Model Choice in Structured Nonparametric Regression and Diffusion Models

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften
der Universität Mannheim

vorgelegt von

Berthold R. Haag

Juni 2006

Dekan:	Prof. Konrad Stahl, Ph.D.
Referent:	Prof. Dr. Enno Mammen
Korreferent:	Prof. Dr. Rainer Dahlhaus
Tag der mündlichen Prüfung:	26. Juli 2006

Acknowledgements

First of all, I have to thank my advisor Enno Mammen for his guidance and supervision of my thesis. He introduced me to the topic of nonparametric regression and my research has benefited a lot from his broad statistical knowledge. It facilitated the progress of the thesis to find his door always open to ask questions, to discuss and to receive intuitive illustrations.

I have to thank Stefan Hoderlein and Krishna Pendakur for accepting me as coauthor in their research project. It was a good start for the dissertation to have a concrete research question at the very beginning. Parts of our joint work build the third chapter of this thesis. The fruitful cooperation on nonparametric demand has led to a follow-up project with Stefan on which the second chapter is based.

I am grateful to Rainer Dahlhaus for agreeing immediately to write the report as second referee and I have to thank him in particular for writing it very fast.

The last two chapters have been improved a lot by numerous comments by Kyusang Yu. I benefited a lot from his experience with additive models. The discussions with Christian Conrad about nonparametric testing have often opened a new view to problems. Furthermore, we cooperated very successfully not only in research but also in dealing with the administrative and editorial work on our theses. I also have to thank Kristina Barth for struggling through my grammatical mistakes – apart from this page.

During the time at the Chair of Statistics I have enjoyed working with my colleagues Christian Conrad, Stefan Hoderlein, Melanie Schienle and Kyusang Yu. It was always pleasant and fruitful to discuss scientific questions or anything else.

I am specially grateful to my mother and Susanne for their love and support.

Mannheim, 23. August 2006

Berthold Haag

Contents

Chapter 1

Introduction

Nonparametric regression techniques have become a broad area of statistical research in various fields. The roots of the methodology date back at least to the middle of the last century (Rosenblatt, 1956; Parzen, 1962; Nadaraya, 1964; Watson, 1964). Because of the computational cost the applicability was limited until twenty years ago. The rapid development of computer technology and the availability of large datasets have abolished this restriction.

In a classical regression framework it is the aim to estimate the functional relation between a set of predictors and a response variable. In a parametric regression the unknown function is parameterized with a certain (finite) number of unknown parameters, which are then estimated based on a sample of observations. Such a globally restricted model is often not flexible enough to analyze data appropriately. In contrast, nonparametric techniques focus on the estimation of the functional relation at a single point and use the information provided by the data in a neighborhood.

Mathematically speaking, both methods try to minimize the distance between the observed realizations of the response variable and values that are predicted by a specific functional relation. If the class of functional relations is not restricted at all, this distance is minimized by any function that passes through all observed data points (if there are no multiple observations of the predictors). Such an estimator may provide very poor predictions at points different from the observed values. However, the quality of predictions is an important task for estimation. Therefore, the possible solutions of the minimization problem should be chosen from a smaller class of functions. This is done by imposing certain restrictions (e. g. a parametric structure) on the functions, which leads to structured regression models.

In nonparametric regression the restrictions are imposed by assuming a regular behavior of the unknown function in small neighborhoods. Using kernel smoothing, which is probably the most popular nonparametric method in theoretical and applied econometrics at the moment, the unknown function is assumed to be differentiable of a certain degree. While smoothing techniques provide very flexible estimators, their use is problematic when the dimension of the predictors increases. To have enough observations in a local (multivariate) neighborhood, the sample sizes have to increase drastically with the dimension. This phenomenon is known as the *curse of dimensionality* of nonparametric regression.

The curse of dimensionality can be circumvented by further restricting the class of functions – but keeping it still flexible enough. One popular way is to impose an additive structure on the unknown function. Thereby the behavior of the function is only restricted in one-dimensional neighborhoods. The quality of an estimation technique is usually analyzed by its asymptotic behavior, if the sample size increases. The curse of dimensionality is then displayed in slower rates of convergence of such estimators to the unknown functional relation. In contrast additive functions can be estimated with one-dimensional rates. In that sense imposing additivity can be seen as a dimension reduction technique.

It is important to investigate if the choice of a certain structured model is appropriate. A justification can arise from theoretical considerations about the functional relation of the observed data. Secondly, the performance of such models can be judged in a statistical sense by implementing testing procedures. This thesis contributes to the development of testing procedures for structured models (Chapters 2, 3 and 5) as well as to the use of additive dimension reduction for estimation and testing (Chapters 4 and 5). The four chapters are self-contained and can be read separately. Each chapter ends with an appendix in which all proofs are collected not to interrupt the outline of the presentation.

Chapter 2 is based on Haag and Hoderlein (2005). A kernel based test statistic is proposed to test for the omission of variables from a nonparametric regression. The applicability of the test is driven by the fact that under the null hypothesis (of a lower-dimensional model) the estimator converges at a faster rate. Since this also holds in (semi-)parametric models, the theoretical results can be extended to this class of models. The advantage of the test statistic is reflected in better bias properties than comparable tests proposed in the literature. A large number of econometric applications involve systems of equations, therefore the results are presented to allow for a multivariate dependent variable.

Because the asymptotic approximations are usually not valid in finite samples a bootstrap procedure is proposed for the class of tests mentioned above. Bootstrap versions of the test statistic do not require pre-estimation of complicated nonparametric objects. It is formally established that the procedure is valid. A simulation study is conducted to investigate the performance of the bootstrap in finite samples. In addition, the tests are extended to include the popular local polynomial estimators. Additionally, the case of dependent data is considered.

The test procedure is applied to testing homogeneity in consumer demand. In a simple model, demand depends on prices and income of different goods. Under the assumption of homogeneity the demand does only depend on relative prices (relative to the income) and therefore the dimension of the predictor variable decreases by one. Using British demand data, the hypothesis of homogeneity is not rejected in a simple model with three aggregated goods.

In the third chapter a structural model is considered that is implied by economic theory. An important rationality restriction in consumer demand is the symmetry of the Slutsky matrix. Not assuming a parametric structure of the demand function, this results in a nonlinear restriction involving the demand function and its derivatives. A test statistic and a bootstrap implementation are proposed and the asymptotic results are presented. This chapter consists of parts of Haag, Hoderlein and Pendakur (2005), where additionally a constraint estimator of the demand function, that imposes symmetry, is presented. An application to Canadian household data can be found there as well.

Chapter 4 is concerned with the nonparametric estimation of diffusion processes. Continuous-time models have been a basic tool in theoretical finance since the 1970s – mainly because they can be analyzed with elegant probabilistic techniques. In consequence the development of statistical methods for continuous-time processes has attracted much attention. Recently, more flexible nonparametric models and their estimation has been studied. Since nonparametric estimation in these models can be considered as a regression problem these methods suffer from the curse of dimensionality.

As pointed out above, additive models provide a powerful technique to overcome this problem and to maintain high flexibility. Estimation of such models requires iterative procedures and the asymptotic analysis is much more complex than in the classical setting. For the estimation of the additive components Mammen, Linton and Nielsen (1999) have introduced smooth backfitting estimation, an iterative procedure that uses a projection interpretation of usual kernel estimators. For the classical nonparametric regression model it has been shown that smooth backfitting based on local linear estimators is oracle efficient, i. e. it has the same bias and variance as the infeasible estimator based on the knowledge of all other components.

In Chapter 4 a multivariate diffusion process is considered and (some or all) elements of the drift vector and the diffusion matrix are modelled as additive functions. Smooth backfitting based on local linear and Nadaraya-Watson estimators is used to estimate the components. Assuming stationarity, the asymptotic properties of all estimators are derived under high frequency sampling. The efficiency results from the standard regression continue to hold. In particular, Nadaraya-Watson based estimators achieve the same variance as the oracle estimator, while the bias is not oracle. The local linear based estimators are shown to be fully oracle efficient. In a simulation study, the finite sample performance of the estimators is investigated. As an illustration, the estimators are applied to interest yield data.

The last chapter returns to the problem of testing for parametric structure. A standard approach is to measure the distance between a parametric and a nonparametric fit with a squared deviation measure. These tests inherit the curse of dimensionality from the nonparametric estimator. This results in a loss of power in finite samples and against local alternatives.

A new test statistic is proposed to circumvent the curse of dimensionality by projecting the residuals under the null hypothesis onto the space of additive functions. To estimate this projection the smooth backfitting estimator is used. The asymptotic behavior of the test statistic is derived under the null hypothesis and local and fixed alternatives. The motivation for the projection approach is to have a data analytic tool if the sample size is too small for a full-dimensional test as in Chapter 2. In that case, the asymptotic approximations are usually not valid and it is advisable to simulate the distributions with the bootstrap.

Therefore, a wild bootstrap procedure is proposed and its validity is established. The finite sample properties of the bootstrap are investigated in a simulation study. The test has good power in different settings and the circumvention of the curse of dimensionality is demonstrated in a high-dimensional model. It is very robust in particular against increasing correlation of the predictors. Finally the test is applied to testing the parametric specification of a consumer demand system.

Chapter 2

Bootstrap Specification Testing in Systems of Equations

2.1 Introduction

Nonparametric specification testing in systems of equations appears throughout Economics. For the proposed test statistics, there are two main areas of application: The first is testing for parametric or semiparametric specification, the second is testing for the significance of regressors. The main focus of this chapter is to test for significance of certain regressors. The application to test (semi-) parametric specifications will appear as a direct extension. Because there exist already a large literature on nonparametric testing, the approach of this chapter has to be integrated.

In the nonparametric testing literature, there are two main strands of work. The first are the integrated conditional moment (ICM) tests. Key contributions for parametric specifications are Bierens (1982, 1990) and Bierens and Ploberger (1997), while Delgado and Gonzales-Manteiga (2001) consider omission of variables. These tests can be viewed as extensions to the conditional moment tests proposed by Newey (1985) and Tauchen (1985).

The second strand of literature considers the L_2-distance between two functions, usually using nonparametric (kernel) estimators at some point. Within this class there are two subclasses that can be classified according to their treatment of the degenerate U-statistic which is at the core of the test statistic. The first subclass avoids dealing with the U-statistic explicitly by using ad hoc methods like reweighting observations or splitting the sample. Contributions include Hidalgo

(1992), Wooldridge (1992), Yatchew (1992) and Whang and Andrews (1993).

The second subclass of tests using nonparametric estimators deals directly with this complication, by applying central limit theorems of Hall (1984) or de Jong (1987). Individual tests include Härdle and Mammen (1993), Hong (1993), Horowitz and Härdle (1994), Fan and Li (1996), Lavergne and Vuong (1996, 2000), Zheng (1996), Li and Wang (1998) and Aït-Sahalia, Bickel and Stoker (2002). Related is also the work of Horowitz and Spokoiny (2001).

The ICM tests and the test that use nonparametric estimators are compared in Fan and Li (2000), who use the notation n, h and d to denote sample size, bandwidth and dimension of all regressors, respectively. The upshot of their discussion is that ICM-tests can detect Pitman type local alternatives that approach the null at order $n^{-1/2}$, whereas the second class can only detect those that approach the null at order $n^{1/2}h^{d/4}$. In contrast, the second type of tests has better power properties against high frequency alternatives. This suggests that the two types of tests should be seen as complements rather than competitors. However, generally speaking, ICM tests have a nonnormal limiting distribution that depends on nuisance parameters. Precisely this dependence makes their application rather cumbersome.

Within the class of L_2-distance tests, the approach that avoids ad hoc modifications may be seen as more natural. Sample splitting for instance is associated with an obvious loss of power (see Fan and Li ,1996, for further discussion on the disadvantages of ad hoc modification). Considering the omission of variables the procedure of this chapter is more closely related to Fan and Li (1996), Lavergne and Voung (2000) and Aït-Sahalia, Bickel and Stoker (2002), while the other tests mentioned above concentrate on the case of a parametric null hypothesis.

There are several extensions in comparison to Fan and Li (1996), Lavergne and Voung (2000) and Aït-Sahalia, Bickel and Stoker (2002). Arguably the biggest is the use of the bootstrap. This helps to avoid the pre-estimation of elements of the limiting distribution. In addition, the bootstrap has the advantage of generating better approximations to the unknown finite sample distribution. Specifically, we adopt a "wild bootstrap" procedure as proposed in Härdle and Mammen (1993), Gozalo (1997) and Li and Wang (1998) for testing parametric specifications. Because of an additional smoothing step in the construction of the test statistic, our specific test statistic is shown to have better bias properties than Fan and Li (1996), Lavergne and Voung (2000) and Aït-Sahalia, Bickel and Stoker (2002). Among other things, this results in weaker assumptions on the bandwidths.

In the next section we introduce the test formally and discuss the conventional asymptotic theory as well as a bootstrap version of the test statistic. The third section will focus on extensions to the basic test statistic of the second section: The implementation of local polynomials, the extension to (semi-)parametric hypotheses, and the case of dependent (i. e. mixing) data. A simulation study will occupy the fourth section. Finally, the method will be applied to testing homogeneity of degree zero in demand analysis, using British data.

2.2 The Test Statistic

2.2.1 Transforming the Hypothesis into a Test Statistic

Throughout this paper, we consider a model that captures the relationship between the random vectors Y, X and Z. Here $Y \in \mathbb{R}^{d_Y}$ is a d_Y-dimensional dependent variable, and $X \in \mathbb{R}^{d_X}, Z \in \mathbb{R}^{d_Z}$ are predictors. The hypothesis to be tested is whether Z can be omitted from the regression of Y on (X, Z). For testing this hypothesis, we define the following functions

$$\begin{aligned} \mu(x,z) &= \mathbf{E}(Y \mid X = x, Z = z) \\ m(x) &= \mathbf{E}(Y \mid X = x). \end{aligned}$$

If it is possible to exclude Z from the regression, then these functions will coincide almost surely. Hence, we will base the test statistic on the null hypothesis

$$H_0 \colon \mathbb{P}(\mu(X,Z) = m(X)) = 1,$$

while the alternative is that they differ on a subset of the support of Z of positive measure. The null is equivalent to the condition that the L_2-distance of the two functions is zero. Using a positive and bounded weighting function $a(x, z)$ this condition can be expressed as

$$\Gamma = \mathbf{E}\Big(\sum_{j=1}^{d_Y}\big(\mu^j(X,Z) - m^j(X)\big)^2 a(X,Z)\Big) = 0. \tag{2.1}$$

Using the fact that $m^j(X) = \mathbf{E}(m^j(X) \mid X, Z)$, we base the test on

$$\Gamma = \mathbf{E}\Big(\sum_{j=1}^{d_Y}\big(\mu^j(X,Z) - \mathbf{E}(m^j(X) \mid X,Z)\big)^2 a(X,Z)\Big). \tag{2.2}$$

As mentioned above, alternative test statistics for the single equation case ($d_Y = 1$) have been proposed in several publications. Aït-Sahalia, Bickel and Stoker (2002) base their test statistic directly on equation (2.1), while Fan and Li (1996) propose to base a test statistic on

$$\mathbf{E}\big((Y - m(X))\,\mathbf{E}(Y - m(X) \mid X, Z) f(X, Z) a(X, Z)\big).$$

To avoid technical problems, Fan and Li (1996) use $a(X, Z) = f^2(X, Z) a'(X, Z)$ and a leave-one-out estimator for the conditional expectation. Another possibility would be to compare residual sums of squares, i. e. basing a test statistic on

$$\mathbf{E}\big(((Y - m(X))^2 - (Y - \mu(X, Z))^2) a(X, Z)\big)$$

which would be an adaptation of the tests by Dette (1999) and Fan, Zhang and Zhang (2001) to the problem of omitting variables. To our knowledge such a test has not yet been implemented. We expect that its local power properties are worse than those of a test based on (2.1) or (2.2) (see Dette, 1999, who shows these worse power properties for the case of a parametric null hypothesis).

2.2.2 Sample Counterpart

The sample counterpart of Γ in (2.2) serves as test statistic. Given a sample of n independent and identically distributed random vectors $(Y_1, X_1, Z_1), \dots,$ (Y_n, X_n, Z_n), we replace the unknown functions $m(x)$ and $\mu(x, z)$ by their Nadaraya-Watson estimators $\widehat{m}_{\widetilde{h}}(x)$ and $\widehat{\mu}_h(x, z)$. Formally, these are defined as vectors with the one-dimensional estimators, $\widehat{m}^j_{\widetilde{h}}(x) = \sum_{i=1}^n K_{\widetilde{h}}(x - X_i) Y_i^j / \sum_{i=1}^n K_{\widetilde{h}}(x - X_i)$ and $\widehat{\mu}^j_h(x, z) = \sum_{i=1}^n K_h(x - X_i, z - Z_i) Y_i^j / \sum_{i=1}^n K_h(x - X_i, z - Z_i)$, where $K_h(u) = K(u/h)/h$ with a kernel K and bandwidths h and $\widetilde{h}$. as individual elements. As an estimator for $\mathbf{E}(m^j(X) \mid X = x, Z = z)$ we propose

$$\widehat{\mathcal{K}_n m^j_{\widetilde{h}}}(x, z) = \frac{\sum_{i=1}^n K_h(x - X_i, z - Z_i) \widehat{m}^j_{\widetilde{h}}(X_i)}{\sum_{i=1}^n K_h(x - X_i, z - Z_i)}.$$

Then, the statistic is given by

$$\widehat{\Gamma}_{\mathcal{K}} = \frac{1}{n} \sum_{j=1}^{d_Y} \sum_{i=1}^{n} \big(\widehat{\mu}^j_h(X_i, Z_i) - \widehat{\mathcal{K}_n m^j_{\widetilde{h}}}(X_i, Z_i)\big)^2 A_i \tag{2.3}$$

with $A_i = a(X_i, Z_i)$. The additional smoothing step associated with $\widehat{\mathcal{K}_n m_{\widetilde{h}}}(x, z)$ produces an artificial bias that eliminates the bias coming from $\widehat{\mu}_h(x, z)$, thereby reducing the number of bias components in the asymptotic expression. This reduction in turn allows to employ less restrictive requirements on the bandwidths. A similar modification was suggested by Härdle and Mammen (1993) for the case of a parametric null hypothesis. The superiority of $\widehat{\Gamma}_{\mathcal{K}}$ over the tests of Aït-Sahalia, Bickel and Stoker (2002) and Fan and Li (1996) can be stated in terms of smoothness conditions and local power properties of the tests and will be discussed after Theorem 2.2.

2.2.3 Asymptotic Distribution of the Test Statistic

In order to treat the asymptotic distribution of the test statistic, we introduce the following assumptions. The first two assumptions are concerned with the data generating process.

Assumption 2.1. *The data $(Y_i, X_i, Z_i), i = 1, \ldots, n$ are independent and identically distributed with density $f(y, x, z)$.*

Assumption 2.2. *For the data generating process*

1. *The continuously differentiable weighting function $a(x, z)$ is positive and bounded with compact support $\mathcal{A} \subset \mathbb{R}^{d_X + d_Z}$.*

2. *$f(y, x, z)$ is r-times continuously differentiable ($r \geq 2$). f and its partial derivatives are bounded and square-integrable on $\mathcal{A}$.*

3. *$\mu(x, z)$ and $m(x)$ are $r + 1$-times continuously differentiable.*

4. *$f(x, z) = \int f(y, x, z)\,\mathrm{d}y$ is bounded from below on $\mathcal{A}$, i. e. $\inf_{(x,z)\in\mathcal{A}} f(x, z) = b > 0$.*

5. *The covariance matrix*

$$\Sigma(x, z) = (\sigma^{ij}(x, z))_{1 \leq i,j \leq d_Y} = \mathbf{E}((Y - \mu(X, Z))(Y - \mu(X, Z))' \mid X = x, Z = z)$$

   is square-integrable (elementwise) on $\mathcal{A}$.

6. *$\mathbf{E}((Y^j - \mu^j(X, Z))^2 (Y^k - \mu^k(X, Z))^2) < \infty$ for every $1 \leq j, k \leq d_Y$.*

The first assumption may be relaxed to allow for dependent data. We will discuss this extension in Section 2.3.3. Assumption 2.2 contains standard differentiability and integrability assumptions that do not have to be discussed.

The following assumptions are concerned with the kernel and the bandwidth sequences. For simplicity, we assume product kernels in both regressions. Therefore we formulate our assumptions for one-dimensional kernel functions. To further simplify things, we assume that we have only one single bandwidth for each regression (h and $\widetilde{h}$) instead of bandwidth vectors $\mathbf{h} \in \mathbb{R}^{d_X+d_Z}$ and $\widetilde{\mathbf{h}} \in \mathbb{R}^{d_X}$.

