Spokoiny, V. G. (2001). An adaptive rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, *69*(3), 599 – 631.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

Härdle, W., Liang, H., & Gao, J. (2000). *Partially Linear Models*. Physica.

Härdle, W. & Linton, O. (1994). *Applied Nonparametric Methods*, volume IV of *Handbook of Econometrics*, chapter 38, (pp. 2295 – 2339). Elsevier Science.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. Springer.

Hsiao, C., Li, Q., & Racine, J. S. (2007). A consistent model specification test with mixed discrete and continuous data. *Journal of Econometrics*, *140*, 802 – 826.

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 271 – 293.

Kiefer, N. & Racine, J. S. (2009). The smooth colonel meets the reverend. *Journal of Nonparametric Statistics*, *21*(5), 521 – 533.

Koenker, R. (2010). Additive models for quantile regression: An analysis of risk factors for malnutrition in india. In H. D. Vinod (Ed.), *Advances in Social Science Research Using R*, volume 196 of *Lecture Notes in Statistics* chapter 2, (pp. 23–33). Springer.

Koenker, R. & Zeileis, A. (2009). On reproducible econometric research. *Journal of Applied Econometrics*, *24*, 833 – 847.

Lee, T.-H. (2000). Neural network test and nonparametric kernel test for neglected nonlinearity in regression models. *Studies in Nonlinear Dynamics & Econometrics*, *4*(4), 169 – 182.

Li, C., Ouyang, D., & Racine, J. S. (2009). Nonparametric regression with weakly dependent data: the discrete and continuous regressor case. *Journal of Nonparametric Statistics*, *21*(6), 697 – 711.

Li, Q. (1996a). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, *15*(3), 261 – 274.

Li, Q. (1996b). On the root-n-consistent semiparametric estimation of partially linear models. *Economics Letters*, *51*, 277 – 285.

Li, Q., Maasoumi, E., & Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics*, *148*(2), 186 – 200.

Li, Q. & Racine, J. S. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*, 266 – 292.

Li, Q. & Racine, J. S. (2004a). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*, 485 – 512.

Li, Q. & Racine, J. S. (2004b). Predictor relevance and extramarital affairs. *Journal of Applied Econometrics*, *19*, 533 – 535.

Li, Q. & Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. & Racine, J. S. (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *26*(4), 423–434.

Li, Q. & Racine, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory*, *26*, 1607 – 1637.

Li, Q., Racine, J. S., & Wooldridge, J. M. (2009). Efficient estimation of average treatment effects with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, *27*(2), 206 – 223.

Li, Q. & Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics*, *87*, 145 – 165.

Loader, C. R. (1999). Bandwidth selection: Classical or plug-in? *The Annals of Statistics*, *27*(2), 415 – 438.

Maasoumi, E. & Racine, J. S. (2009). A robust entropy-based test of asymmetry for discrete and continuous processes. *Econometric Reviews*, *28*(1-3), 246 – 261.

Maasoumi, E., Racine, J. S., & Stengos, T. (2007). Growth and convergence: A profile of distribution dynamics and mobility. *Journal of Econometrics*, *136*, 483 – 508.

Miles, D. & Mora, J. (2003). On the performance of nonparametric specification tests in regression models. *Computational Statistics & Data Analysis*, *42*, 477 – 490.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, *9*(1), 141 – 142.

Ouyang, D., Li, Q., & Racine, J. S. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, *18*(1), 69 – 100.

Ouyang, D., Li, Q., & Racine, J. S. (2009). Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory*, *25*, 1 – 42.

Pagan, A. & Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press.

Parmeter, C. F., Henderson, D. J., & Kumbhakar, S. C. (2007). Nonparametric estimation of a hedonic price function. *Journal of Applied Econometrics*, *22*, 695 – 699.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Racine, J. S. (1997). Consistent specification testing for nonparametric regression. *Journal of Business and Economic Statistics*, *15*(3), 369 – 378.

Racine, J. S. (2008). Nonparametric econometrics: A primer. *Foundations and Trends in Econometrics*, *3*(1), 1 – 88.

Racine, J. S., Hart, J., & Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, *25*(4), 523 – 544.

Racine, J. S. & Ker, A. (2006). Rating crop insurance policies with efficient nonparametric estimators that admit mixed data types. *Journal of Agricultural and Resource Economics*, *31*(1), 27 – 39.

Racine, J. S. & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*, 99 – 130.

Racine, J. S., Li, Q., & Zhu, X. (2004). Kernel estimation of multivariate conditional distributions. *Annals of Economics and Finance*, *5*, 211 – 235.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*(4), 931 – 954.

Rousseeuw, P. J. & van Zomeren, B. C. (1990). Unmasing multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*(411), 633 – 639.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on statistics and applied probability. Chapman and Hall.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677 – 680.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*, *25*(2), 613 – 641.

Vinod, H. D. (2008). *Hands-on intermediate econometrics using R*. World Scientific Publishing.

Wand, M. P. & Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall.

Wang, M.-C. & van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, *68*(1), 301 – 309.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, *26*(4), 359 – 372.

Whang, Y.-J. & Andrews, D. W. (1993). Tests of specification for parametric and semiparametric models. *Journal of Econometrics*, *57*, 277 – 318.

Wilson, P. W. & Carey, K. (2004). Nonparametric analysis of returns to scale in the US hospital industry. *Journal of Applied Econometrics*, *19*, 505 – 524.

Wolf, H. P. (2009). *relax: R Editor for Literate Analysis and lateX*. R package version 1.2.1.

Wolf, H. P. (2010). *aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, and some slider functions*. R package version 1.2.3.

Yang, L. & Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(4), 793 – 815.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, *75*, 263 – 289.

Zheng, J. X. (1998). A consistent nonparametric test of parametric regression models under conditional quantile restrictions. *Econometric Theory*, *14*, 123 – 138.