We shall make use of the following notation: Define kernel constants

$$\kappa_k = \int u^k K(u)\,\mathrm{d}u \qquad \text{and} \qquad \kappa_k^2 = \int u^k K(u)^2\,\mathrm{d}u$$

$$\kappa_* = \int \Big(\int K(u)K(u-v)\,\mathrm{d}u \Big)^2 \mathrm{d}v.$$

Then, our assumptions regarding kernels and bandwidths are as follows:

Assumption 2.3. *The one-dimensional kernel is Lipschitz continuous, bounded, has compact support, is symmetric around 0 and of order r (i. e. $\int u^k K(u)\,\mathrm{d}xu = 0$ for all $k < r$ and $\int u^r K(u)\,\mathrm{d}u < \infty$).*

Assumption 2.4. *For the bandwidths*

1. *For $n \to \infty$, the bandwidth sequence $h = O(n^{-1/\delta})$ satisfies*

$$d_X + d_Z < \delta. \tag{2.4}$$

2. *For $n \to \infty$, the bandwidth sequence $\widetilde{h} = O(n^{-1/\widetilde{\delta}})$ satisfies*

$$2\delta \frac{d_X}{d_X + d_Z} < \widetilde{\delta}. \tag{2.5}$$

3. *For the order r of the kernel holds*

$$\widetilde{\delta} \frac{2\delta - d_X - d_Z}{4\delta} < r. \tag{2.6}$$

While the assumptions on the kernel are standard, the assumptions on the bandwidths do merit some discussion. Observe first that the optimal rate for estimating the full dimensional regression function $\mu(x, z)$, given by

$$\delta_{opt} = (d_X + d_Z) + 2r.$$

is not excluded from inequality (2.4). Under the null hypothesis, $\mu(x,z)$ does not depend on z. Then, the derivatives with respect to z are zero and the corresponding bias terms disappear. It follows that under H_0 the optimal bandwidth in the z-directions is infinite. But under the alternative and in the x-directions there exists an optimal bandwidth.

If we want to make use of this bandwidth, through employing data-driven methods of bandwidth choice in the full dimensional regression (e. g. cross validation), then the inequalities (2.5) and (2.6) impose restrictions on the bandwidth $\widetilde{h}$ of the dimension-reduced regression function $m(x)$. More specifically, because of (2.5), it might be necessary to use a larger-than-optimal bandwidth, and because of (2.6), to employ higher order kernels. As an example, take $d_X = 1, d_Z = 1$. It is not possible to use both δ_{opt} and $\widetilde{\delta}_{opt}$ for any choice of r, because inequality (2.5) yields the restriction $\delta < \widetilde{\delta}$.

An alternative representation of (2.4)–(2.6) may be given in terms of n and h. We obtain $nh^{d_X+d_Z} \to \infty$ (necessary for consistency of the kernel density estimator), $h^{d_X+d_Z}\widetilde{h}^{-d_X} \to 0$ and $nh^{(d_X+d_Z)/2}\widetilde{h}^{2r} \to 0$.[1] The last two conditions ensure that the estimation error of the dimension-reduced regression does not dominate the test statistic.

The restrictions on the bandwidths are much weaker than those restrictions assumed by Aït-Sahalia, Bickel and Stoker (2002). In their case the optimal rate for estimation is excluded for all regressions and higher order kernels are always needed, provided $d_X + d_Z \geq 3$. In contrast, our assumptions allow to trade the use of higher order kernel for a larger-than-optimal bandwidth.

In practise we propose to calculate data-driven bandwidths (by cross-validation) for the dimension reduced regression. In case the optimal rate is excluded, we suggest to adjust the bandwidth by $n^{1/\widetilde{\delta}_{opt}-1/\widetilde{\delta}}$. Although we do not formally address the issue of data-driven bandwidths $\widehat{h}$ we assume that our results will hold if $\widehat{h}/h \xrightarrow{P} 1$.

For the first theorem, we introduce the following quantities

$$\sigma_\Gamma^{ij} = \iint \sigma^{ij}(x,z)^2 a(x,z)^2 \,\mathrm{d}x\,\mathrm{d}z \qquad b_\Gamma^i = \iint \sigma^{ii}(x,z)a(x,z)\,\mathrm{d}x\,\mathrm{d}z.$$

The asymptotic normality of the test statistic is given by the following

[1] Note that these restrictions imply $n\widetilde{h}^{d_X} \to \infty$, which ensures the consistency of the dimension reduced regression.

Theorem 2.1. *Let Assumptions 2.1–2.4 hold. Then we have that under H_0*

$$\Sigma_{\mathcal{K}}^{-1}(nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}} - h^{-(d_X+d_Z)/2}B_{\mathcal{K}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

where

$$\Sigma_{\mathcal{K}}^2 = 2(\kappa_*)^{d_X+d_Z}\Big(\sum_{i=1}^{d_Y}\sigma_\Gamma^{ii} + 2\sum_{i<j}\sigma_\Gamma^{ij}\Big) \qquad B_{\mathcal{K}} = (\kappa_0^2)^{d_X+d_Z}\sum_{i=1}^{d_Y} b_\Gamma^i.$$

Simplifying the proofs in the appendix to one line, the test statistic can be written as

$$\widehat{\Gamma}_{\mathcal{K}} = \Gamma + I_n + U_n,$$

where $\Gamma = 0$ under H_0, U_n depends upon the uniform rate of convergence of the restricted estimator, and I_n is a degenerated U-statistic which dominates asymptotically. This U-statistic converges at the rate $nh^{(d_X+d_Z)/2}$, which is faster than $n^{1/2}$, under the admissible bandwidth sequence.

Next, we investigate the behavior of the test statistic under the alternative. There are a number of efficiency measures (e. g. Bahadur efficiency or Hodges-Lehman efficiency) to compare two test statistics. The most common one is the asymptotic relative efficiency (Pitman efficiency) which compares the behavior of the tests under local alternatives. To this end, define a sequence of alternatives

$$H_{1n}\colon \mu(x,z) = m(x) + \varepsilon_n(x,z)$$

where $\varepsilon_n(x,z)$ is a converging sequence of functions. Note that fixed alternatives are included for $\varepsilon_n(x,z) = \varepsilon(x,z) \neq 0$.

Theorem 2.2. *Let Assumptions 2.1–2.4 hold. If there exists a constant B_L such that*

$$\lambda_n \sum_{k=1}^{d_Y}\frac{1}{n}\sum_{j=1}^{n}\Big(\frac{\varepsilon_n^k(X_j,Z_j)}{f(X_j,Z_j)}\Big)^2 a(X_j,Z_j) \xrightarrow{P} B_L$$

for $\lambda_n = O(nh^{(d_X+d_Z)/2})$. Then we have that under H_{1n}

$$\Sigma_{\mathcal{K}}^{-1}(nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}} - h^{-(d_X+d_Z)/2}(B_{\mathcal{K}} + \kappa_0^2 B_L)) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

For a fixed alternative it holds that $nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}} \to \infty$.

The test cannot detect alternatives that converge to zero at a rate faster than $n^{-1/2}h^{-(d_X+d_Z)/4}$. This means that the test suffers from the curse of dimensionality because the rate decreases as the number of dimensions increase. Aït-Sahalia, Bickel and Stoker (2002) and Fan and Li (1996) establish local power properties of their tests and both obtain the same rate. Theorem 2.2 holds for the test of Aït-Sahalia, Bickel and Stoker (2002) in an analogous fashion. A comparison with the test of Fan and Li (1996) is only possible using $a(x,z)f(x,z)^{-2}$ as a weighting function, since Fan and Li (1996) use density weighting. The asymptotic variance differs through a kernel related constant. Because $\kappa_* < \kappa_0^2$ for a density K, our test is asymptotically relatively more efficient than the test of Fan and Li (1996).

2.2.4 Bootstrap-Implementation

The direct way to implement the test is to estimate the expected value $B_{\mathcal{K}}$ and the variance $\Sigma^2_{\mathcal{K}}$. This requires the estimation of integrals like

$$\int \sigma^{jj'}(x,z)^k a(x,z)^k \,\mathrm{d}x\,\mathrm{d}z \qquad k=1,2,\ j,j'=1,\ldots,d_Y. \tag{2.7}$$

Therefore estimators of the conditional (co)variances are needed. A Nadaraya-Watson-type estimator may be defined as

$$\widehat{\sigma}_h^{jj'}(x,z) = \frac{\sum_{i=1}^n K_h(x-X_i,z-Z_i)\big(Y_i^j - \widehat{\mu}_h^j(X_i,Z_i)\big)\big(Y_i^{j'} - \widehat{\mu}_h^{j'}(X_i,Z_i)\big)}{\sum_{i=1}^n K_h(x-X_i,z-Z_i)}.$$

This estimator has better properties than the difference between estimators of the second and the squared first conditional moment of Y given X and Z (see Fan and Yao, 1998). Now the integral in (2.7) can be calculated numerically. To ensure consistency of the standardized test statistic the underlying (co)variance estimators (as well as the density estimator) have to be chosen such that

$$\sup_{(x,z)\in\mathcal{A}} |\widehat{\sigma}_h^{jj'}(x,z) - \sigma^{jj'}(x,z)| = o_P(h^{-(d_X+d_Z)/2}).$$

Estimating the components of the asymptotic distribution of $\widehat{\Gamma}_{\mathcal{K}}$ is cumbersome. Moreover, it is also problematic: In the proof of the asymptotic normality of the test statistic many terms of lower magnitude are omitted. Asymptotic approximations involving U-statistics work often very poorly in a finite sample, as was pointed out by Hjellvik and Tjøstheim (1995). To avoid this problem we propose a wild bootstrap procedure to derive critical values for the test statistic, as in Härdle and Mammen (1993). In our setting this is performed in the following way

1. Calculate (multivariate) residuals $\widehat{\varepsilon}_i = Y_i - \widehat{m}_{\widetilde{h}}(X_i)$.

2. For each i randomly draw $\varepsilon_i^* = (\varepsilon_i^{1,*}, \dots, \varepsilon_i^{d_Y,*})'$ from a distribution $\widehat{F}_i$ that mimics the first three moments of $\widehat{\varepsilon}_i$.

3. Generate the bootstrap sample $(Y_i^*, X_i^*, Z_i^*), i = 1, \dots, n$ by $Y_i^* = \widehat{m}_{\widetilde{h}}(X_i) + \varepsilon_i^*$ and $X_i^* = X_i, Z_i^* = Z_i$.

4. Calculate $\widehat{\Gamma}_{\mathcal{K}}^*$ from the bootstrap sample $(Y_i^*, X_i^*, Z_i^*), i = 1, \dots, n$.

5. Repeat steps 2 to 4 often enough to obtain critical values for $\widehat{\Gamma}_{\mathcal{K}}$.

Assumption 2.5. *For the bootstrap distribution*
The bootstrap residuals $\varepsilon_i^*, i = 1, \dots, n$ are drawn independently from distributions $\widehat{F}_i$, such that $\mathbf{E}_{\widehat{F}_i}\, \varepsilon_i^* = 0, \mathbf{E}_{\widehat{F}_i}\, \varepsilon_i^*(\varepsilon_i^*)' = \widehat{\varepsilon}_i \widehat{\varepsilon}_i'$ and $\mathbf{E}_{\widehat{F}_i}(\varepsilon_i^{k,*})^4 < \infty$ for all $k = 1, \dots, d_Y$.

This set of admissible distributions is very general. Apart from the simple wild bootstrap, a smooth conditional moment bootstrap as in Gozalo (1997) may also be used. In the classical wild bootstrap, residuals are drawn from a two-point distribution that takes the value $\widehat{\varepsilon}_i(1-\sqrt{5})/2$ with probability $(5+\sqrt{5})/10$ and $\widehat{\varepsilon}_i(1+\sqrt{5})/2$ else (see Härdle and Mammen, 1993). Assumption 2.5 is fulfilled for discrete distributions, distributions with compact support and – among others – for the normal distribution. These are the most commonly used distributions in practice.

The theoretical result concerning this bootstrap procedure is given in

Theorem 2.3. *Let Assumptions 2.1–2.5 be true. Under H_0, it holds that*

$$\Sigma_{\mathcal{K}}^{-1}(nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}}^* - h^{-(d_X+d_Z)/2}B_{\mathcal{K}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

conditional on the data $(Y_1, X_1, Z_1), \dots, (Y_n, X_n, Z_n)$ with probability tending to one.

To prove theorem 2.3 it is sufficient to assume that the bootstrap distribution $\widehat{F}_i$ mimics the first two moments of $\widehat{\varepsilon}_i$. Using an Edgeworth expansion in the proof, we conjecture that matching the first three moments yields a higher order approximation. In our simulation study we find evidence that this improves the finite sample properties. Therefore we recommend to mimic three moments in applications.

2.3 Extensions

In this section we discuss extensions to the test statistic along three lines. First, we explore the use of local polynomial estimators to replace the Nadaraya-Watson estimator. Second, we extend the test statistic to semi-parametric hypotheses. Last, but not least, we investigate the behavior of the test in the case of dependent data. In all cases, we focus on the respective modifications of Theorem 2.1. Changes in the proofs of the bootstrap result and the local power properties are straightforward.

2.3.1 Local Polynomials

In nonparametric regression analysis the superiority of local polynomial estimators to Nadaraya-Watson estimators is well known (see Fan and Gijbels, 1996). Therefore it is a natural extension to use local polynomial estimators for $\mu(x, z)$ and $m(x)$ in the test statistic. Recall that they are defined via minimizing

$$\sum_{i=1}^{n} \Big(Y_i^j - \sum_{0 \le |\mathbf{k}| \le p} b_{\mathbf{k}}(x, z)(X_i - x, Z_i - z)^{\mathbf{k}} \Big)^2 K_h(X_i - x, Z_i - z), \tag{2.8}$$

with respect to all $b_{\mathbf{k}}$. For vectors $\mathbf{k} = (k_1, \dots, k_{d_X + d_Y})$ we have utilized the notation $|\mathbf{k}| = \sum_j k_j$ and $x^{\mathbf{k}} = \prod_j (x^j)^{k_j}$. Then $\widehat{\mu}_h^{j,LP}(x, z)$ is defined as the solution for $b_{\mathbf{0}}$. Introducing the quantities

$$\widehat{t}_{\mathbf{k}}^j(x, z) = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \Big(\frac{(X_i - x, Z_i - z)}{h} \Big)^{\mathbf{k}} K_h(X_i - x, Z_i - z),$$

$$\widehat{f}_{h,\mathbf{k}}(x, z) = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{(X_i - x, Z_i - z)}{h} \Big)^{\mathbf{k}} K_h(X_i - x, Z_i - z),$$

which are arranged in a vector $\widehat{T}^j(x, z) = (\widehat{t}_{\mathbf{k}}^j(x, z))_{\mathbf{k}}$ and a matrix $\widehat{S}(x, z) = (\widehat{f}_{h,\mathbf{k}+\mathbf{j}}(x, z))_{\mathbf{k},\mathbf{j}}$ in a lexicographical order.[2] With this notation, the estimator can be written explicitly as

$$\widehat{\mu}_h^{j,LP}(x, z) = \lfloor \widehat{S}^{-1}(x, z) \widehat{T}^j(x, z) \rfloor_1,$$

where $\lfloor \cdot \rfloor_1$ extracts the first element of a vector. $\widehat{m}_h^{j,LP}(x)$ is defined analogously. The local polynomial version of $\mathbf{E}(m^j(X) \mid X, Z)$ is defined as the solution to (2.8)

[2] Addition is in the Hadamard-sense, i. e. $\mathbf{j} + \mathbf{k} = (j_1 + k_1, \dots, j_{d_X + d_Z} + k_{d_X + d_Z})$.

where Y_i^j is replaced with $\widehat{m}_{\widetilde{h}}^{j,LP}(X_i)$. Explicitly it can be written as

$$\widehat{\mathcal{K}_n m_{\widetilde{h}}}^{j,LP}(x,z) = \lfloor \widehat{S}^{-1}(x,z)\widetilde{T}^j(x,z)\rfloor_1,$$

where the elements of the vector $\widetilde{T}^j(x,z)$ are given by

$$\widetilde{t}_{\mathbf{k}}^j(x,z) = \frac{1}{n}\sum_{i=1}^n \widehat{m}_{\widetilde{h}}^{j,LP}(X_i)\Big(\frac{(X_i - x, Z_i - z)}{h}\Big)^{\mathbf{k}} K_h(X_i - x, Z_i - z).$$

The new test statistic is then the analog to (2.3)

$$\widehat{\Gamma}_{\mathcal{K}}^{LP} = \frac{1}{n}\sum_{j=1}^{d_Y}\sum_{i=1}^n \big(\widehat{\mu}_h^{j,LP}(X_i,Z_i) - \widehat{\mathcal{K}_n m_{\widetilde{h}}}^{j,LP}(X_i,Z_i)\big)^2 A_i.$$

To define the kernel constants arising in the bias and variance parts of the asymptotic distribution, we have to define the matrix $M = (\kappa_{\mathbf{j}+\mathbf{k}})_{\mathbf{j},\mathbf{k}}$ with entries $\kappa_{\mathbf{k}} = \int u^{\mathbf{k}} K(u)\,\mathrm{d}u$. In an abuse of notation we denote with $\kappa_{\mathbf{k}}^{-1}$ the elements of the first row of M^{-1}. This enables to define

$$\kappa_\Sigma = \int\Big(\int\Big(\sum_{1\le\mathbf{k}\le r}(u-v)^{\mathbf{k}}\kappa_{\mathbf{j}}^{-1}K(u-v)\Big)\Big(\sum_{1\le\mathbf{k}\le r}u^{\mathbf{k}}\kappa_{\mathbf{j}}^{-1}K(u-v)\Big)\,\mathrm{d}u\Big)^2\,\mathrm{d}v,$$

$$\kappa_B = \int\Big(\sum_{1\le\mathbf{k}\le r}u^{\mathbf{k}}\kappa_{\mathbf{j}}^{-1}K(u)\Big)^2\,\mathrm{d}u,$$

which we require for the derivation of the asymptotic distribution of $\widehat{\Gamma}_{\mathcal{K}}^{LP}$ in the following theorem:

Theorem 2.4. *Let Assumptions 2.1–2.3 hold. Let Assumption 2.4 hold for $r = p+1$ for p odd and $r = p+2$ for p even, where p is the order of the local polynomial estimator. Then we have that under H_0*

$$\Sigma_L^{-1}(nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}}^{LP} - h^{-(d_X+d_Z)/2}B_L) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\Sigma_L^2 = 2(\kappa_\Sigma)^{d_X+d_Z}\Big(\sum_{i=1}^{d_Y}\sigma_\Gamma^{ii} + 2\sum_{i<j}\sigma_\Gamma^{ij}\Big) \qquad B_L = (\kappa_B)^{d_X+d_Z}\sum_{i=1}^{d_Y} b_\Gamma^i.$$

Note, that if an even order of the local polynomial fulfills the requirements of Assumption 2.4, then the subsequent odd order polynomial fulfills also these requirements. The use of one additional order gives therefore no gain in flexibility when choosing the bandwidth sequences. Therefore, in contrast to estimation it is natural to use an even order local polynomial for testing. If we replace the corresponding kernel constants with κ_Σ and κ_B the results of Theorems 2.2 and 2.3 continue to hold. This can be seen directly from the proof of Theorem 2.4.

2.3.2 Semiparametric Modelling

The asymptotic distribution of the test is driven by the fact that the low-dimensional estimator $\widehat{m}_{\tilde{h}}(x)$ converges faster than the full-dimensional estimator $\widehat{\mu}_h(x,z)$. This remains true for semiparametric hypotheses, i. e.

$$H_{0S}\colon \mathbb{P}(\mu(x,z) = m(x) + G(z,\theta)) = 1,$$

where $G(z,\theta) = \big(G^1(z,\theta) + \cdots + G^{d_Y}(z,\theta)\big)$ is a known function depending on a finite-dimensional parameter vector $\theta \in \Theta$. Denote with $\widehat{\theta}$ a parametric estimator that allows us to construct estimators of the nonparametric regression part under H_{0S}, i. e.,

$$\widehat{m}^k_{\tilde{h}}(x,\widehat{\theta}) = \frac{\sum_{i=1}^n K_{\tilde{h}}(x-X_i)(Y_i^k - G^k(Z_i,\widehat{\theta}))}{\sum_{i=1}^n K_{\tilde{h}}(x-X_i)}.$$

Then we propose to use as test statistic

$$\widehat{\Gamma}^S_{\mathcal{K}} = \frac{1}{n}\sum_{j=1}^{d_Y}\sum_{i=1}^{n}\big(\widehat{\mu}_h(X_i,Z_i) - \widehat{\mathcal{K}_n m_{\tilde{h}}}^{j,S}(X_i,Z_i)\big)^2 A_i,$$

with

$$\widehat{\mathcal{K}_n m_{\tilde{h}}}^{j,S}(x,z) = \frac{\sum_{i=1}^n K_h(x-X_i,z-Z_i)(\widehat{m}^k_{\tilde{h}}(x,\widehat{\theta}) + G^k(Z_i,\widehat{\theta}))}{\sum_{i=1}^n K_h(x-X_i,z-Z_i)}.$$

To obtain an asymptotic result we require the following assumption on the speed of convergence of the semiparametric estimator:

Assumption 2.6. $G^k(z,\theta) - G^k(z,\widehat{\theta}) = o_P(n^{-1/2}h^{(d_X+d_Z)/4})$ *for all* $k = 1,\dots,d_Y$ *uniformly in* $\mathcal{A}_Z = \{z \mid \exists\ x\ s.t.\ (x,z) \in \mathcal{A}\}$ *and* $\theta \in \Theta$.

This assumption is stated in a very general fashion. It has to be checked for a specific model and estimation problem. As an example, consider the linear model with $d_X = 0$ and $G(z,\theta) = \theta' z$. The least squares estimator is known to be root-n consistent and Assumption 2.6 is fulfilled. Moreover, as a special case for $d_Y = 1$ we obtain the test introduced by Härdle and Mammen (1993).

The asymptotic distribution of the test is stated in the following

Theorem 2.5. *Let Assumptions 2.1–2.4 and 2.6 hold. Then we have that under* H_{0S}

$$\Sigma_{\mathcal{K}}^{-1}(nh^{(d_X+d_Z)/2}\widehat{\Gamma}^S_{\mathcal{K}} - h^{-(d_X+d_Z)/2}B_{\mathcal{K}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\Sigma_{\mathcal{K}}$ *and* $B_{\mathcal{K}}$ *is given as in Theorem 2.1.*

2.3.3 Dependent Data

The assumption of independent and identically distributed data is very restrictive. In many data sets in practice time series effects are present. To deal with this complication, we extend the results of the previous sections to the case of mixing random variables. For a time series $W_i = (Y_i, X_i, Z_i), i = 1, \dots, n$ we define the sigma algebras $\mathcal{F}_s^t = \sigma(W_s, W_{s+1}, \dots, W_t$ with $-\infty \leq s < t \leq \infty$ and the β-mixing coefficients

$$\beta(n) = \sup_{t \in \mathbb{Z}} \mathbf{E}\big(\sup_{A \in \mathcal{F}_{t+n}^{\infty}} |\mathbb{P}(A \mid \mathcal{F}_{-\infty}^t) - \mathbb{P}(A)| \big)$$

A process is called absolutely regular if $\beta(n) \to 0$ for $n \to \infty$. To derive the asymptotic normality of the test statistic, we invoke the following additional assumptions

Assumption 2.7. *For dependent data*

1. *The data $W_i = (Y_i, X_i, Z_i), i = 1, \dots, n$ are strictly stationary and absolutely regular with mixing coefficients $\beta(n)$. The stationary density is denoted by $f(w)$.*

2. *The density of the joint distribution of (W_q, W_r, W_s, W_t) is bounded and continuously differentiable for all q, r, s, t.*

3. *For some $\nu > 1$ it holds that $\mathbf{E}\,|Y^j|^{4\nu} < \infty$ for all $j = 1, \dots, d_Y$.*

4. *For the mixing coefficients we have the summability conditions*

$$\sum_{i=1}^{\infty} \beta(i)^{1-2/\nu} < \infty \quad \text{and} \quad \sum_{i=1}^{\infty} i^{a'} \beta(i)^{1-2/a},$$

with $2 < a < 4\nu$ and $a' > 1 - 2/a$.

It holds that $\sum_{n=1}^{\infty} \psi(n) < \infty$ where

$$\psi(n) = \frac{nL(n)}{r(n)} \Big(\frac{nT(n)^2}{\widetilde{h}^{d_X} \log n} \Big)^{1/4} \beta(r(n)),$$

with $L(n) = (nT(n)^2 / (\widetilde{h}^{d_X + 2} \log n))^{d_X/2}$, $r(n) = (n\widetilde{h}^{d_X} / \log n)^{1/2} / T(n)$ and $T(n) = \big(n \log n (\log \log n)^{1+\epsilon}\big)^{1/4\nu}$.

For $m = n^{1/\delta_m}$ with $\delta_m > 4\delta$ and $1/\delta + 1/\delta_m < 3/2$ it holds that

$$n^6 h^2 (m^2 \beta(m)^{1-1/\nu} + n^2 \beta(m)^{2-2/\nu}) \to 0,$$

as $n \to \infty$.

These assumptions are not restrictive: Many well-known time series models were shown to be absolutely regular, most of them with exponentially decaying mixing coefficients. For mixing coefficients with geometric decay, the requirements of Assumption 2.7 are directly fulfilled (for some $\nu > 1$).

The dependence structure of Y, X and Z is only modelled in terms of differentiability assumptions on their joint density. This is general enough to cover the cases where X and Z are lagged values of Y. Beside time series regression, the test can be used to determine the order of a nonparametric AR-process as well as to test for parametric AR-structure.

This assumption enables us to state the following extension to the previous theorems.

Theorem 2.6. *Theorems 2.1–2.4 remain valid, if we replace Assumption 2.1 by Assumption 2.7.*

Asymptotic results under mixing assumptions are obtained by a trade-off between the number of existing moments and the decaying rate of the mixing coefficients. This is given in terms of the parameter ν. The use of a larger bandwidth may also reduce the requirements on the rate of decay (and the moment conditions). Here, this is given in terms of the sequence $\psi(n)$.

2.4 Monte Carlo Simulation Study

In this section we examine the finite sample behavior of the test statistic $\widehat{\Gamma}_{\mathcal{K}}$ by means of a simulation study. Under the null hypothesis, we simulate from the model

$$Y_i = m(X_i) + \sigma(X_i) U_i, \qquad i = 1, \ldots, n,$$

where $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $X_i \overset{\text{iid}}{\sim} \mathcal{U}(-\pi, \pi)$ independent of U_i. The additional regressor $Z_i \overset{\text{iid}}{\sim} \mathcal{U}(-\pi, \pi), i = 1, \ldots, n$ is simulated independently from U_i and X_i. We consider two different models for the regression function, given by $m(x) = (x/\pi)^2$ and $m(x) = \cos(x)$. Moreover, we consider the case of homoscedastic ($\sigma(x) = 0.5$) and heteroscedastic ($\sigma(x) = 0.5 \exp(-(x/\pi)^2)$) errors.

Table 2.1: Simulation Results for $n = 100$

h_0	1.00	1.50	2.00	2.50	3.00
Panel A: $m(x) = \cos(x)$					
$\sigma(x) = 0.5$					
0.10	0.069	0.060	0.065	0.095	0.106
0.05	0.009	0.008	0.023	0.035	0.038
p-value	0.632	0.593	0.580	0.575	0.650
$\sigma(x) = 0.5\exp(-(x/\pi)^2)$					
0.10	0.036	0.039	0.060	0.081	0.105
0.05	0.006	0.011	0.013	0.020	0.047
p-value	0.701	0.623	0.602	0.594	0.612
Panel B: $m(x) = (x/\pi)^2$					
$\sigma(x) = 0.5$					
0.10	0.039	0.071	0.070	0.085	0.099
0.05	0.006	0.018	0.023	0.031	0.041
p-value	0.689	0.632	0.592	0.577	0.572
$\sigma(x) = 0.5\exp(-(x/\pi)^2)$					
0.10	0.054	0.051	0.078	0.091	0.131
0.05	0.011	0.016	0.024	0.045	0.056
p-value	0.692	0.616	0.572	0.565	0.603

For the nonparametric regression we use the forth order kernel $K(u) = \frac{15}{32}(7x^4 - 10x^2 + 3)\mathbb{1}_{[-1,1]}(x)$. The bandwidth sequences are chosen by the simple plug-in

Table 2.2: Simulation Results for $n = 100$,$m(x) = \cos(x), \sigma(x) = 0.5$

h_0	1.00	1.50	2.00	2.50	3.00
			$\delta = 4, \widetilde{\delta} = 6$		
0.10	0.031	0.032	0.048	0.057	0.081
0.05	0.001	0.005	0.006	0.017	0.030
p-value	0.714	0.644	0.603	0.588	0.584
			$\eta_i^* \sim \mathcal{N}(0, 1)$		
0.10	0.026	0.028	0.052	0.068	0.095
0.05	0.001	0.005	0.007	0.026	0.036
p-value	0.688	0.636	0.602	0.572	0.604

rules[3] $h_X = h_0 n^{-1/\delta} \widehat{sd}(X), h_Z = h_0 n^{-1/\delta} \widehat{sd}(Z)$ and $\widetilde{h}_X = h_0 n^{-1/\widetilde{\delta}} \widehat{sd}(X)$, where we use different values for h_0 to investigate the performance of the test for different bandwidths. For simplicity we use the same constant h_0 for all three bandwidths. Unless otherwise stated we use $\delta = 5$ and $\widetilde{\delta} = 6$ which clearly fulfill Assumption 2.4 for a kernel of order $r = 4$.

The bootstrap is implemented with $B = 199$ iterations. The residuals are drawn from the classical two-point distribution given in Section 2.2.4. All tables report the proportion of rejection based on 1 000 Monte-Carlo iterations.

The results for a sample size of $n = 100$ are displayed in Table 2.1. We find that the test tends to be too conservative for small bandwidths. However we observe no severe distortion of the level for the chosen bandwidth constants. Given that $n = 100$ is a relative small sample size for a two-dimensional nonparametric regression problem, the empirical level of the test is surprisingly accurate for a wide range of bandwidths.

In Table 2.2 we investigate two deviations from the general setting. In the upper panel we report the rejection rates, if we choose $\delta = 4$ and $\widetilde{\delta} = 6$. This

[3] We denote the empirical standard deviation of a random variable X with $\widehat{sd}(X) = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}$.

choice is admissible by Assumption 2.4 and we observe that the test is even more conservative for this choice of bandwidth. Overall we observe the need for selecting a rather large bandwidth. The implementation of a forth order kernel reduces the bias even for large values of h. This means that the residuals can be estimated more accurately, which allows to approximate the expected value and variance of the test statistic better than with small bandwidths.

In the lower panel of Table 2.2 we use a different distribution in the bootstrap procedure. Here, the resampled residuals are $\varepsilon_i^* = \widehat{\varepsilon}_i \eta_i^*$ where $\eta_i^* \sim \mathcal{N}(0,1)$. In contrast to the setting in Table 2.1 this bootstrap distribution only matches the first two moments of the empirical residuals. While this is sufficient to prove the asymptotic validity of the bootstrap, we find that using the normal distribution produces more conservative results. This provides further evidence to our conjecture that mimicking more moments of the residuals leads to a higher order approximation of the finite sample distribution of $\widehat{\Gamma}_{\mathcal{K}}$ by the bootstrap.

Next, we increase the sample size and simulate with $n = 200$. We return to the general setting ($\delta = 5, \widetilde{\delta} = 6$ and mimicking three moments in the bootstrap) and give the results in Table 2.3. Obviously, the empirical level stabilizes to its desired value for a wide range of bandwidths, but we still observe underrejection for small values of h_0.

In the second part of the Monte Carlo study we simulate the empirical power of the test under H_1. To this end, we use the model

$$Y_i = \cos(X_i) + g_\lambda(Z_i) + 0.5U_i, \quad i = 1, \dots, n,$$

and retain the previous setup, i. e., $X_i, Z_i \overset{\text{iid}}{\sim} \mathcal{U}(-\pi, \pi)$ and $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ independent of (X_i, Z_i). The parameter λ measures the deviation from the null hypothesis in the L_2-sense. For the nonparametric regression and the bootstrap procedure we use the same setting as under the null hypothesis. We have restricted the simulation study to homoscedastic errors and a sample size of $n = 100$.

As a first specification for the alternative, we choose $g_\lambda(z) = \lambda \cos(z)$. In this case the L_2-distance is $\mathbf{E}(\mu(X,Z) - m(X))^2 = \lambda^2/2$. In Figure 2.1 the empirical power for this experiment is displayed for four different values of h_0 and for levels of $\alpha = 0.10$ and $\alpha = 0.05$. We see that for all bandwidths the test is consistent against this alternative, but observe lower power for small bandwidths.

Next, we consider $g_\lambda(z) = \lambda(z/\pi)^2$, which leads to $\mathbf{E}(\mu(X,Z) - m(X))^2 = 4\lambda^2/45$. To obtain the same L_2-distance as in the cosine specification, a different range of λ is selected in Figure 2.2. A comparison of the two experiments shows

Table 2.3: Simulation Results for $n = 200$

h_0	1.00	1.50	2.00	2.50	3.00
Panel A: $m(x) = \cos(x)$					
$\sigma(x) = 0.5$					
0.10	0.045	0.061	0.071	0.112	0.114
0.05	0.009	0.018	0.033	0.041	0.055
p-value	0.642	0.588	0.560	0.573	0.584
$\sigma(x) = 0.5 \exp(-(x/\pi)^2)$					
0.10	0.035	0.061	0.083	0.102	0.127
0.05	0.003	0.021	0.025	0.044	0.066
p-value	0.635	0.588	0.574	0.568	0.601
Panel B: $m(x) = (x/\pi)^2$					
$\sigma(x) = 0.5$					
0.10	0.038	0.061	0.079	0.107	0.112
0.05	0.007	0.025	0.026	0.040	0.058
p-value	0.648	0.595	0.575	0.572	0.586
$\sigma(x) = 0.5 \exp(-(x/\pi)^2)$					
0.10	0.039	0.073	0.074	0.101	0.131
0.05	0.010	0.015	0.027	0.044	0.067
p-value	0.624	0.582	0.570	0.564	0.594

that the empirical power of the test is even better for the quadratic alternative if we keep the L_2-distance constant.

In contrast, high frequency alternatives are difficult to detect. As an example

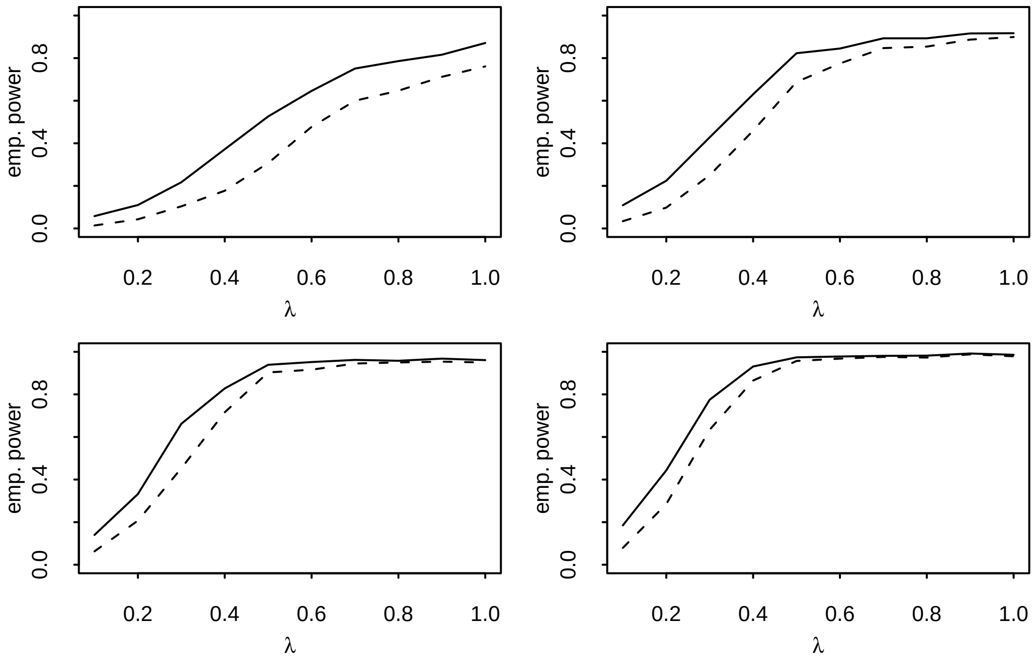

Figure 2.1: Simulated power for $g_\lambda(z) = \lambda \cos(z)$. The bandwidth constants are $h_0 = 1.50, 2.00, 2.50, 3.00$ (upper left, upper right, lower left, lower right). Levels are given by $\alpha = 0.10$ (solid) and $\alpha = 0.05$ (dashed)

we use $g_\lambda(z) = \cos(\lambda z)$ where $\lambda \in \mathbb{Z}$. The L_2-distance for such alternatives is constant $\mathbf{E}(\mu(X,Z) - m(X))^2 = 1/2$, but for higher values of λ estimation becomes difficult. Since the two-dimensional regression $\widehat{\mu}_h(x,z)$ estimates high frequencies poorly, the test breaks down in finite samples. In this case, small bandwidths are favorable, because they enable a better approximation for high frequencies. The simulated power in Figure 2.3 underscores this.

Finally we look at low frequency alternatives, given by $g_\lambda(z) = \cos(\lambda z)$ with $\lambda \in [0,1]$. The L_2-distance for these alternatives varies between 0 and 0.6, given by $\mathbf{E}(\mu(X,Z) - m(X))^2 = (1 - \sin(2\lambda\pi)/(2\lambda\pi))/2$. In contrast to the high frequency alternatives low bandwidths have lower power, which can be seen in Figure 2.4. The test is consistent for all bandwidths, but this type of alternatives is more difficult to detect than the specifications in Figures 2.1 or 2.2.

This simulation study underlines that the proposed test statistic produces reliable results for moderate sample sizes. the different alternatives under consideration highlight the role of the bandwidth. While small bandwidths are ad-

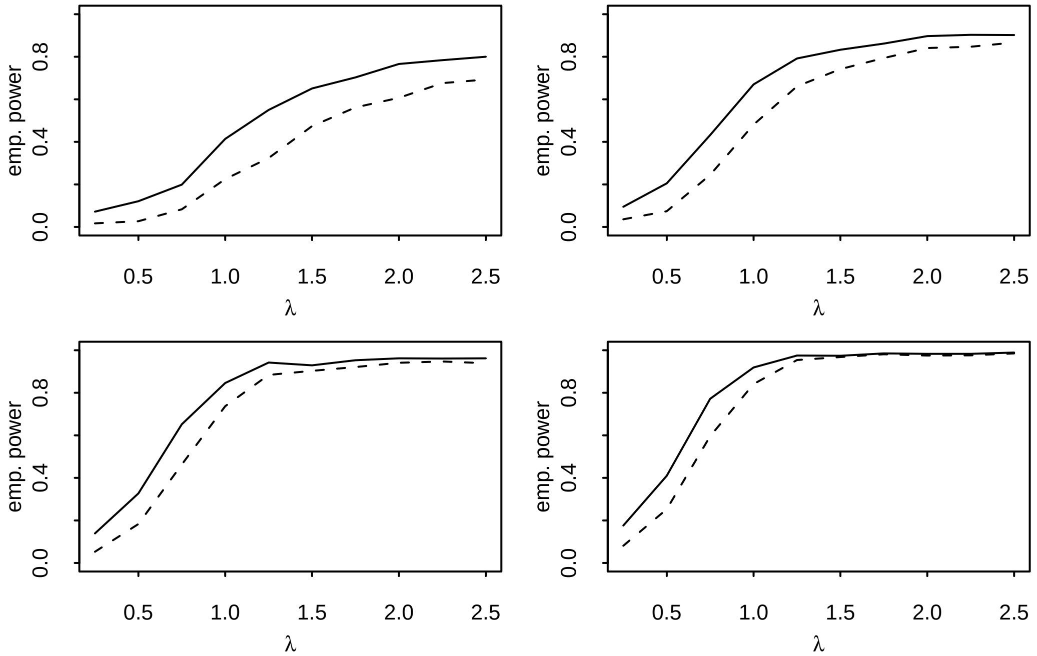

Figure 2.2: Simulated power for $g_\lambda(z) = \lambda(z/\pi)^2$. The bandwidth constants are $h_0 = 1.50, 2.00, 2.50, 3.00$ (upper left, upper right, lower left, lower right). Levels are given by $\alpha = 0.10$ (solid) and $\alpha = 0.05$ (dashed)

vantageous to detect high frequency deviations from the null hypothesis, large bandwidths have better power against low frequency alternatives. A data adaptive bandwidth selection procedure similarly to Horowitz and Spokoiny (2001) could be preferable, but this extension is left for future research.

2.5 An Empirical Application: Homogeneity in Consumer Demand

It is an implication of a linear budget set that individual demand is homogeneous of degree zero. Formally, in the standard formulation involving budget shares $Y \in [0,1]^{d_Y}$, log income $X \in \mathbb{R}$ as well as log prices $P \in \mathbb{R}^{d_Y}$, and the relationship $Y = \mu(P, X) + U$, with $\mathbf{E}(U \mid P, X) = 0$, we obtain that

$$\mu(P, X) = \mu(P - X, 0) = m(\widetilde{P}),$$

where $\widetilde{P} = P - X$. Hence, a test of whether $\mathbf{E}(Y \mid \widetilde{P}, X) = m(\widetilde{P})$ can be seen as a test for homogeneity.

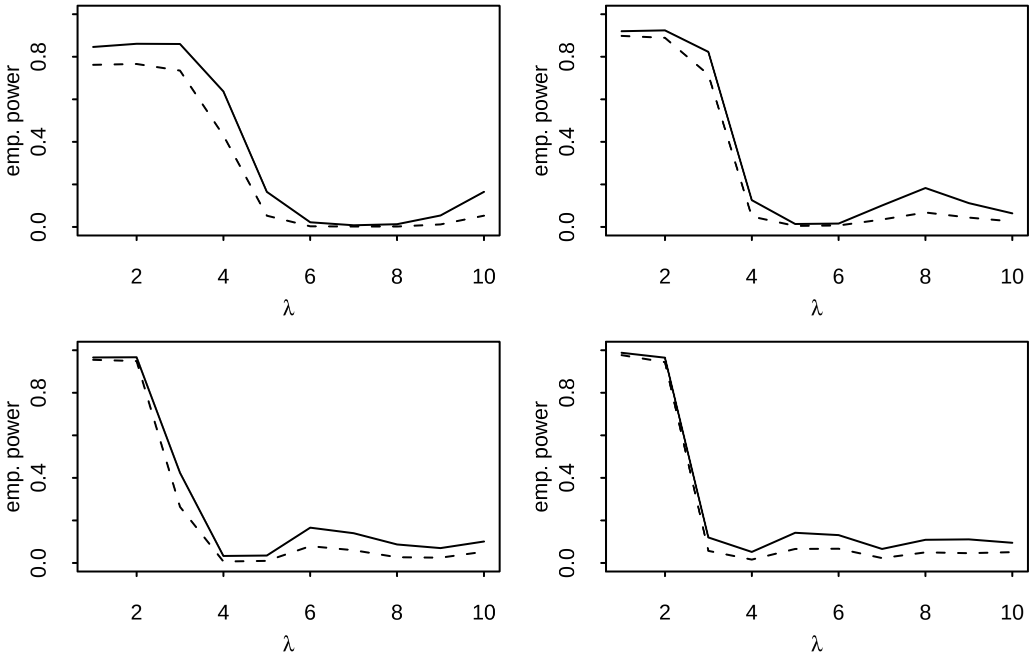

Figure 2.3: Simulated power for $g_\lambda(z) = \cos(\lambda z)$. The bandwidth constants are $h_0 = 1.50, 2.00, 2.50, 3.00$ (upper left, upper right, lower left, lower right). Levels are given by $\alpha = 0.10$ (solid) and $\alpha = 0.05$ (dashed)

The testing procedure will be applied to this specific problem of testing the significance of a regressor using British household data. Every year, the Family Expenditure Survey (FES) reports the income, expenditures, demographic composition and other characteristics of about 7 000 households. The sample surveyed represents about 0.05 % of all households in the United Kingdom. The information is collected partly by interview and partly by records. Records are kept by each household member, and include an itemized list of expenditures during 14 consecutive days. The periods of data collection are evenly spread out over the year. The information is then compiled and provides a repeated series of yearly cross-sections.

All the goods are grouped into three categories, Group 1 to 3. The first category is related to food consumption and consists of the subcategories food bought, food out (catering) and tobacco, which are self-explanatory. The second category contains expenditures which are related to the house, namely housing (a more heterogeneous category; it consists of rent or mortgage payments), furniture as well as household goods and services. Finally, the last group consists of motor-

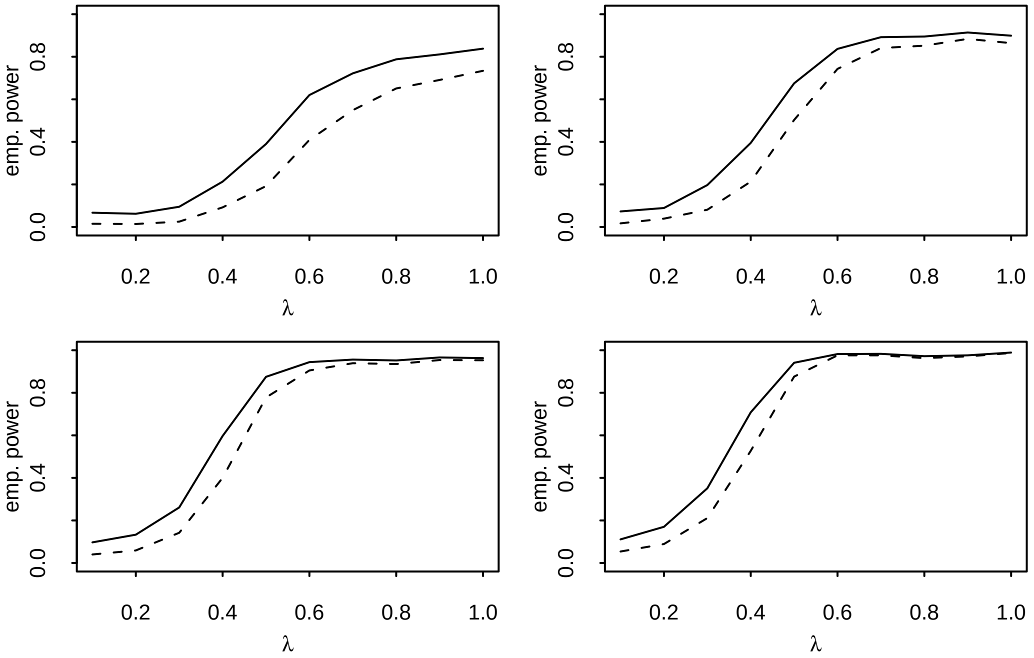

Figure 2.4: Simulated power for $g_\lambda(z) = \cos(\lambda z)$. The bandwidth constants are $h_0 = 1.50, 2.00, 2.50, 3.00$ (upper left, upper right, lower left, lower right). Levels are given by $\alpha = 0.10$ (solid) and $\alpha = 0.05$ (dashed)

Table 2.4: Value of Test Statics for different Choice of Bandwidth Constant and Time Interval

	Period					
	1974–1980		1981–1988		1989–1994	
h_0	$\widehat{\Gamma}_{\mathcal{K}}$	p-value	$\widehat{\Gamma}_{\mathcal{K}}$	p-value	$\widehat{\Gamma}_{\mathcal{K}}$	p-value
1.5	1.4214	(0.60)	1.2263	(0.05)	1.2392	(0.23)
2.0	0.7830	(0.71)	0.7192	(0.12)	0.5663	(0.60)
2.5	0.5674	(0.45)	2.7167	(0.26)	0.3310	(0.81)
3.0	0.4093	(0.27)	2.9898	(0.21)	0.1979	(0.80)

ing and fuel expenditures, categories that are often related to energy prices. For brevity, we call these categories food, housing and energy. These broader categories are formed since more detailed accounts suffer from infrequent purchases (recall that the recording period is 14 days) and are thus underreported. Together they account for 50-60 % of expenditures. We removed outliers by excluding the upper and lower 2.5 % of the population in the three groups.

"Income" in demand analysis is total expenditure, under an additive separability assumption of preferences over time and decisions. It is obtained by adding up all expenditures, with a few exceptions which are known to have measurement error like tobacco. This is done to define nominal income; real income is then obtained by dividing through the retail price indices.

In this paper, we stratify the population to obtain more homogeneous subpopulations. More specifically, like much of the demand literature we focus on one subpopulation, namely two person households, sampled in a certain time interval, both adults, at least one of which is working and the head of household is a white-collar worker. This focus is also justified because other subpopulations are much more prone to measurement problems. It is likely that there is remaining preference heterogeneity. However, we abstract from this problem here, but see Hoderlein (2005) on this issue.

To test whether X can be omitted from the regression of Y on $\widetilde{P}$ the following specifications are used: In accordance with our assumptions, we set $\delta = 4$ and $\widetilde{\delta} = 6$, and determine h and $\widetilde{h}$ through $h = h_0 n^{-1/\delta} \widehat{sd}(\widetilde{P})$, and $\widetilde{h} = h_0 n^{-1/\widetilde{\delta}} \widehat{sd}(\widetilde{P}, X)$. The same forth order kernel as in the simulation study was used. Table 2.4 shows the result of our test statistic for homogeneity of degree zero for various values of the bandwidth constant h_0 and time periods. The p-values based on 199 bootstrap implementations are in brackets. We conjecture that homogeneity is generally accepted, as there is only one test rejected at the level of 0.05. Of course, if we perform such a high number of tests, one test is likely to reject. Somehow, we would like to correct for the dependence between the tests.

2.6 Conclusion

The bootstrap simplifies relatively complicated nonparametric procedures and makes them therefore accessible for applications. At the same time, the bootstrap helps improving the small sample properties. This chapter, which considers nonparametric specification testing, underscores these advantages in our specific setting. In particular, we show that our bootstrap-based tests are simple, easy to implement and work well in quite small samples, where they outperform other comparable tests in the literature.

In addition to these small sample advantages we also establish by asymptotic arguments that our tests are at least as good as other existing tests proposed in the literature. In particular it should be noted that Aït-Sahalia, Bickel and Stoker (2001) use in their simulation study sample sizes of more than 500 observations. Finally, we provide new extensions that are important for applications, like allowing for local polynomial estimators, dependent data and systems of equations under the same format.

Appendix

Proof of Theorem 2.1

For abbreviation we introduce $V_i = (X_i, Z_i)$ and $W_i = (Y_i, X_i, Z_i)$ and decompose the statistic in the following way

$$\begin{aligned}\widehat{\Gamma}_{\mathcal{K}} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}\frac{K_h(V_i-V_j)}{\widehat{f}_h(V_i)}(Y_j^k-\widehat{m}^k_{\tilde{h}}(X_j))\Big)^2 A_i \\ &= \widehat{\Gamma}_{\mathcal{K}1}+\widehat{\Gamma}_{\mathcal{K}2}+\widehat{\Gamma}_{\mathcal{K}3}+\widehat{\Gamma}_{\mathcal{K}4}, \qquad (2.9)\end{aligned}$$

where

$$\begin{aligned}\widehat{\Gamma}_{\mathcal{K}1} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(V_i-V_j)\frac{Y_j^k-\mu^k(V_j)}{\widehat{f}_h(V_i)}\Big)^2 A_i \\ \widehat{\Gamma}_{\mathcal{K}2} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(V_i-V_j)\frac{\mu^k(V_j)-m^k(X_j)}{\widehat{f}_h(V_i)}\Big)^2 A_i \\ \widehat{\Gamma}_{\mathcal{K}3} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(V_i-V_j)\frac{m^k(X_j)-\widehat{m}^k_{\tilde{h}}(X_j)}{\widehat{f}_h(V_i)}\Big)^2 A_i\end{aligned}$$

and $\widehat{\Gamma}_{\mathcal{K}4}$ contains all cross terms. Note that under H_0 we have that $\widehat{\Gamma}_{\mathcal{K}2} = 0$ almost surely. We start by investigating $\widehat{\Gamma}_{\mathcal{K}1}$, which yields the asymptotic distribution and show that $\widehat{\Gamma}_{\mathcal{K}3}$ and $\widehat{\Gamma}_{\mathcal{K}4}$ are of lower order afterwards.

First, we write

$$\begin{aligned}\widehat{\Gamma}_{\mathcal{K}1} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(V_i - V_j)\frac{Y_j^k - \mu^k(V_j)}{f(V_i)}\Big)^2\Big(\frac{f(V_i)}{\widehat{f}_h(V_i)}\Big)^2 A_i\\ &= (I_{\mathcal{K}n} + \Delta_{\mathcal{K}n})(1 + o_P(1)),\end{aligned}$$

where we have defined

$$I_{\mathcal{K}n} = \int\sum_{k=1}^{d_Y}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(v - V_j)\frac{Y_j^k - \mu^k(V_j)}{f(v)}\Big)^2 a(v)f(v)\,\mathrm{d}v \tag{2.10}$$

$$\Delta_{\mathcal{K}n} = \int\sum_{k=1}^{d_Y}\Big(\frac{1}{n}\sum_{j=1}^{n}K_h(v - V_j)\frac{Y_j^k - \mu^k(V_j)}{f(v)}\Big)^2 a(v)(\widehat{f}_e(v) - f(v)\,\mathrm{d}v, \tag{2.11}$$

and $\widehat{f}_e = \frac{1}{n}\sum_{i=1}^{n}\delta_{(V_i)}(v)$ denotes the empirical distribution of the sampled data (where $\delta_{(V_i)}$ is the Dirac-measure at V_i).
Starting with the leading term, we rearrange $I_{\mathcal{K}n}$ to obtain

$$\begin{aligned}I_{\mathcal{K}n} &= \frac{1}{n^2}\sum_{i<j}\sum_{k=1}^{d_Y}\int K_h(v - V_i)\frac{Y_i^k - \mu^k(V_i)}{f(v)}K_h(v - V_j)\frac{Y_j^k - \mu^k(V_j)}{f(v)}a(v)f(v)\,\mathrm{d}v\\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{d_Y}\int\big(K_h(v - V_i)\frac{Y_i^k - \mu^k(V_i)}{f(v)}\big)^2 a(v)f(v)\,\mathrm{d}v\\ &= I_{\mathcal{K}n,1} + I_{\mathcal{K}n,2}.\end{aligned} \tag{2.12}$$

Now it remains to show

$$nh^{(d_X+d_Z)/2}I_{\mathcal{K}n,1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_{\mathcal{K}}^2) \tag{2.13}$$

$$nh^{(d_X+d_Z)/2}I_{\mathcal{K}n,2} - h^{-(d_X+d_Z)/2}B_{\mathcal{K}} \xrightarrow{P} 0 \tag{2.14}$$

$$nh^{(d_X+d_Z)/2}\Delta_{\mathcal{K}n} \xrightarrow{P} 0. \tag{2.15}$$

From this the statement of the theorem follows.

Proof of (2.13) Write

$$I_{\mathcal{K}n,1} = \sum_{i<j}h_n(W_i, W_j)$$

as U-statistic with kernel

$$h_n(W_i, W_j) = \frac{2}{n^2 h^{d_X+d_Z}} \sum_{k=1}^{d_Y} (Y_i^k - \mu^k(V_i))(Y_j^k - \mu^k(V_j)) \\ \times \int K(u)K(u + (V_i - V_j)/h) \frac{a(V_i + uh)}{f(V_i + uh)} \,\mathrm{d}u.$$

where a change of variables has been applied. Asymptotic normality is shown by using a central limit theorem for generalized U-statistics (see Lemma 3.1 by de Jong, 1987). Under the conditions

$$\text{(2.16)} \qquad \frac{\max_{1\le i\le n} \sum_{j=1}^n \mathbf{E}\, h_n(W_i, W_j)}{\mathbf{var}\, I_{\mathcal{K}n,1}} \xrightarrow{P} 0 \qquad \text{and} \qquad \frac{\mathbf{E}\, I_{\mathcal{K}n,1}^4}{(\mathbf{var}\, I_{\mathcal{K}n,1})^2} \xrightarrow{P} 3$$

it follows that

$$\sqrt{2} \frac{I_{\mathcal{K}n,1}}{\sqrt{\mathbf{var}\, I_{\mathcal{K}n,1}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

It is immediate to see that the kernel is degenerate, symmetric and centered. Now, we introduce $\sigma_n^2 = \mathbf{E}\, h_n(W_i, W_j)^2$. As we have independent and identically distributed data we can write

$$\max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^n \mathbf{E}\, h_n(W_i, W_j)^2 = (n-1)\sigma_n^2$$

and

$$\begin{aligned} \mathbf{var}\, I_{\mathcal{K}n,1} &= \sum_{i_1<i_2} \mathbf{var}\, h_n(W_{i_1}, W_{i_2}) \\ &\quad + \sum_{i_1<i_2} \sum_{\substack{i_3<i_4\\ (i_3,i_4)\neq(i_1,i_2)}} \mathbf{cov}(h_n(W_{i_1}, W_{i_2}), h_n(W_{i_3}, W_{i_4})) \\ &= \frac{n(n-1)}{2}\sigma_n^2 \end{aligned}$$

because $h_n(\cdot,\cdot)$ is centered. From these two results the first condition in equation (2.16) is established. For the second calculate

(2.17)

$$\begin{aligned} \mathbf{E}\, I_{\mathcal{K}n,1}^4 = \sum_{i_1<i_2} \mathbf{E}\, h_n(W_{i_1}, W_{i_2})^4 + 3 \sum_{i_1<i_2} \sum_{\substack{i_3<i_4\\ (i_3,i_4)\neq(i_1,i_2)}} \mathbf{E}\, h_n(W_{i_1}, W_{i_2})^2 h_n(W_{i_3}, W_{i_4})^2 \\ + 24 \sum_{i_1<i_2} \sum_{i_3\neq i_1,i_2} \mathbf{E}\, h_n(W_{i_1}, W_{i_2})^2 h_n(W_{i_1}, W_{i_3}) h_n(W_{i_2}, W_{i_3}) \\ +3 \sum_{i_1} \sum_{i_2\neq i_1} \sum_{i_3\neq i_1,i_2} \sum_{i_4\neq i_1,i_2,i_3} \mathbf{E}\, h_n(W_{i_1}, W_{i_2}) h_n(W_{i_2}, W_{i_3}) h_n(W_{i_3}, W_{i_4}) h_n(W_{i_4}, W_{i_1}) \end{aligned}$$

where all vanishing terms (with $\mathbf{E}\, h_n(W_{i_1}, W_{i_2}) = 0$) are omitted. To show the second condition, the remaining terms have to be calculated. Starting with the denominator, we have to calculate

$$\sigma_n^2 = \mathbf{E}\, h_n(W_1, W_2)^2. \tag{2.18}$$

Resolving the square and changing variables[4] to $\widetilde{v} = (v - v_1)/h$ together with expanding $a(\cdot)$ and $f(\cdot)$ yields

$$\begin{aligned}
\sigma_n^2 = \frac{4}{n^4 h^{2(d_X+d_Z)}} \sum_{k,k'} \iint & K(\widetilde{v}) \frac{y_1^k - \mu^k(v_1)}{f(v_1)} \\
& \times K(\widetilde{v} + (v_1 - v_2)/h) \frac{y_2^k - \mu^k(v_2)}{f(v_1)} a(v_1) f(v_1)\, \mathrm{d}\widetilde{v} \\
\times \int K(\widetilde{v}) \frac{y_1^{k'} - \mu^{k'}(v_1)}{f(v_1)} & K(\widetilde{v} + (v_1 - v_2)/h) \frac{y_2^{k'} - \mu^{k'}(v_2)}{f(v_1)} a(v_1) f(v_1)\, \mathrm{d}\widetilde{v} \\
& \times f(y_1, v_1) f(y_2, v_2)\, \mathrm{d}y_1\, \mathrm{d}v_1\, \mathrm{d}y_2\, \mathrm{d}v_2 (1 + O(h))
\end{aligned}$$

Now substitute $\widetilde{\widetilde{v}} = (v_1 - v_2)/h$ to obtain

$$\begin{aligned}
&= \frac{4(\kappa_*)^{d_X+d_Z}}{n^4 h^{d_X+d_Z}} \sum_{k,k'} \int (y_1^k - \mu^k(v_1))(y_2^k - \mu^k(v_1))(y_1^{k'} - \mu^{k'}(v_1))(y_2^{k'} - \mu^{k'}(v_1)) \\
&\qquad \times \Big(\frac{a(v_1)}{f(v_1)}\Big)^2 f(y_1, v_1) f(y_2, v_1)\, \mathrm{d}y_1\, \mathrm{d}y_2\, \mathrm{d}v_1 (1 + O(h)) \\
&= \frac{4(\kappa_*)^{d_X+d_Z}}{n^4 h^{d_X+d_Z}} \sum_{k,k'} \int \Big(\int (y_1^k - \mu^k(v_1))(y_1^{k'} - \mu^{k'}(v_1)) \frac{f(y_1, v_1)}{f(v_1)}\, \mathrm{d}y_1 \Big)^2 a(v_1)^2\, \mathrm{d}v_1 \\
&\qquad \times (1 + O(h)) \\
&= \frac{2}{n^4 h^{d_X+d_Z}} \Sigma_{\mathcal{K}}^2 (1 + O(h)).
\end{aligned}$$

Similar calculations show that

$$\begin{aligned}
\mathbf{E}\, h_n(W_1, W_2)^4 &= O(n^{-8} h^{-3(d_X+d_Z)}) \\
\mathbf{E}\, h_n(W_1, W_2)^2 h_n(W_1, W_3)^2 &= O(n^{-8} h^{-2(d_X+d_Z)}) \\
\mathbf{E}\, h_n(W_1, W_2)^2 h_n(W_1, W_3) h_n(W_2, W_3) &= O(n^{-8} h^{-2(d_X+d_Z)}) \\
\mathbf{E}\, h_n(W_1, W_2) h_n(W_2, W_3) h_n(W_3, W_4) h_n(W_1, W_4) &= O(n^{-8} h^{-(d_X+d_Z)}).
\end{aligned}$$

[4] Here the notation is simplified. As v_1 is $d_X + d_Z$-dimensional one has to apply $d_X + d_Z$ substitutions.

Using combinatorical arguments it can be established from equation (2.17) that $\mathbf{E}\, I^4_{\mathcal{K}n,1}$ is asymptotically dominated by terms with $\mathbf{E}\, h_n(W_1, W_2)^2 h_n(W_3, W_4)^2 = (\mathbf{E}\, h_n(W_1, W_2)^2)^2$. Therefore the second condition in equation 2.16 is fulfilled as

$$\frac{\mathbf{E}\, I^4_{\mathcal{K}n,1}}{(\mathbf{var}\, I_n)^2} = \frac{12n^{-4}h^{-2(d_X+d_Z)}\Sigma^4_{\mathcal{K}}(1+o(1))}{(2n^{-2}h^{-(d_X+d_Z)}\Sigma^2_{\mathcal{K}}(1+o(1)))^2} \longrightarrow 3$$

and weak convergence of $I_{\mathcal{K}n,1}$ is established.

Proof of (2.14) The expected value of the test statistic is given by

$$\mathbf{E}\, I_{\mathcal{K}n,2} = \frac{1}{n}\sum_{k=1}^{d_Y}\iint\Big(K_h(v-v_1)\frac{y_1^k-\mu^k(v_1)}{f(v)}\Big)^2 a(v)f(v)\,\mathrm{d}v f(y_1,v_1)\,\mathrm{d}y_1\,\mathrm{d}v_1.$$

Changing variables and expanding yields

$$\begin{aligned}&= \frac{\kappa_0^2}{nh^{d_X+d_Z}}\sum_{k=1}^{d_Y}\int(y_1^k-\mu^k(v_1))^2\frac{a(v_1)}{f(v_1)}f(y_1,v_1)\,\mathrm{d}v_1(1+O(h))\\ &= n^{-1}h^{-(d_X+d_Z)}B_{\mathcal{K}}(1+O(h^r)).\end{aligned}$$

Convergence in probability follows from Markov's inequality with second moments, which requires to calculate

$$\frac{1}{n^4}\Big(\int\sum_{k=1}^{d_Y}\big(K_h(v-v_1)(y_1^k-\mu^k(v_1))\big)^2\frac{a(v)}{f(v)}\,\mathrm{d}v\Big)^2 f(y_1,v_1)\,\mathrm{d}y_1\,\mathrm{d}v_1.$$

Changing variables as before results in

$$\frac{\kappa_0^2}{n^4h^{2(d_X+d_Z)}}\sum_{k,k'}\int(y_1^k-\mu^k(v_1))^2(y_1^{k'}-\mu^{k'}(v_1))^2\frac{a(v_1)^2}{f(v_1)^2}f(y_1,v_1)\,\mathrm{d}y_1\,\mathrm{d}v_1(1+o(1)),$$

which is bounded by Assumption 2.2. In total this yields

$$\mathbf{E}\, I^2_{\mathcal{K}n,2} = O(n^{-3}h^{-2(d_X+d_Z)}) = o(n^{-2}h^{-(d_X+d_Z)})$$

and convergence in probability of $I_{\mathcal{K}n,2}$ follows.

Proof of (2.15) For this statement we will restrict to the case when $d_Y = 1$. Then convergence in probability has to be shown for

$$\Delta_{\mathcal{K}n} = \frac{1}{n^3}\sum_{i,j,k}\gamma_n(W_i, W_j, W_k),$$

where

$$\gamma_n(W_i, W_j, W_k) = \widetilde{\gamma}_n(W_i, W_j, W_k) - \int \widetilde{\gamma}_n(W_i, W_j, w) f(w)\, \mathrm{d}w,$$

with

$$\widetilde{\gamma}_n(W_i, W_j, W_k) = K_h(V_k - V_i) \frac{Y_i - \mu^1(V_i)}{f(V_k)} a(V_k) K_h(V_k - V_j) \frac{Y_j - \mu^1(V_j)}{f(V_k)} a(V_k).$$

First we show that the expectation tends to zero

$$\mathbf{E}\, \Delta_{\mathcal{K}n} = \frac{1}{n^3} \sum_{i,j,k} \mathbf{E}\, \gamma_n(W_i, W_j, W_k) = o(n^{-1} h^{-(d_X + d_Z)/2}),$$

where only the cases $i = k \neq j, j = k \neq i$ and $i = j = k$ have to be considered, all others have expectation zero. In the remaining cases, two (resp. one) substitution can be applied and their total contribution is $O(n^{-1} h^{2(d_X+d_Z)} + n^{-2} h^{d_X+d_Z})$.
Then, Markov's inequality is applied with the second moments and we have to investigate

$$\mathbf{E}\, \Delta^2_{\mathcal{K}n} = \frac{1}{n^6} \sum_{ijk} \mathbf{E}\, \gamma_n(W_i, W_j, W_k)^2 + \frac{2}{n^6} \sum_{ijk} \sum_{i'j'k'} \mathbf{E}\, \gamma_n(W_i, W_j, W_k) \gamma_n(W_{i'}, W_{j'}, W_{k'}).$$

The covariance parts vanish, whenever $k \neq k'$. If $k = k'$ the covariance terms are zero by the conditional independence of the error terms, in all cases where $i \neq i'$ or $j \neq j'$. For the remaining cases we have to distinguish if the number of different indices is $N = 2, 3$. Then, the overall contribution of these terms is $O(n^{N-6} h^{-4(d_X+d_Z)} h^{N(d_X+d_Z)}) = o(n^{-2} h^{-(d_X+d_Y)})$.
Next, consider the variance terms. If there are three different indices, two changes of variables can be applied and the overall contribution is $O(n^{-3} h^{-2(d_X+d_Z)}) = o(n^{-2} h^{-(d_X+d_Z)})$. If there are two different indices, one change of variables can be applied and we obtain terms of order $O(h^{-3(d_X+d_Z)})$ with a total contribution of $O(n^{-4} h^{-3(d_X+d_Z)}) = o(n^{-2} h^{-(d_X+d_Z)})$. If $i = j = k$ one change of variables is still possible and the contribution is $O(n^{-5} h^{-3(d_X+d_Z)}) = o(n^{-2} h^{-(d_X+d_Z)})$. This completes the proof of equation (2.15).

Convergence in Probability of $\widehat{\Gamma}_{\mathcal{K}3}$ For the third term in (2.9) it holds that

$$\begin{aligned}|\widehat{\Gamma}_{\mathcal{K}3}| &\leq \max_{k=1,\dots,d_Y} \sup_{x\in\mathcal{A}} |m^k(X_j) - \widehat{m}^k_{\widetilde{h}}(X_j)|^2 \sup_{v\in\mathcal{A}} |a(v)| \\ &= O_P(\widetilde{h}^{2r} + \frac{\log n}{n\widetilde{h}^{d_X}}) \\ &= o_P\big(n^{-1}h^{-(d_X+d_Z)/2}\big)\end{aligned}$$

under Assumption 2.4.3.

Convergence in Probability of $\widehat{\Gamma}_{\mathcal{K}4}$ The non-zero parts are given by

$$\begin{aligned}\widehat{\Gamma}_{\mathcal{K}4} = \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}&\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(V_i - V_j)\frac{Y_j^k - \mu^k(V_j)}{\widehat{f}_h(V_i)}\Big) \\ &\times \Big(\frac{1}{n}\sum_{j'=1}^{n} K_h(V_i - V_{j'})\frac{m^k(X_{j'}) - \widehat{m}^k_{\widetilde{h}}(X_{j'})}{\widehat{f}_h(V_i)}\Big)A_i \\ = \sum_{i,j,j'} \gamma_{ijj'}.&\end{aligned}$$

Because

$$\mathbf{E}\,\varepsilon_j^k\big(m^k(V_{j'}) - \widehat{m}_{\widetilde{h}}(V_{j'})\big) \mid V_1,\dots,V_n) = n^{-1}K_{\widetilde{h}}(V_j - V_{j'})\sigma^2(V_j)$$

we have that

$$\mathbf{E}\,\widehat{\Gamma}_{\mathcal{K}4} = O(n^{-1}) = o(n^{-1}h^{-(d_X+d_Z)/2}).$$

It follows from similar considerations as done to show (2.15) that

$$\mathbf{E}\,\widehat{\Gamma}^2_{\mathcal{K}4} = o(n^{-2}h^{-(d_X+d_Z)}).$$

This completes the proof of the theorem □

Proof of Theorem 2.2

Under H_{1n} the decomposition (2.9) remains valid and the asymptotic analysis of $\widehat{\Gamma}_{\mathcal{K}1}$ and $\widehat{\Gamma}_{\mathcal{K}3}$ is unchanged. However $\widehat{\Gamma}_{\mathcal{K}2}$ is not zero any longer. If it holds that $\mu(x,z) = m(x) + \varepsilon_n(x,z)$, we have that

$$\begin{aligned}\widehat{\Gamma}_{\mathcal{K}2} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(V_i - V_j)\frac{\varepsilon_n^k(V_j)}{f(V_i)}\Big)^2 A_i(1 + o_P(1)) \\ &= \sum_{k=1}^{d_Y}\int\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(v - V_j)\frac{\varepsilon_n^k(V_j)}{f(v)}\Big)^2 a(v)\,\mathrm{d}v + o_P(n^{-1}h^{(d_X+d_Z)/2}).\end{aligned}$$

This follows from similar calculations as to show (2.15). Omitting the lower order terms it holds that

$$\begin{aligned} nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}2} &= \int K(u)^2\,\mathrm{d}u \sum_{k=1}^{d_Y}\frac{1}{n}\sum_{j=1}^{n}\Big(\frac{\varepsilon_n^k(V_j)}{f(V_j)}\Big)^2 a(V_j) + o_P(1) \\ &\xrightarrow{P} h^{-(d_X+d_Z)/2}B_L + o_P(1). \end{aligned}$$

The last convergence holds by assumption if $\lambda_n = O(nh^{(d_X+d_Z)/2}$. In particular for any fixed alternative, the convergence does not apply and $nh^{(d_X+d_Z)/2}\widehat{\Gamma}_{\mathcal{K}2} = O(n)$ and diverges. This yields consistency of the test statistic. □

Proof of Theorem 2.3

In the proof of this theorem we use the notation $\mathbf{E}^*$ and $\mathbf{var}^*$ to denote expectation and variance conditional on the data. Decompose

$$\begin{aligned} \widehat{\Gamma}^*_{\mathcal{K}} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(X_i-X_j, Z_i-Z_j)\frac{Y_j^{k,*}-\widehat{m}_{\widetilde{h}}^{k,*}(X_i)}{\widehat{f}_h(X_j,Z_j)}\Big)^2 A_i \\ &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n} K_h(X_i-X_j, Z_i-Z_j) \\ &\qquad \times \Big(\frac{\varepsilon^{k,*}}{\widehat{f}_h(X_j,Z_j)} + \frac{\widehat{m}_{\widetilde{h}}^{k}(X_i)-\widehat{m}_{\widetilde{h}}^{k,*}(X_i)}{\widehat{f}_h(X_j,Z_j)}\Big)\Big)^2 A_i \\ &= (I^*_{\mathcal{K}n} + \Delta^*_{\mathcal{K}n})(1+o_P(1)) + \Gamma^*_{\mathcal{K}3} + \Gamma^*_{\mathcal{K}4}, \end{aligned}$$

where $I^*_{\mathcal{K}n}$ and $\Delta^*_{\mathcal{K}n}$ are defined as in (2.10) and (2.11) by replacing $Y_j^k - \mu^k(X_j)$ with $\varepsilon_j^{k,*}$. $\Gamma^*_{\mathcal{K}3}$ can be bounded by showing that

$$\sup_{x\in\mathcal{A}} |\widehat{m}_{\widetilde{h}}^{k}(x) - \widehat{m}_{\widetilde{h}}^{k,*}(x)| = O_P\Big(\widetilde{h}^r + \Big(\frac{\log n}{nh^{d_X}}\Big)^{1/2}\Big). \tag{2.19}$$

Decomposing $I^*_{\mathcal{K}n}$ as in equation (2.12) into $I^*_{\mathcal{K}n,1}$ and $I^*_{\mathcal{K}n,2}$ it remains to show that

$$nh^{(d_X+d_Z)/2} I^*_{\mathcal{K}n,1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma^2_{\mathcal{K}}), \tag{2.20}$$

conditional on the data with probability tending to one and

$$nh^{(d_X+d_Z)/2} I^*_{\mathcal{K}n,2} - h^{-(d_X+d_Z)/2}B_{\mathcal{K}} \xrightarrow{P} 0 \tag{2.21}$$

$$nh^{(d_X+d_Z)/2}\Delta^*_{\mathcal{K}n} \xrightarrow{P} 0. \tag{2.22}$$

Then the statement of the theorem follows.

Proof of (2.19) First note that

$$\begin{aligned}\sup_{x\in\mathcal{A}} |\widehat{m}^k_{\widetilde{h}}(x) - \widehat{m}^{k,*}_{\widetilde{h}}(x)| &= \sup_{x\in\mathcal{A}} |(\widehat{f}^k_{\widetilde{h}}(x))^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)(Y_i^k - Y_i^{k,*})| \\ &\le \sup_{x\in\mathcal{A}} |m^k(x) - \widehat{m}^k_{\widetilde{h}}(x)| + \sup_{x\in\mathcal{A}} |(\widehat{f}^k_{\widetilde{h}}(x))^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\varepsilon_i^{k,*}| \\ &\qquad + \sup_{x\in\mathcal{A}} |(\widehat{f}^k_{\widetilde{h}}(x))^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\varepsilon_i^{k}|.\end{aligned}$$

The first term already has the desired rate. Because $\widehat{f}^k_{\widetilde{h}}(x)$ is consistent and $f(x)$ is bounded from below on $\mathcal{A}$ further analysis can be restricted to the numerator. Since the analysis of the second and the third term in analogous, we concentrate on the second term. First, a truncation argument is applied. Define $\widetilde{\varepsilon}_i^{k,*} = \mathbf{1}_{\{\varepsilon_i^{k,*}\le n\widetilde{h}^{d_X}\}}$, which allows to decompose

$$\begin{aligned}(2.23)\quad \frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\varepsilon_i^{k,*} &= \frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\widetilde{\varepsilon}_i^{k,*} \\ &\qquad + \frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\varepsilon_i^{k,*}\mathbf{1}_{\{\varepsilon_i^{k,*}> n\widetilde{h}^{d_X}\}}.\end{aligned}$$

Starting with the second term, note that it holds that $\mathbf{E}\,|\varepsilon_i^{k,*}\mathbf{1}_{\{\varepsilon_i^{k,*}> n\widetilde{h}^{d_X}\}}| = O(n^{-2}\widetilde{h}^{-2d_X})$, because the forth moment of $\varepsilon_i^{k,*}$ is finite. Then, the second term on the right side of (2.23) can be bounded with Markov's inequality with first moments

$$\begin{aligned}\mathbf{E}|\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\varepsilon_i^{k,*}\mathbf{1}_{\{\varepsilon_i^{k,*}> n\widetilde{h}^{d_X}\}}| &\le \mathbf{E}\,|K_{\widetilde{h}}(x-X_1)\varepsilon_1^{k,*}\mathbf{1}_{\{\varepsilon_1^{k,*}> n\widetilde{h}^{d_X}\}}| \\ &\le \sup_u |K(u)|\,\mathbf{E}\,|\varepsilon_i^{k,*}\mathbf{1}_{\{\varepsilon_i^{k,*}> n\widetilde{h}^{d_X}\}}|(1+O(\widetilde{h})) \\ &= O(n^{-2}\widetilde{h}^{-2d_X}),\end{aligned}$$

from which the desired rate follows.
Finally, we turn to the first term in (2.23). Covering the compact set $\mathcal{A}$ with N cubes $\mathcal{A}_l = \{x \mid \|x - x_l\| < \eta_N\}, l = 1, \dots, N, \eta_N = O(N^{-1/d_X})$ we write

$$\begin{aligned}(2.24)\quad \sup_{x\in\mathcal{A}}|\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x-X_i)\widetilde{\varepsilon}_i^{k,*}| &\le \max_l|\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x_l-X_i)\widetilde{\varepsilon}_i^{k,*}| \\ &\qquad + \max_l \sup_{x\in\mathcal{A}_l}|\frac{1}{n}\sum_{i=1}^{n}(K_{\widetilde{h}}(x-X_i) - K_{\widetilde{h}}(x_l-X_i))\widetilde{\varepsilon}_i^{k,*}|.\end{aligned}$$

Using the Lipschitz-continuity of the kernel, one directly obtains that the second term on the right hand side is of $O_P(\eta_N \widetilde{h}^{-d_X/2} n^{-1/2}) = o_P(n^{-1/2}\widetilde{h}^{d_X/2}(\log n)^{1/2})$. The first term on the is bounded using Bonferroni's inequality first and then Bernstein's inequality

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x_l - X_i)\widetilde{\varepsilon}_i^{k,*}\Big| > \Big(\frac{\log n}{n\widetilde{h}^{d_X}}\Big)^{1/2}\frac{c_1}{2}\Big)$$

$$\leq 2\exp\Big(-\frac{c_1^2(\log n)/(4n\widetilde{h}^{d_X})}{4\sum_{i=1}^{n}\mathbf{E}(\frac{1}{n}K_{\widetilde{h}}(x - X_i)\widetilde{\varepsilon}_i^{k,*})^2 + c_2(\frac{\log n}{n^3\widetilde{h}^{3d_X}})^{3/2}}\Big),$$

where c_2 is the constant arising from Cramer's conditions on the distribution of $\widetilde{\varepsilon}_i^{k,*}$. It follows from standard arguments that $\sum_{i=1}^{n}\mathbf{E}(\frac{1}{n}(K_{\widetilde{h}}(x - X_i)\widetilde{\varepsilon}_i^{k,*}$ and so we get that

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} K_{\widetilde{h}}(x - X_i)\widetilde{\varepsilon}_i^{k,*}\Big| > \Big(\frac{\log n}{n\widetilde{h}^{d_X}}\Big)^{1/2}\frac{c}{2}\Big) \leq O(n^{-1}).$$

Then, for $N = o(n)$ the desired rate of convergence is obtained.

Proof of (2.20) To derive the asymptotic distribution of

$$I^*_{\mathcal{K}n,1} = \sum_{i<j} h_n(W_i^*, W_j^*),$$

given the data with probability tending to one, again Lemma 3.1 will be applied. This is done by showing that the conditions hold with probability tending to one, i.e.

$$\frac{\max_{1\leq i\leq n}\sum_{j=1}^{n}\mathbf{E}^*\, h_n(W_i^*, W_j^*)^2}{\mathbf{var}^*\, I^*_{\mathcal{K}n1}} \xrightarrow{P} 0 \quad \text{and} \quad \frac{E^*(I^*_{\mathcal{K}n,1})^4}{(\mathbf{var}^*\, I^*_{\mathcal{K}n,1})^2} \xrightarrow{P} 3.$$

Here,

$$h_n(W_i^*, W_j^*) = \frac{2}{n^2}\int K_h(v - V_i)K_h(v - V_j)\frac{a(v)}{f(v)}\,\mathrm{d}v \sum_{k=1}^{d_Y}\varepsilon_i^{k,*}\varepsilon_j^{k,*}.$$

First we analyze

$$\mathbf{E}^* h_n(W_i^*, W_j^*)^2 = \frac{4}{n^4}\Big(\int K_h(v-V_i)K_h(v-V_j)\frac{a(v)}{f(v)}\,\mathrm{d}v\Big)^2 \times \Big(\sum_{k=1}^{d_Y}(Y_i^k - \widehat{m}_{\widetilde{h}}^k(X_i))(Y_j^k - \widehat{m}_{\widetilde{h}}^k(X_j))\Big)^2$$

(2.25)

$$= \frac{4}{n^4}\Big(\int K_h(v-V_i)K_h(v-V_j)\frac{a(v)}{f(v)}\,\mathrm{d}v\Big)^2 \times \Big(\sum_{k=1}^{d_Y}(Y_i^k - \mu^k(V_i))(Y_j^k - \mu^k(V_j))\Big)^2 (1 + O_P\Big(\widetilde{h}^r + \Big(\frac{\log n}{n\widetilde{h}^{d_X}}\Big)^{1/2}\Big))$$
$$= h_n(W_i, W_j)^2 + o_p(n^{-4}h^{-2(d_X+d_z)}).$$

This holds because under H_0 we have that $m^k(X_i) = \mu^k(X_i, Z_i)$ almost surely. Starting with the numerator, we utilize the conditional independence of the bootstrap residuals to see that

$$\mathbf{var}^* I_{\mathcal{K}n,1} = \sum_{i<j} \mathbf{E}^* h_n(W_i^*, W_j^*).$$

To bound this in probability, apply Markov's inequality with the first moment

$$\mathbf{E}\Big|\sum_{i<j}\mathbf{E}^* h_n(W_i^*, W_j^*)^2\Big| = \sum_{i<j}\mathbf{E}\, h_n(W_i^*, W_j^*)^2 = n^{-2}h^{-(d_X+d_Z)}2\Sigma_{\mathcal{K}}^2(1+o(1)),$$

from which it follows that

$$\mathbf{var}^* I_{\mathcal{K}n,1} \xrightarrow{P} \mathbf{var}\, I_{\mathcal{K}n,1}.$$

This is now used to show the first condition. Together with (2.16) and (2.25) we obtain

$$\frac{\max_{i=1,\dots,n}\sum_{j=1,j\neq i}^n \mathbf{E}^* h_n(W_i^*, W_j^*)^2}{\mathbf{var}\, I_{\mathcal{K}n,1}}$$
$$= \frac{\max_{i=1,\dots,n}\sum_{j=1,j\neq i}^n h_n(W_i, W_j)^2 + O_P(n^{-3}(\widetilde{h}^r + (\log n/(n\widetilde{h}^{d_X}))^{1/2}))}{\mathbf{var}\, I_{\mathcal{K}n,1}}$$
$$= \frac{\max_{i=1,\dots,n}\sum_{j=1,j\neq i}^n h_n(W_i, W_j)^2}{\mathbf{var}\, I_{\mathcal{K}n,1}} + O_P(n^{-1}(\widetilde{h}^r + (\log n/(n\widetilde{h}^{d_X}))^{1/2}))$$
$$= \quad o_P(1).$$

For the second condition we again use the convergence of the denominator. Then using the first moment to bound the probability leads to similar calculations as done in the proof of equation (2.13).

Proof of (2.21) The proof of equation (2.21) consists of using iterated expectations and use there the same calculations as to proof equation (2.20).

Proof of (2.22) As $\mathbf{E}^* \varepsilon_j^{k,*} = 0$ the same arguments as for $\Delta_{\mathcal{K}n}$ remain to hold for $\Delta^*_{\mathcal{K}n}$. □

Proof of Theorem 2.4

From Masry (1996) it is known that

$$\sup_{v\in\mathcal{A}} |\widehat{S}(v) - f(v)M| = O_P\Big(h^2 + \Big(\frac{\log n}{nh^{d_X+d_Y}}\Big)^{1/2}\Big),$$

therefore we can write

$$\begin{aligned}\widehat{\Gamma}^L_{\mathcal{K}} &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}(\lfloor f(V_i)^{-1}M^{-1}(\widehat{T}^k(V_i) - \widetilde{T}^k(V_i))\rfloor_1)^2 A_i(1+o_P(1))\\ &= \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{0\le|\mathbf{j}|\le p}\kappa_{\mathbf{j}}^{-1}\sum_{l=1}^{n}\Big(\frac{V_l-V_i}{h}\Big)^{\mathbf{j}}K_h(V_l-V_i)\frac{Y_l^k - \widehat{m}^{k,L}_{\widetilde{h}}(X_l)}{f(V_i)}\Big)^2\\ &= \widehat{\Gamma}^L_{\mathcal{K}1} + \widehat{\Gamma}^L_{\mathcal{K}2} + \widehat{\Gamma}^L_{\mathcal{K}3} + \widehat{\Gamma}^L_{\mathcal{K}4},\end{aligned}$$

where we decompose according to $Y_l^k - \widehat{m}^{k,L}_{\widetilde{h}}(X_l) = Y_l^k - \mu^k(V_l) + \mu^k(V_l) - m^k(V_l) + m^k(X_l) - \widehat{m}^{k,L}_{\widetilde{h}}(X_l)$ and transfer all cross terms to $\widehat{\Gamma}^L_{\mathcal{K}4}$. Then $\widehat{\Gamma}^L_{\mathcal{K}2} = 0$ almost surely under H_0. And $\widehat{\Gamma}^L_{\mathcal{K}3} = O_P(h^{2r} + \log n/(nh^{d_X}))$ by applying results from Masry (1996) for the density estimator and the local linear estimator. Next, we decompose

$$\widehat{\Gamma}^L_{\mathcal{K}1} = I^L_{\mathcal{K}n,1} + I^L_{\mathcal{K}n,2} + \Delta^L_{\mathcal{K}n},$$

where the quantities are given as in (2.10), (2.11) and (2.12) and the kernel is replaced by

$$\widetilde{K}_h(u) = \sum_{1\le\mathbf{j}\le p}\Big(\frac{u}{h}\Big)^{\mathbf{j}}\kappa_{\mathbf{j}}^{-1}K_h(u).$$

Since this kernel satisfies the assumptions which are necessary to show (2.13)–(2.15) (note that higher order properties of the kernel are not used there), the statement of the theorem follows. □

Proof of Theorem 2.5

In this case we can decompose the test statistic into

$$\widehat{\Gamma}^S_{\mathcal{K}} = \widehat{\Gamma}^S_{\mathcal{K}1} + \widehat{\Gamma}^S_{\mathcal{K}2} + \widehat{\Gamma}^S_{\mathcal{K}3} + \widehat{\Gamma}^S_{\mathcal{K}4} + \widehat{\Gamma}^S_{\mathcal{K}5},$$

where $\widehat{\Gamma}^S_{\mathcal{K}1} = \widehat{\Gamma}_{\mathcal{K}1}$,

$$\widehat{\Gamma}^S_{\mathcal{K}2} = \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}\frac{K_h(X_i - X_j, Z_i - Z_j)}{\widehat{f}_h(X_i, Z_i)} \\ \times (\mu^k(X_j, Z_j) - m^k(X_j) - G^k(Z_j, \theta))\Big)^2 A_i$$

$$\widehat{\Gamma}^S_{\mathcal{K}3} = \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}\frac{K_h(X_i - X_j, Z_i - Z_j)}{\widehat{f}_h(X_i, Z_i)}(m^k(X_j) - \widehat{m}^k_{\widetilde{h}}(X_j))\Big)^2 A_i$$

$$\widehat{\Gamma}^S_{\mathcal{K}4} = \frac{1}{n}\sum_{k=1}^{d_Y}\sum_{i=1}^{n}\Big(\frac{1}{n}\sum_{j=1}^{n}\frac{K_h(X_i - X_j, Z_i - Z_j)}{\widehat{f}_h(X_i, Z_i)}(G^k(Z_j, \theta) - G^k(Z_j, \widehat{\theta})\Big)^2 A_i$$

and $\widehat{\Gamma}^S_{\mathcal{K}5}$ contains all cross terms. Under H_{0S} we have that $\widehat{\Gamma}^S_{\mathcal{K}2} = 0$ almost surely and the other two terms are bounded by uniform convergence rates. Start with

$$\begin{aligned} |\widehat{\Gamma}^S_{\mathcal{K}3}| &\leq \max_{k=1,\dots,d_Y}\sup_{x\in\mathcal{A}_X}|m^k(x) - \widehat{m}^k_{\widetilde{h}}(x)|^2 \sup_{(x,z)\in\mathcal{A}}|a(x,z)| \\ &\leq \sup_{(x,z)\in\mathcal{A}}|a(x,z)|\Big(\max_{k=1,\dots,d_Y}\sup_{x\in\mathcal{A}_X}|m^k(x) - \widetilde{m}^k_{\widetilde{h}}(x)|^2 \\ &\quad + \max_{k=1,\dots,d_Y}\sup_{z\in\mathcal{A}_Z}|G^k(z,\theta) - G^k(z,\widehat{\theta})|^2\Big) \\ &= o_P(n^{-1}h^{(d_X+d_Z)/2}). \end{aligned}$$

The quantity $\widetilde{m}^k_{\widetilde{h}}(X_j)$ denotes a nonparametric regression of the unobserved variable $Y_i - G(Z_i, \theta)$ on X_i. The standard uniform convergence rate holds for this estimator and by our assumptions it converges faster than the test statistic. For the parametric function $G(z, \theta)$ the convergence rate was assumed. From this assumption we also obtain

$$|\widehat{\Gamma}^S_{\mathcal{K}4}| = o_P(n^{-1}h^{(d_X+d_Z)/2}).$$

The asymptotic distribution of $\widehat{\Gamma}_{\mathcal{K}1}$ was derived in the proof of Theorem 2.1. □

Proof of Theorem 2.6

For dependent data, decomposition (2.9) still applies and under H_0 it holds that $\widehat{\Gamma}_{\mathcal{K}2} = 0$. Because β-mixing implies α-mixing, the results in Masry (1996) hold under Assumption 2.7. This means, we have that

$$\sup_{x\in\mathcal{A}}|m^k(x) - \widehat{m}_{\widetilde{h}}(x)| = O_P(\widetilde{h}^r + \Big(\frac{\log n}{n\widetilde{h}^{d_X}}\Big)^{1/2})$$

and the same rate holds for the kernel density estimator. Therefore it remains to analyze $\widehat{\Gamma}_{\mathcal{K}1}$ and to show (2.13)–(2.15) for the dependent case.

Proof of (2.13) To derive the asymptotic distribution, we regard $I_{\mathcal{K}n,1}$ still as U-Statistic, and apply Theorem 2.1 in Fan and Li (1999). To apply this central limit theorem a large number of assumptions have to be checked. For brevity we concentrate on those that influence the rates. Denote with $(\widetilde{W_i})_{i=1,\dots,n}$ a sequence of independent and identically distributed random variables with the same marginal distribution as $(W_i)_{i=1,\dots,n}$

$$\begin{aligned} \frac{m}{n^{3/2}} \frac{\mathbf{E}\, h_n(\widetilde{W}_1, \widetilde{W}_2)^4}{(\mathbf{E}\, h_n(\widetilde{W}_1, \widetilde{W}_2)^2)^2} &= O\Big(\frac{m}{n^{3/2}h}\Big) \\ m^4 \frac{\max_{t>1} \mathbf{E}(\int h_n(w, W_1) h_n(w, W_t) f(w)\, \mathrm{d}w)^2}{(\mathbf{E}\, h_n(\widetilde{W}_1, \widetilde{W}_2)^2)^2} &= O(m^4 h) \\ n^2 \beta(m)^{1-1/\nu} \frac{m^2 + n^2 \beta(m)^{1-1/\nu}}{(\mathbf{E}\, h_n(\widetilde{W}_1, \widetilde{W}_2)^2)^2} &= O\big(n^6 h^2 (m^2 \beta(m)^{1-1/\nu} + n^2 \beta(m)^{2-2/\nu})\big) \end{aligned}$$

Together with the assumptions on the number of existing moments of Y and the kernel function ($\mathbf{E}\, Y^{4\nu} < \infty, \kappa_0^{4\nu}$), this yields (2.13).

Proof of (2.14) It is easy to see that $\mathbf{E}\, I_{\mathcal{K}n,2}$ is unchanged. To show convergence in probability using the second moment of $I_{\mathcal{K}n,2}$, the covariances have to be bounded. Writing

$$I_{\mathcal{K}n,2} = \sum_{i=1}^{n} h'_n(X_i),$$

with

$$h'_n(W_i) = \frac{1}{nh^{-(d_X+d_Z)}} \sum_{k=1}^{d_Y} \int \Big(K(u,v)(Y_1^k - m^k(X_1)) \Big)^2 \frac{a(X_1+uh, Z_1+vh)}{f(X_1+uh, Z_1+vh)}\, \mathrm{d}u\, \mathrm{d}v.$$

We then use the covariance inequality for strongly dependent processes ($\nu > 1$)

$$\mathbf{cov}(h'_n(W_i), h'_n(W_j)) \le c\big(\mathbf{E}(h'_n(W_1))^\nu\big)^{2/\nu} \beta(j-i)^{1-2/\nu}.$$

As $\big(\mathbf{E}(h'_n(W_1))^\nu\big)^{2/\nu} = O(n^{-2}h^{-2(d_X+d_Z)})$ (if $\mathbf{E}\, Y^{2\nu} < \infty$ and $\kappa_0^{2\nu} < \infty$) the convergence follows if $\sum_{i=1}^\infty \beta(i)^{1-2/\nu} < \infty$.

Proof of (2.15) To show that the expected value converges we use

$$|\,\mathbf{E}\, \gamma_n(W_i, W_j, W_k)| \le 4M^{1/\nu} \beta(\min\{i-k, j-k\})^{1-1/\nu},$$

where $M = \max\{\mathbf{E}\, \widetilde{\gamma}_n(W_i, W_j, W_k)^\nu, \mathbf{E} \int \widetilde{\gamma}_n(W_i, W_j, w)^\nu f(w)\, \mathrm{d}w\}$. (Lemma A.1 in Dette and Spreckelsen, 2004).

Convergence in probability is shown by using the first absolute moment and Lemma A.0 in Fan and Li (1999) to obtain

$$\mathbf{E}\,|\gamma_n(W_i, W_j, W_k)| \leq 4M^{1/\nu}\beta(\min\{i-k, j-k\})^{1-1/\nu},$$

with M as above. Some tedious calculations show that convergence in probability is established if $\mathbf{E}\,Y^{2\nu} < \infty$ and $\sum_{i=1}^{\infty}\beta(i)^{1-1/\nu} < \infty$.
This establishes the asymptotic distribution. □

Chapter 3

Testing Slutsky Symmetry in Nonparametric Demand Systems

3.1 Introduction

This chapter extracts parts of an article by Haag, Hoderlein and Pendakur (2005) which is devoted to testing and imposing rationality restrictions to consumer demand systems. In particular a nonparametric test for Slutsky symmetry is proposed. The restriction of symmetry is a set of nonlinear restrictions on the functions and derivatives of the expenditure share vector function. Using kernel regression techniques to estimate the unknown demand function, the test will be dominated by the estimators of the derivatives. This is the case because the estimated derivatives converge slower than the estimated functions themselves. Based on this insight, we provide a new test of symmetry, its asymptotic distribution and guidance on bootstrap simulation of its sampling distribution. A closely related test for symmetry in nonparametric demand systems was proposed in Lewbel (1995). The test proposed in this chapter invokes much weaker smoothness assumptions.

Beside the testing procedure, Haag Hoderlein and Pendakur (2005) propose a nonparametric estimation method that imposes Symmetry on the demand function and derive its asymptotic properties. Furthermore, both methods are applied to Canadian household data and their usage in empirical work is highlighted.

In this chapter the presentation is restricted to the test for symmetry. In Section 3.2 the test statistic is motivated and the asymptotic results are stated. In the third section a bootstrap implementation is provided and its validity is derived formally. To reduce technical details in the outline, all proofs and necessary assumptions are deferred to the appendix.

3.2 The Test Statistic and the Asymptotic Distribution

Define the cost function $C(p, u)$ to give the minimum cost to attain utility level u facing the M-vector of log-prices $p = (p^1, \dots, p^M)'$. Similarly, define the Marshallian demand function by $m(p, x)$, where x denotes log total expenditure, and let $z = (p', x)'$. The Slutsky matrix $S(z) = (s^{jk}(z))_{1 \le j \le M-1, 1 \le k \le M-1}$ is defined as the Hessian of the cost function with respect to (unlogged) prices. The elements may be expressed in terms of log-price and log-expenditure (rather than utility) as

$$s^{jk}(z) = \partial_k m^j(z) + m^k(z)\partial_x m^j(z) + m^j(z)m^k(z) + \delta^j(k)m^k(z)$$

where $\delta^j(k)$ denotes the Kronecker function to indicate a diagonal element and $\partial_x = \frac{\partial}{\partial x}$ and $\partial_k = \frac{\partial}{\partial p^k}$ are used for abbreviation (Mas-Colell, Whinston and Green, 1995). If the Slutsky matrix is continuously symmetric over a region of the z space, then Young's Theorem guarantees the existence of a cost function whose derivatives could produce the observed demand system over this region (see, e. g., Mas-Colell, Whinston and Green, 1995). The aim is to test whether the Slutsky matrix is symmetric.

In this setting, symmetry will be tested without testing homogeneity (which is also required for rationality). Other articles test homogeneity, see especially Kim and Tripathi (2003). For a test of homogeneity see also chapter 2.

Given homogeneity, symmetry is neccessary and sufficient for the existence of a cost function which could rationalise demands. However, marshallian demands could satisfy homogeneity without satisfying symmetry. In this case, although consumers do not suffer from money illusion, their demands cannot be rationalised by a cost function.

In the data, we are given observations on the $2M + 1$-dimensional random vector $Y = (W, Z)'$ where $W \in \mathbb{R}^M$ is an M-vector of expenditure shares and $Z = (P^1, \dots, P^M, X)'$ is the vector of log-prices $P = (P^1, \dots, P^M)'$ and household log-expenditure X. We define the regression function

$$m(z) = (m^1(z), \dots, m^M(z))' = \mathbf{E}(W \mid Z = z).$$

The fact that we have identified the mean regression function with an individual's demand function is not entirely innocuous. This problem is treated in detail in Lewbel (2001) and Hoderlein (2004), who establish that for $m(\cdot)$ to inherit symmetry certain (untestable) assumptions about the heterogeneity of individual preferences have to be fulfilled. One trivial case where this identification is valid is when, conditional on observables, there is no preference heterogeneity, and the difference $W - \mathbf{E}(W \mid Z = z)$ stems for example from orthogonal measurement error.

Under the assumption that $m(\cdot)$ indeed inherits rationality properties, we propose a test using the L_2-distance of those elements of the Slutsky matrix which are the same under symmetry. That is, we integrate and add up the squared distance between $S(z)$ and $S(z)'$. Here, the null hypothesis is

$$H_0 \colon \mathbb{P}(s^{jk}(Z) = s^{kj}(Z), \forall\, j \neq k) = 1$$

and the alternative is that there is at least one pair (j,k) with $s^{jk}(z) \neq s^{kj}(z)$ over a significant range. We may express the alternative as

$$H_1 \colon \mathbb{P}(s^{jk}(Z) = s^{kj}(Z), \forall\, j \neq k) < 1.$$

The null hypothesis is equivalent to the condition that the L_2-distance of these functions is zero. Using a positive and bounded weighting function $a(z)$ this can be written as

$$\Gamma_S = \mathbf{E}\Big(\sum_{j<k} (s^{jk}(Z) - s^{kj}(Z))^2 a(Z)\Big) = 0.$$

A test statistic may be constructed by the analogy principle. Observing a sample of n independent and identically distributed random vectors $(W_1, Z_1), \dots, (W_n, Z_n)$ we replace the unknown functions $m^j(z)$ by their Nadaraya-Watson estimators $\widehat{m}_h^j(z) = \sum_i K_h(z - Z_i) Y_i^j / \sum_i K_h(z - Z_i)$, where $K(\cdot)$ is a $M+1$-variate kernel function and $K_h(u) = (\det H)^{-1/2} K(H^{-1/2} u)$ with a bandwidth matrix $H^{1/2}$. For simplicity of notation we assume that the bandwidth matrix is diagonal with identical bandwidth h in each direction and that the kernel is a product kernel with properties defined in detail in the appendix below. The derivatives of the estimator $\partial_k \widehat{m}_h^j(z)$ are used as estimators for the derivatives $\partial_k m_j(z)$. We then obtain

$$\begin{aligned}(3.1)\quad \widehat{\Gamma}_S = \frac{1}{n} \sum_{j=1}^{M-2} \sum_{k=j+1}^{M-1} \sum_{i=1}^{n} (\partial_k \widehat{m}_h^j(Z_i) + \widehat{m}_h^k(Z_i) \partial_x \widehat{m}_h^j(Z_i) \\ - \partial_j \widehat{m}_h^k(Z_i) - \widehat{m}_h^j(Z_i) \partial_x \widehat{m}_h^k(Z_i))^2 A_i\end{aligned}$$

where $A_i = a(Z_i)$. This test statistic is a nonlinear combination of the function and its derivatives. However, since the estimator of the derivative converges slower than the estimator of the function, the asymptotic distribution is dominated by the derivative estimator and the function can assumed to be known.

To define the expected value and variance of the test statistic we have to introduce the covariance matrix $(\sigma^{ij}(z))_{1\le i,j\le M-1} = \mathbf{E}((W-m(Z))(W-m(Z))' \mid Z=z)$ and need the following lengthy notation:

$$\sigma_S^{jkj'k'} = \sum_{\substack{C,C'\in\{j,k\}\\ D,D'\in\{j',k'\}}} (-1)^{|\{C,C',D,D'\}|} \int \sigma^{CD}(z)\sigma^{C'D'}(z)a(z)^2 K^{CC'DD'}(z)\,\mathrm{d}z$$

with

$$K^{jkj'k'}(z) = \iint K^k(z,v)K^j(z,v-u)\,\mathrm{d}v \cdot \int K^{k'}(z,w)K^{j'}(z,w-u)\,\mathrm{d}w\,\mathrm{d}u$$

and

$$K^j(z,v) = \frac{\partial K}{\partial v^j}(v) + m^j(z)\frac{\partial K}{\partial x}(v).$$

Similarly define

$$\begin{aligned} b_S^{jk} = \int \sigma^{jj}(z)a(z)\int (K^k(z,u))^2\,\mathrm{d}u\,\mathrm{d}z + \int \sigma^{kk}(z)a(z)\int (K^j(z,u))^2\,\mathrm{d}u\,\mathrm{d}z \\ -2\int \sigma^{jk}(z)a(z)\int K^j(z,u)K^k(z,u)\,\mathrm{d}u\,\mathrm{d}z. \end{aligned}$$

The asymptotic distribution is given in the following

Theorem 3.1. *Let the model be as defined above and let Assumptions 3.1–3.4 hold. Under H_0,*

$$\sigma_S^{-1}(nh^{(M+5)/2}\widehat{\Gamma}_S - h^{-(M+1)/2}B_S) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\sigma_S^2 = 2\big(\sum_{j<k}\sigma_S^{jkjk} + 2\sum_{j<k}\sum_{\substack{j'<k'\\ (j,k)<(j',k')^1}} \sigma_S^{jkj'k'}\big) \qquad B_S = \sum_{j<k} b_S^{jk}.$$

Simplifying the proofs in the appendix to one line, the test statistic can be written as

$$\widehat{\Gamma}_H = \Gamma_S + U_n + \Delta_n$$

[1] The ordering is in a lexicographic sense.

where Γ_S is 0 under H_0, Δ_n depends upon the uniform rate of convergence of the estimators and U_n is a degenerated U-statistic. This U-statistic converges at the rate $nh^{(M+5)/2}$, which might be faster than $n^{1/2}$ depending on the choice of the bandwidth sequence. Under the alternative Γ_S is a positive constant and after multiplying the test statistic with $nh^{(M+5)/2}$ this term tends to infinity. Therefore we obtain consistency of the test against alternatives with $\Gamma_S > 0$.

Another test for Slutsky symmetry based on kernel regression has been proposed by Lewbel (1995). This test procedure is based on the integrated conditional moment (ICM) test of Bierens (1982), which uses the fact that H_1 is equivalent to

$$\mathbf{E}(s^{jk}(Z) - s^{kj}(Z))w(\Xi' Z) \mid \Xi = \xi) \neq 0$$

for a set of ξ with nonzero Lebesgue-measure, where the weighting function has to be chosen appropriately (see Bierens and Ploberger, 1997). Lewbel (1995) uses a Kolmogorov-Smirnov-type test-statistic and derives asymptotic normality under stringent smoothness assumptions. Assuming that the unknown function and the density are $r+1$-times continuously differentiable (see assumption in the appendix), Lewbel requires $r > 2(M+1)$ whereas our test requires only $r > \frac{3}{4}(M+1)$. Although the smoothness class of an unknown function is difficult to establish in practice, this is a substantial relaxation of assumptions. Fan and Li (2000) discuss in detail the question under which circumstances the ICM-test or the kernel based test of our type has greater benefits. Their results should carry over to our situation.

In nonparametric regression analysis the advantages of local polynomial estimators over Nadaraya-Watson estimators are well known, especially in derivative estimation (see Fan and Gijbels, 1996). If we use higher order local polynomial estimators for m and its first partial derivatives, our results continue to hold when $K(\cdot)$ is replaced by its equivalent kernel. The rate of convergence of the test statistic remains the same, only some of the kernel constants arising in the bias and variance expression will change. In our application the advantages of the local polynomial based test is clear. For $M = 4$, we need $r > 3.75$. Using Nadaraya-Watson estimators, a kernel of order 4 has to be implemented. Using a local quadratic estimator, for example, the order only has to be 2. However, the smoothness assumptions remain unchanged.

3.3 Practical Implementation Using the Bootstrap

The direct way to implement the test is to estimate the expected value B_S and the variance σ_S^2. This requires the estimation of integrals like

$$\int \sigma^{jj}(z)a(z)\int (K^k(z,u))^2\,\mathrm{d}u\,\mathrm{d}z$$

or even more complex combinations in the variance parts. Therefore estimators of the conditional variances and covariances are needed. A Nadaraya-Watson-type estimator is given by

$$\widehat{\sigma}_h^{jj'}(z) = \frac{\sum_{i=1}^n K_h(z-Z_i)(W_i^j - \widehat{m}_h^j(Z_i))(W_i^{j'} - \widehat{m}_h^{j'}(Z_i))}{\sum_{i=1}^n K_h(z-Z_i)}$$

Given the large number of bias and variance components in Theorem 3.1, the asymptotic approach to the test is difficult to implement. Moreover, these asymptotic approximations can work very poorly in a finite sample.

To avoid the problems noted above, one might instead use a bootstrap procedure to derive critical values. To bootstrap the test statistic, note that the estimator of the derivative can be written as a weighted average

$$\partial_k \widehat{m}_h^j(z) = \sum_{i=1}^n \widetilde{V}_{ni}^k(z)W_i^j \tag{3.2}$$

where $\widetilde{V}_{ni}^k$, $i=1,\dots,n$ is a set of weights giving the kth price derivative of the jth expenditure share at z when applied to the data W_i^j. Using this in the definition of $\widehat{\Gamma}_S$ we obtain

$$\widehat{\Gamma}_S = \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=j+1}^{M-1}\sum_{i=1}^{n}\Big(\sum_{l=1}^{n} V_{nl}^k(Z_i)W_l^j - V_{nl}^j(Z_i)W_l^k\Big)^2 A_i$$

with $V_{nl}^k(Z_i) = \widetilde{V}_{nl}^k(Z_i) + \widehat{m}_h^k(Z_i)\widetilde{V}_{nl}^x(Z_i)$.

Next we exploit the fact that the estimator of the function converges faster than the estimator of the derivative. Plugging in $W_l^j = m^j(Z_l) + \varepsilon_l^j$ and noting that for large n it holds under the assumption of symmetry that

$$\begin{aligned}(3.3)\quad &\sum_{l=1}^n V_{nl}^{jk}(Z_i)m^j(Z_l) - V_{nl}^{kj}(Z_i)m^k(Z_l) \approx \\ &\qquad \partial_k m^j(Z_i) + m^k(Z_i)\partial_x m^j(Z_i) - \partial_j m^k(Z_i) - m^j(Z_i)\partial_x m^k(Z_i) = 0.\end{aligned}$$

Therefore the test statistic can under H_0 be approximated by

$$\widehat{\Gamma}_S \approx \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=j+1}^{M-1}\sum_{i=1}^{n}\Big(\sum_{l=1}^{n} V_{il}^{jk}\varepsilon_l^j + V_{il}^{kj}\varepsilon_l^k\Big)^2 A_i.$$

The bootstrap is based on this equation and is described as follows

1. Construct (multivariate) residuals $\widehat{\varepsilon}_i = W_i - \widehat{m}_h(Z_i)$.
2. For each i randomly draw $\varepsilon_i^* = (\varepsilon_i^{1*}, \ldots, \varepsilon_i^{M-1,*})'$ from a distribution $\widehat{F}_i$ that mimics the first three moments of $\widehat{\varepsilon}_i$.
3. Calculate $\widehat{\Gamma}_S^*$ from the bootstrap sample $(\varepsilon_i^*, Z_i), i = 1, \ldots, n$ by
$$\widehat{\Gamma}_S^* = \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=j+1}^{M-1}\sum_{i=1}^{n}\Big(\sum_{\substack{l=1\\ l\neq i}}^{n} V_{il}^{jk}\varepsilon_l^{j*} - V_{il}^{kj}\varepsilon_l^{k*}\Big)^2 A_i.$$
4. Repeat this often enough to obtain critical values.

To approximate the distribution by the bootstrap, usually the restriction of the null hypothesis is imposed in the construction of the residuals. Because symmetry implies a complicated restriction to the demand function and its derivatives, this is not directly possible. Therefore the restriction is imposed in the construction of the test statistic by using equation (3.3).

The theoretical result concerning the bootstrap procedure is given in the following

Theorem 3.2. *Let the model be as defined above and let Assumptions 3.1–3.5 hold. Under H_0, conditional on the data $(W_i, Z_i), i = 1, \ldots, n$ it hold that*

$$\sigma_S^{-1}(nh^{(M+5)/2}\widehat{\Gamma}_S^* - h^{-(M+5)/2}B_S) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

with probability tending to one.

The set of admissible distributions $\widehat{F}_i$ is very general. One may use a wild bootstrap (Härdle and Mammen, 1993) or smooth conditional moment bootstrap (Gozalo, 1997) suitably modified to account for cross-equation correlations (see Haag, Hoderlein and Pendakur, 2005) for details.

To prove the asymptotic result of Theorem 3.2 it is sufficient to assume that the bootstrap distribution $\widehat{F}_i$ mimics the first two moments of $\widehat{\varepsilon}_i$. Matching the first three moments as suggested in the algorithm we propose could yield higher order approximations of the Edgeworth expansion of the test statistic. Although we do not consider such expansions, we believe that this should improve the finite sample properties and use therefore three moments in the application.

Appendix

General Assumptions

Assumption 3.1. *The data* $Y_i = (W_i, Z_i), i = 1, \dots, n$ *are independent and identically distributed with density* $f(y)$*.*

This assumption can be relaxed to dependent data. All proofs can be extended to α-mixing processes in the case of estimation and β-mixing processes in the case of testing. Changes in the proofs are briefly discussed below. The validity of the bootstrap is not affected by dependent data, if we assume that $\mathbf{E}(\varepsilon_t \mid Y_{t-1}, \dots, Y_1) = 0$. Then, the bootstrap works because the residuals are mean independent and we only use the residuals for resampling (see also Kreiß, Neumann and Yao, 2002).

Assumption 3.2. *For the data generating process*

1. $f(y)$ *is* $r+1$*-times continuously differentiable (*$r \geq 2$*).* f *and its first partial derivatives are bounded and square-integrable.*

2. $m(z)$ *is* $r+1$*-times continuously differentiable.*

3. $f(z) = \int f(w, z)\,\mathrm{d}w$ *is bounded from below on the compact support* $\mathcal{A}$ *of* $a(z)$*, i. e.* $\inf_{z \in \mathcal{A}} f(z) = b > 0$.

4. *The covariance matrix*
$$\Sigma(z) = (\sigma^{ij}(z))_{1 \leq i,j \leq M-1} = \mathbf{E}((W - m(Z))(W - m(Z))' \mid Z = z)$$
is square-integrable (elementwise) on $\mathcal{A}$*.*

$\mathbf{E}((W^j - m^j(Z))^2 (W^k - m^k(Z))^2) < \infty$ *for every* $1 \leq j, k \leq M-1$*.*

Assumption 3.3. *For the kernel regression*
The kernel is a $M+1$*-dimensional function* $K\colon [-1,1]^{M+1} \to \mathbb{R}$*, symmetric around 0 with* $\int K(u)\,\mathrm{d}u = 1$*,* $\int |K(u)| du < \infty$ *and of order* r *(i. e.* $\int u^k K(u)\,\mathrm{d}u = 0$ *for all* $k < r$ *and* $\int u^r K(u)\,\mathrm{d}u < \infty$*). Further* $\|x\|^{M+1} K(x) \to 0$ *for* $\|x\| \to \infty$*.*

Our results should continue to hold for arbitrary kernel functions. See Ruppert and Wand (1994) for details about the implementation of general multivariate kernels. Further assumptions have to be made on the rate of convergence of the bandwidth sequence and the smoothness r of the unknown demand function $m(\cdot)$.

Assumption 3.4. *For the bandwidth sequence*

1. *For the order r of the kernel, we require*

$$r > \frac{3}{4}(M+1). \tag{3.4}$$

2. *For $n \to \infty$, the bandwidth sequence $h = O(n^{-1/\delta})$ satisfies*

$$2(M+1) < \delta < (M+1)/2 + 2r \tag{3.5}$$

The asymptotic distribution of the test statistic is derived under the above conditions on the bandwidth sequence. It is important to note that the optimal rate for estimation, given by

$$\delta_{opt} = (M+1) + 2r$$

is excluded. Here, a smaller bandwidth is needed to obtain the asymptotic distribution. In practice we calculate a data-driven bandwidth (by cross-validation) and adjust it by $n^{1/\delta_{opt}-1/\delta}$. Although we do not formally address the issue of data-driven bandwidths $\widehat{h}$ we assume that our results will hold if $\widehat{h}/h \xrightarrow{P} 1$.

Assumption 3.5. *For the bootstrap distribution*
The bootstrap residuals $\varepsilon_i^*, i = 1, \ldots, n$ are drawn independently from distributions $\widehat{F}_i$, such that $\mathbf{E}_{\widehat{F}_i}\, \varepsilon_i^* = 0, \mathbf{E}_{\widehat{F}_i}\, \varepsilon_i^*(\varepsilon_i^*)' = \widehat{\varepsilon}_i\widehat{\varepsilon}_i'$ and $\mathbf{E}_{\widehat{F}_i}(\varepsilon_i^{k,*})^4 < \infty$ for all $k = 1, \ldots, d_Y$.

Usually the bootstrap residuals are constructed by $\varepsilon_i^* = \eta_i^*\widehat{\varepsilon}_i$. Then, the assumption is fulfilled for discrete distributions, distributions with compact support and among others for the normal distribution, which are the most often used distributions for η_i^* in practice.

Proof of Theorems 3.1 and 3.2

The proof uses a functional expansion method applied to $\widehat{\Gamma}_S$ similar to the method in Aït-Sahalia Bickel and Stoker (2001). This leads to a von Mises expansion where the first order term is zero under H_0. The second order term is usually an infinite weighted sum of chi-squared distributed random variables. Here, a Feller-type condition is fulfilled which ensures the asymptotic negligibility of all summands. This condition is stated in the central limit theorem for degenerate U-statistics by de Jong (1987), which we use to derive the asymptotic normality for both theorems. The extension to β-mixing random variables follows by using Theorem 2.1 in Fan and Li (1999). Apart from this, the difference consists in tedious calculations of covariances where essentially a summability condition of the β-mixing coefficients is necessary.

Preliminary Lemmata

For probability measures the following notation is used: $\mathbb{P}^X$ is the measure of the distribution of X, $\mathbb{P}^{W|Z}$ is the measure of the conditional distribution of W given Z. And $\mathbf{E}^*$ denotes the conditional expectation of the bootstrap sample given the data. Marginal densities are defined by the list of the arguments and with a superscript indicating the element of w which is part of the argument. Kernel density estimators are defined in the same way.

Next define the seminorms

$$\|f^k\|_f = \max\{\sup_{z\in\mathcal{A}} |f(z)|,\ \sup_{z\in\mathcal{A}} |\int w^k f^k(w^k, z)\,\mathrm{d}w^k|\}$$

$$\|f^k\|_d = \max\{\max_{p=1,\dots,M+1} \sup_{z\in\mathcal{A}} |\partial_p f(z)|, \max_{p=1,\dots,M+1} \sup_{z\in\mathcal{A}} |\partial_p \int w^k f^k(w^k, z)\,\mathrm{d}w^k|\}$$

for density functions.

Lemma 3.1 (de Jong, 1987). *Let $Y_1, \dots, Y_n$ be a sequence of independent and identically distributed random variables. Suppose that the U-statistic $U_n = \sum_{1\le i<j\le n} h_n(Y_i, Y_j)$ with a symmetric function h_n is centered (i. e. $\mathbf{E}\, h_n(Y_1, Y_2) = 0$) and degenerate (i. e. $\mathbf{E}(h_n(Y_1, Y_2) \mid Y_1) = \mathbf{E}(h_n(Y_1, Y_2) \mid Y_2) = 0$, $\mathbb{P}$-a. s.). Then if*

$$\frac{\max_{1\le i\le n} \sum_{j=1, j\ne i} \mathbf{E}\, h_n(Y_i, Y_j)^2}{\mathbf{var}\, U_n} \longrightarrow 0 \qquad \text{and} \qquad \frac{\mathbf{E}\, U_n^4}{(\mathbf{var}\, U_n)^2} \longrightarrow 3$$

we have that

$$\frac{U_n}{\sqrt{\mathbf{var}\, U_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

Lemma 3.2. *Under the assumptions we have that for any $k = 1, \dots, M-1$*

$$\|\widehat{f}_h^k - f^k\|_d = O_P(h^{r-1} + (\log n/(nh^{M+3}))^{1/2})$$
$$\|\widehat{f}_h^k - f^k\|_f = O_P(h^r + (\log n/(nh^{M+1}))^{1/2}).$$

Proof. Noting that

$$(3.6)\quad \sup_{z\in\mathcal{A}} \Big| \int w^k(\widehat{f}_h^k(w^k,z) - f^k(w^k,z))\,\mathrm{d}w^k \Big| \\ \leq \|\widehat{f}_h^k(z)\|_\infty \|\widehat{m}_h^k(z) - m^k(z)\|_\infty + \|m^k(z)\|_\infty \|\widehat{f}_h^k(z) - f^k(z)\|_\infty$$

where $\|\cdot\|_\infty$ denotes the supremum-norm. Then the lemma follows from the uniform rates of convergence for density estimates, their derivatives, regression estimates and their derivatives. They can be found in Härdle (1990) or Masry (1996). □

Proof of Theorem 3.1

For simplification, denote $\partial_{M+1} = \partial_x$ and $\bar{K}_h(z) = \sum_{p=1}^{M+1} \partial_p K_h(z)$ which is of $O(h^{-(M+2)})$ because of the inner derivative. Start by expanding the statistic

$$\begin{aligned}
\widehat{\Gamma}_S =& \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=1}^{M-1}\sum_{i=1}^{n}(\partial_k \widehat{m}_h^j(Z_i) + m^k(Z_i)\partial_x \widehat{m}_h^j(Z_i) \\
&- \partial_j \widehat{m}_h^k(Z_i) - m^j(Z_i)\partial_x \widehat{m}_h^k(Z_i))^2 A_i \\
&+ \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=1}^{M-1}\sum_{i=1}^{n}((\widehat{m}_h^k(Z_i) - m^k(Z_i))\partial_x \widehat{m}_h^j(Z_i) \\
&- (\widehat{m}_h^j(Z_i) - m^j(Z_i))\partial_x \widehat{m}_h^k(Z_i))^2 A_i \\
&+ \frac{1}{n}\sum_{j=1}^{M-2}\sum_{k=1}^{M-1}\sum_{i=1}^{n}(\partial_k \widehat{m}_h^j(Z_i) + m^k(Z_i)\partial_x \widehat{m}_h^j(Z_i) - \partial_j \widehat{m}_h^k(Z_i) - m^j(Z_i)\partial_x \widehat{m}_h^k(Z_i)) \\
&\quad \times ((\widehat{m}_h^k(Z_i) - m^k(Z_i))\partial_x \widehat{m}_h^j(Z_i) - (\widehat{m}_h^j(Z_i) - m^j(Z_i))\partial_x \widehat{m}_h^k(Z_i))A_i
\end{aligned}$$

$$(3.7)\quad = \widehat{\Gamma}_{S1} + \widehat{\Gamma}_{S2} + \widehat{\Gamma}_{S3}$$

By Chebychev, $\widehat{\Gamma}_{S2} = O_p(h^{2r} + n^{-1}h^{-(M+1)}) = o_p(n^{-1}h^{-(M+5)/2})$ and an application of Cauchy-Schwarz shows that the third term is also of $o_p(n^{-1}h^{-(M+5)/2})$ provided that $\widehat{\Gamma}_{S1}$ has the limiting distribution of the theorem. So it is left to derive the asymptotic distribution of $\widehat{\Gamma}_{S1}$ has.
Start by looking at the theoretical version

$$\Gamma_{S1} = \sum_{j<k} \Gamma_{S1}^{jk}$$

For the beginning it suffices to investigate the case $j = 1, k = 2$ and to note that the other terms can be treated in the same way.

Consider Γ^{12}_{S1} as a functional of two $M+2$-dimensional density functions $f_1(w^1,z)$, $f_2(w^2,z)$, two $M+1$-dimensional functions $c_1(z)$ and $c_2(z)$ and a $M+1$-dimensional density $f_3(z)$:

$$\begin{aligned}\Gamma^{12}_{S1}(f_1,f_2,c_1,c_2,f_3) = \int (\partial_2 \frac{\int w^1 f_1(w^1,z)\,\mathrm{d}w^1}{f_1(z)} + c^2(z)\partial_{M+1}\frac{\int w^1 f_1(w^1,z)\,\mathrm{d}w^1}{f_1(z)} \\ - \partial_1 \frac{\int w^2 f_2(w^2,z)\,\mathrm{d}w^2}{f_2(z)} - c^1(z)\partial_{M+1}\frac{\int w^2 f_2(w^2,z)\,\mathrm{d}w^2}{f_2(z)})^2 a(z) f_3(z)\,\mathrm{d}z.\end{aligned}$$

Then the following functional expansion holds

Lemma 3.3. *Let $|g_1(w^1,z)|, |g_2(w^2,z)| < b/2$ be bounded functions and $\mathcal{G}^b$ $(\mathbb{R}^{M+2},\mathbb{R})$ the set of all such functions. Then under H_0 and Assumption 3.2 Γ^{12}_{S1} has an extension on $\mathcal{G}^b \times \mathcal{G}^b$ around (f^1,f^2) given by*

$$\begin{aligned}(3.8)\quad \Gamma^{12}_{S1}(f^1+g_1, f^2+g_2, m^1, m^2, f_3) = \Gamma^{12}_{S1}(f^1,f^2,m^1,m^2,f_3) \\ + \int \Bigg(\partial_2 \int \alpha^1(w,z) g_1(w^1,z)\,\mathrm{d}w + c_2(z)\partial_{M+1}\int \alpha^1(w,z)g_1(w^1,z)\,\mathrm{d}w^1 \\ - \partial_1 \int \alpha^2(w,z) g_2(w^2,z)\,\mathrm{d}w^2 - \partial_{M+1}\int \alpha^2(w,z) g_1(w^2,z)\,\mathrm{d}w^2\Bigg)^2 a(z) f_3(z)\,\mathrm{d}z \\ + O(\|g_1\|_d^2\|g_1\|_f + \|g_2\|_d^2\|g_2\|_f),\end{aligned}$$

with

$$\alpha^k(w,z) = \frac{w^k - m^k(z)}{f(z)}.$$

Let $\widehat{f}_e(\widetilde{p},x) = n^{-1}\sum_{i=1}^n \delta_{\{\widetilde{P}_i=\widetilde{p}, X_i=x\}}(\widetilde{p},x)$ denote the empirical distribution of the sampled data and extend the test statistic in the following way

$$\begin{aligned}\widehat{\Gamma}^{12}_{S1} &= \Gamma^{12}_{S1}(\widehat{f}^1_h, \widehat{f}^2_h, m^1, m^2, \widehat{f}_e) \\ &= \Gamma^{12}_{S1}(f^1 + \widehat{f}^1_h - f^1, f^2 + \widehat{f}^2_{\tilde{h}} - f^2, m^1, m^2, f) \\ &\quad + \Gamma^{12}_{S1}(f^1 + \widehat{f}^1_h - f^1, f^2 + \widehat{f}^2_{\tilde{h}} - f^2, m^1, m^2, \widehat{f}_e - f)\end{aligned}$$

Applying Lemma 3.3 to $g_i = \widehat{f}^i_h - f^i, i = 1,2$ allows to write

$$(3.9)\qquad = I^1_{Sn} + \Delta^1_{Sn} + O_p(\|\widehat{f}^1_h - f^1\|_d^2 \|\widehat{f}^1_h - f^1\|_f + \|\widehat{f}^2_h - f^2\|_d^2\|\widehat{f}^2_h - f^2\|_f)$$

using $\Gamma^{12}_{S1}(f^1,f^2,m^1,m^2,f) = 0$ under H_0 and where

$$\begin{aligned}I^{12}_{Sn} &= \int \Bigg(\sum_{i=1}^n r^{12}_{Sn}(W_i,Z_i;z)\Bigg)^2 a(z) f(z)\,\mathrm{d}z \\ \Delta^{12}_{Sn} &= \int \Bigg(\sum_{i=1}^n r^{12}_{Sn}(W_i,Z_i;z)\Bigg)^2 a(z)(\widehat{f}_e(z) - f(z))\,\mathrm{d}z\end{aligned}$$

with

$$
\begin{aligned}
r^{jk}_{S_n}(W_i,Z_i;z) &= \partial_k\alpha^j(W_i,z)K_h(z-Z_i) + m^k(z)\partial_{M+1}\alpha^j(W_i,z)K_h(z-Z_i)\\
&\quad - \partial_j\alpha^k(W_i,z)K_h(z-Z_i) - m^j(z)\partial_{M+1}\alpha^k(W_i^k,z)K_h(z-Z_i).
\end{aligned}
\tag{3.10}
$$

Here it has also been used that for all k it holds that

$$
\int \alpha^k(w,z)f^k(w,z)\,\mathrm{d}w^k = \int \frac{w^k}{f(z)}f^k(w,z)\,\mathrm{d}w^k - \int \frac{m^k(z)}{f(z)}f^k(w,z)\,\mathrm{d}w^k = 0.
\tag{3.11}
$$

The lower order terms in the extension (3.9) are bounded by Lemma 3.2 and the following

Lemma 3.4. *Under the assumptions we have that*

$$
\Delta^{12}_{Sn} = o_p(n^{-1}h^{-(M+5)/2}).
$$

Using the results on Γ^{ij}_{S1} for $1 \le j < k \le M-1$ allows to write the test statistic as

$$
\widehat{\Gamma}_{S1} = \sum_{j<k}\widehat{\Gamma}^{jk}_{S1} = \sum_{j<k} I^{jk}_{Sn} + o_p(n^{-1}h^{-(M+5)/2}).
\tag{3.12}
$$

Defining the centered random variables

$$
\widetilde{r}^{jk}_{S_n}(W_i,Z_i;z) = r^{jk}_{S_n}(W_i,Z_i;z) - \mathbf{E}\, r^{jk}_{S_n}(W_i,Z_i;z)
$$

the following decomposition applies

$$
\begin{aligned}
I_{Sn} &= \sum_{j<k} I^{jk}_{Sn}\\
&= \sum_{j<k}\int\left(\frac{1}{n}\sum_{i=1}^{n} r^{jk}_{Sn}(W_i,Z_i;z)\right)^2 a(z)f(z)\,\mathrm{d}z\\
&= \frac{2}{n^2}\sum_{j<k}\sum_{i_1<i_2}^{n}\int \widetilde{r}^{jk}_{Sn}(W_{i_1},Z_{i_1};z)\widetilde{r}^{jk}_{Sn}(W_{i_2},Z_{i_2};z)a(z)f(z)\,\mathrm{d}z\\
&\quad + \frac{1}{n^2}\sum_{j<k}\sum_{i=1}^{n}\int (r^{jk}_{Sn}(W_i,Z_i;z))^2 a(z)f(z)\,\mathrm{d}z\\
&\quad + \frac{2(n-1)}{n^2}\sum_{j<k}\sum_{i=1}^{n}\int \widetilde{r}^{jk}_{Sn}(W_i,Z_i;z)\ \mathbf{E}\, r^{jk}_{Sn}(W_i,Z_i;z)a(z)f(z)\,\mathrm{d}z\\
&\quad - \frac{n(n-1)}{n^2}\sum_{j<k}\int\left(\mathbf{E}\, r^{jk}_{Sn}(W_i,Z_i;z)\right)^2 a(z)f(z)dz\\
&= I_{Sn1} + I_{Sn2} + I_{Sn3} + I_{Sn4}.
\end{aligned}
\tag{3.13}
$$

These terms are analyzed by the following

Lemma 3.5. *Under the assumptions we have that under H_0*

$$\begin{aligned} nh^{(M+5)/2} I_{Sn1} &\xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma_S^2) \\ nh^{(M+5)/2} I_{Sn2} - h^{-(M+1)/2} B_S &\xrightarrow{P} 0 \\ nh^{(M+5)/2} I_{Sn3} &\xrightarrow{P} 0 \\ nh^{(M+5)/2} I_{Sn4} &\longrightarrow 0. \end{aligned}$$

Together with equation (3.12) this states the asymptotic result of the theorem. □

Proof of Theorem 3.2

Analogously to equation (3.7) we get

$$\widehat{\Gamma}_S^* = \widehat{\Gamma}_{S1}^* + \widehat{\Gamma}_{S2}^* + \widehat{\Gamma}_{S3}^*$$

where the last two terms are of lower order (use Lemma 3.6). Again we decompose

$$\widehat{\Gamma}_{S1}^* = \sum_{j<k} \widehat{\Gamma}_{S1}^{jk*}$$

and investigate wlog the case $j = 1, k = 2$. Let now $\widehat{f}_h^{1*}$ denote the kernel density estimator of $(\varepsilon_i^{1*}, Z_i)_{i=1,\dots,n}$. Replacing W^1, W^2 with $\varepsilon^{1*}, \varepsilon^{2*}$ in the definition of Γ_{S1}^{12} and applying Lemma 3.3 with $g^i = \widehat{f}_h^{i,*} - f^i, i = 1, 2$ we can decompose

$$\begin{aligned} \widehat{\Gamma}_{S1}^{12*} &= \Gamma_{S1}^{12}(\widehat{f}_h^{1*}, \widehat{f}_h^{2*}, m^1, m^2, \widehat{f}_e) \\ &= \Gamma_{S1}^{12}(f^1, f^2, m^1, m^2, f) + I_{Sn}^{12*} + \Delta_{Sn}^{12*} \\ &\quad + O_P(\|\widehat{f}_h^{1*} - f^1\|_d^2 \|\widehat{f}_h^{1*} - f^1\|_f + \|\widehat{f}_h^{2*} - f^2\|_d^2 \|\widehat{f}_h^{2*} - f^2\|_f). \end{aligned}$$

As $f^i(\varepsilon^*, z) = f^1(\varepsilon^*) f(z)$, we have that $\Gamma_{S1}^{12}(f^1, f^2, m^1, m^2, f) = 0$. Note that this property allows to construct the distribution of $\widehat{\Gamma}_S$ under H_0 by the bootstrap. Here

$$\begin{aligned} I_{Sn}^{12*} &= \int \left(\sum_{i=1}^n r_{Sn}^{12*}(\varepsilon_i^*, Z_i; z) \right)^2 a(z) f(z) \, \mathrm{d}z \\ \Delta_{Sn}^{12*} &= \int \left(\sum_{i=1}^n r_{Sn}^{12*}(\varepsilon_i^*, Z_i; z) \right)^2 a(z) (\widehat{f}_e(z) - f(z)) \, \mathrm{d}z \end{aligned}$$

and analogously to equation 3.10

$$r^{12*}_{S_n}(\varepsilon^*_i, Z_i; z) = \varepsilon^{1*}_i \partial_2 \frac{K_h(z-Z_i)}{f(z)} + \varepsilon^{1*}_i m^2(z)\partial_{M+1}\frac{K_h(z-Z_i)}{f(z)} \\ - \varepsilon^{2*}_i \partial_j \frac{K_h(z-Z_i)}{f(z)} - \varepsilon^{2*}_i m^j(z)\partial_{M+1}\frac{K_h(z-Z_i)}{f(z)}.$$

Next, lower order terms are bounded.

Lemma 3.6. *Under the assumptions we have that for any* $k = 1, \dots, M-1$

$$\|\widehat{f}^{k*}_h - f^k\|_d = O_P(h^{r-1} + (\log n/(nh^{M+3}))^{1/2})$$
$$\|\widehat{f}^{k*}_h - f^k\|_f = O_P(h^{r} + (\log n/(nh^{M+1}))^{1/2}).$$

The proof of

$$\Delta^{12*}_{S1} = o_P(n^{-1}h^{-(M+5)/2})$$

is omitted because changes in the proof of Lemma 3.4 are essentially the same as changes in the proof of in Lemma 3.5, which we will give in Lemma 3.7 below. Together we have that

$$\widehat{\Gamma}^*_{S1} = \sum_{j<k} I^{jk*}_{Sn} + o_P(n^{-1}h^{-(M+5)/2}).$$

Next, the same decomposition as in (3.13) applies

$$I^*_{Sn} = \sum_{j<k} I^{jk*}_{Sn} = I^*_{Sn1} + I^*_{Sn2} + I^*_{Sn3} + I^*_{Sn4},$$

where

$$\widetilde{r}^{jk*}_{Sn}(\varepsilon^*_i, Z_i; z) = r^{jk*}_{Sn}(\varepsilon^*_i, Z_i; z) - \mathbf{E}^* \, r^{jk*}_{Sn}(\varepsilon^*_i, Z_i; z)$$

and all expectations in the I^*_{Sni} are replaced with expectations conditional on the data.

Lemma 3.7. *Under the assumptions we have that*

$$nh^{(M+5)/2} I^*_{Sn1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_S)$$

under H_0 *conditional on the data with probability tending to one and*

$$nh^{(M+5)/2} I^*_{Sn2} - h^{-(M+1)/2} B_S \xrightarrow{P} 0$$
$$nh^{(M+5)/2} I^*_{Sn3} \xrightarrow{P} 0$$
$$nh^{(M+5)/2} I^*_{Sn4} \xrightarrow{P} 0.$$

This concludes the proof. □

Proof of the Lemmata

Proof of Lemma 3.3 Consider $\Psi(t) = \Gamma_S^{12}(f^1 + tg_1, f^2 + tg_2, m^1, m^2, f_3)$ as a function of t and write the Taylor expansion around $t = 0$

$$\Psi(t) = \Psi(0) + t\Psi'(0) + t^2\Psi''(0)/2 + t^3\Psi'''(\vartheta(t))/6.$$

Here

$$\Psi(t) = \int (\psi(t))^2 a(z) f_3(z)\,\mathrm{d}z,$$

with

$$\psi(t) = \varphi_1^2(t,z) + m^2(z)\varphi_{M+1}^1(t,z) - \varphi_2^1(t,z) - m^1(z)\varphi_{M+1}^2,$$

and

$$\varphi_k^i(t,z) = \partial_k \frac{\int w^i(f^i(w,z) + tg_i(w,z))\,\mathrm{d}w^i}{f^i(z) + tg_i(z)}.$$

Calculating the derivatives of $\Psi(t)$ requires the derivatives of $\varphi_k^i(t,z)$. Setting $i = 1$, they are given by

$$\begin{aligned}
\partial_t \varphi_k^1(t,z) &= \frac{\partial_k F(z)}{G(t,z)^2} + 2\frac{F(z)\partial_k G(t,z)}{G(t,z)^3}\\
\partial_t^2 \varphi_k^1(t,z) &= -2\frac{F(z)}{G(t,z)^3}\partial_k g_1(z) - 2g_1(z)\frac{\partial_k F(z)}{G(t,z)^3} - 6g_1(z)\frac{F(z)\partial_k G(t,z)}{G(t,z)^4}\\
\partial_t^3 \varphi_k^1(t,z) &= 12g_1(z)\frac{F(z)}{G(t,z)^4}\partial_k g_1(z) + 6(g_1(z))^2\frac{\partial_k F(z)}{G(t,z)^4}\\
&\qquad + 24(g_1(z))^2\frac{F(z)\partial_k G(t,z)}{G(t,z)^5},
\end{aligned}$$

where

$$\begin{gathered}
F(z) = f^1(z)\int w g_1(w,z)\,\mathrm{d}w - g_1(z)\int w f^1(w,z)\,\mathrm{d}w\\
G(t,z) = f^1(z) + tg_1(z)
\end{gathered}$$

are introduced for abbreviation. Starting with the first derivative

$$\Psi'(0) = 2\int \psi(0,z)\partial_t\psi(0,z)a(z)f_3(z)\,\mathrm{d}z = 0,$$

because $\psi(0,z) = 0$ $\mathbb{P}$-a.s. under H_0. The second derivative is given by

$$\Psi''(0) = 2\int (\psi(0,z)\partial_t^2\psi(0,z) - \partial_t\psi(0,z)^2)a(z)f_3(z)\,\mathrm{d}z,$$

where the first integrand is again zero under H_0 almost surely. For the second integrand, it has to be summed up over

$$\begin{aligned}\partial_t\varphi_k^1(0,z) &= \partial_k\Big(\int\frac{wg_1(w,z)\,\mathrm{d}w}{f^1(z)} - \frac{g_1(z)}{f^1(z)}\int\frac{wf^1(w,z)\,\mathrm{d}w}{f^1(z)}\Big)\\ &= \partial_k\Big(\int\frac{w}{f^1(z)}g_1(w,z)\,\mathrm{d}w - \int g_1(w,z)\,\mathrm{d}w\frac{m^1(z)}{f^1(z)}\Big)\\ &= \partial_k\int\alpha^1(w,z)g_1(w,z)\,\mathrm{d}w,\end{aligned}$$

from which the second term in equation (3.8) is obtained. Finally, the third derivative has to be bounded

$$\Psi'''(t) = 2\int(\psi(t,z)\partial_t^3\psi(t,z) + 3\partial_t\psi(t,z)\partial_t^2\psi(t,z))a(z)f_3(z)\,\mathrm{d}z.$$

Note that $|G(t,z)^{-1}| \le |G(1,z)^{-1}| \le 2/b$ because $f^1(z) > b$ and $|g_1(z)| \le b/2$. Therefore only the numerators of the derivatives of $\psi(t,z)$ have to be bounded. By noting that the derivatives of g_1, F and G are bounded by $\|g_1\|_d$ and g_1 and F are bounded by $\|g_1\|_f$ we obtain

$$\Psi'''(\vartheta(t)) = O(\|g_1\|_d^2\|g_1\|_f + \|g_2\|_d^2\|g_2\|_f),$$

which completes the proof of the lemma. □

Proof of Lemma 3.4 Convergence in probability has to be shown for

$$\Delta_{Sn}^{12} = \frac{1}{n}\sum_{ijk}\gamma_{ijk}^{12},$$

where

$$\begin{aligned}\gamma_{ijk}^{12} &= (\alpha^1(W_i,Z_k)K_h^2(Z_k,Z_k-Z_i) - \alpha^2(W_i,Z_k)K_h^1(Z_k,Z_k-Z_i))\\ &\quad\times(\alpha^1(W_j,Z_k)K_h^2(Z_k,Z_k-Z_j) - \alpha^2(W_j,Z_k)K_h^1(Z_k,Z_k-Z_j))a(Z_k)\\ &\qquad-\int(\alpha^1(W_i,z)K_h^2(z,z-Z_i) - \alpha^2(W_i,z)K_h^1(z,z-Z_i))\\ &\qquad\times(\alpha^1(W_j,z)K_h^2(z,z-Z_j) - \alpha^2(W_j,z)K_h^1(z,z-Z_j))a(z)f(z)\,\mathrm{d}z.\end{aligned}$$

Multiplying out gives

$$\gamma_{ijk}^{12} = \bar\gamma_{ijk}^{11} + \bar\gamma_{ijk}^{22} - \bar\gamma_{ijk}^{12} - \bar\gamma_{ijk}^{21},$$

where

$$\begin{aligned}\bar{\gamma}_{ijk}^{lm} &= \alpha^l(W_i, Z_k)K_h^m(Z_k, Z_k - Z_i)\alpha^m(W_j, Z_k)K_h^l(Z_k, Z_k - Z_j)a(Z_k)\\ &\qquad - \int \alpha^l(W_j, z)K_h^m(z, z - Z_j)\alpha^m(W_j, z)K_h^l(z, z - Z_j)a(z)f(z)\,\mathrm{d}z.\end{aligned}$$

This enables to write

$$\Delta_{Sn}^{12} = \bar{\Delta}_{11} + \bar{\Delta}_{22} - \bar{\Delta}_{21} - \bar{\Delta}_{12}.$$

All four terms have the same structure and so we restrict to

$$\mathbf{E}\,\bar{\Delta}_{11} = \frac{1}{n^3}\sum_{i,j,k}\mathbf{E}\,\bar{\gamma}_{ijk}^{11} = o(n^{-1}h^{-(M+5)/2}),$$

where only the cases $i = k \neq j, j = k \neq i$ and $i = j = k$ have to be considered, all others have expectation zero. In these cases, two (resp. one) change of variables can be applied and the statement follows.

To show the convergence in probability, Markov's inequality is applied with the second moments and it has to be investigated

$$\mathbf{E}(\bar{\Delta}_{11})^2 = \frac{1}{n^6}\sum_{ijk}\mathbf{E}(\bar{\gamma}_{ijk}^{11})^2 + \frac{2}{n^6}\sum_{ijk}\sum_{i'j'k'}\mathbf{E}\,\bar{\gamma}_{ijk}^{11}\bar{\gamma}_{i'j'k'}^{11}.$$

The covariance parts vanish, if there are six different indices. In the case with five different indices we have $O(n^5)$ terms where $k = k'$ (other combinations vanish) and they give a total contribution of $O(n^{-1}h^{r-4}) = o(n^{-1}h^{-(M+5)/2})$, since the leading terms have expectation zero. If there are four different indices, by change of variables they are in the worst case (when the leading terms do not cancel, e. g. if $i = i'$ and $j = j'$) of order of $O(h^{-4})$. As there are $O(n^3)$ such terms their overall contribution is $O(n^{-3}h^{-4}) = o(n^{-1}h^{-(M+5)/2})$. If the number of different indices is $N = 2, 3$ the overall contribution of these terms is $O(h^{-4(M+2)}h^{N(M+1)}n^{N-6}) = o(n^{-1}h^{-(M+5)})$.

Next consider the variance terms. If there are three different indices, three changes of variables can be applied and the total contribution is $O(h^{-4-2(M+1)}n^{-3}) = o(n^{-1}h^{-(M+5)/2})$. If there are two different indices, one change of variables cannot be applied and we obtain terms of order $O(h^{-4-3(M+1)})$ with a contribution of $O(h^{-4-3(M+1)}n^{-4}) = o(n^{-1}h^{-(M+5)/2})$. If $i = j = k$ one change of variables is still possible and the contribution is $O(h^{-4-3(M+1)}n^{-5}) = o(n^{-1}h^{-(M+5)/2})$.

This completes the proof that $\bar{\Delta}_{Sn}^{12} = o_p(n^{-1}h^{-(M+5)/2})$. □

Proof of Lemma 3.5 Before showing the statements of this lemma, we start by investigating $\widetilde{r}_{Sn}(\cdot)$. Calculating the derivatives in this equation, one needs

$$\begin{aligned}(3.14)\quad \partial_k\alpha^j(W_i,z)K_h(z-Z_i) &= \alpha^j(W_i,z)\partial_k K_h(z-Z_i)\\ &+\alpha^j(W_i,z)K_h(z-Z_i)\frac{\partial_k f(z)}{f(z)}+\frac{K_h(z-Z_i)}{f(z)}\partial_k m^j(z)\end{aligned}$$

to obtain

$$\begin{aligned}(3.15)\quad r^{jk}_{Sn}(W_i,Z_i;z) &= \alpha^j(W_i,z)h^{-(M+2)}K^k(z,(z-Z_i)/h)\\ &-\alpha^k(W_i,z)h^{-(M+2)}K^j(z,(z-Z_i)/h)\\ &+\alpha^j(W_i,z)K_h(z-Z_i)\big(\frac{\partial_k f(z)}{f(z)}+m^k(z)\frac{\partial_{M+1}f(z)}{f(z)}\big)\\ &-\alpha^k(W_i,z)K_h(z-Z_i)\big(\frac{\partial_j f(z)}{f(z)}-m^j(z)\frac{\partial_{M+1}f(z)}{f(z)}\big).\end{aligned}$$

Because the sum over the third terms in (3.14) is zero under H_0. Here the last two terms converge faster, as the chain rule applied to the kernel brings an extra h to the first two terms.

Asymptotic Normality of I_{Sn1} I_{Sn1} can be written as U-Statistic by

$$I_{Sn1}=\sum_{i_1<i_2}h_{Sn}(Y_{i_1},Y_{i_2}),$$

with

$$h_n(Y_{i_1},Y_{i_2})=\frac{2}{n^2}\sum_{j<k}\int\widetilde{r}^{jk}_{Sn}(W_{i_1},Z_{i_2};z)\widetilde{r}^{jk}_{Sn}(W_{i_2},Z_{i_2};z)a(z)f(z)\,\mathrm{d}z.$$

Asymptotic normality is shown by using Lemma 3.1. First note that as we have independent and identically distributed data we can define $\sigma_n^2=\mathbf{E}\,h_n(Y_i,Y_j)^2$ and get for the first condition of the lemma

$$\max_{1\le i\le n}\sum_{\substack{k=1\\k\neq i}}^{n}\mathbf{E}\,h_n(Y_i,Y_k)^2=n\sigma_n^2$$

and

$$\begin{aligned}\mathbf{var}\,I_{Sn1}=\sum_{i_1<i_2}\mathbf{var}\,h_n(Y_{i_1},Y_{i_2})+\sum_{i_1<i_2}\sum_{\substack{i_3<i_4\\(i_3,i_4)\neq(i_1,i_2)}}\\ +\mathbf{cov}(h_n(Y_{i_1},Y_{i_2}),h_n(Y_{i_3},Y_{i_4}))=\frac{n(n-1)}{2}\sigma_n^2,\end{aligned}$$

because $h_n(\cdot,\cdot)$ is centered. This implies directly the first condition of Lemma 3.1. For the second we need

$$\begin{aligned}(3.16)\quad \mathbf{E}\, I_{Sn1}^4 = \sum_{i_1<i_2} \mathbf{E}\, h_n^4(Y_{i_1},Y_{i_2}) + 3 \sum_{i_1<i_2} \sum_{\substack{i_3<i_4\\(i_3,i_4)\neq(i_1,i_2)}} \mathbf{E}\, h_n(Y_{i_1},Y_{i_2})^2 h_n(Y_{i_3},Y_{i_4})^2 \\ + 24 \sum_{i_1<i_2} \sum_{i_3\neq i_1,i_2} \mathbf{E}\, h_n(Y_{i_1},Y_{i_2})^2 h_n(Y_{i_1},Y_{i_3}) h_n(Y_{i_2},Y_{i_3}) \\ + 3 \sum_{i_1} \sum_{i_2\neq i_1} \sum_{i_3\neq i_1,i_2} \sum_{i_4\neq i_1,i_2,i_3} \mathbf{E}\, h_n(Y_{i_1},Y_{i_2}) h_n(Y_{i_2},Y_{i_3}) h_n(Y_{i_3},Y_{i_4}) h_n(Y_{i_4},Y_{i_1}).\end{aligned}$$

To show the second condition, these terms have to be calculated. Starting with the denominator, we have to calculate

$$(3.17)\qquad \sigma_n^2 = \mathbf{E}\, h_n(Y_1,Y_2)^2.$$

Resolving the square and multiplying the $\widetilde{r}_{Sn}^{j}(\cdot)$ gives four terms, where the first is given by

$$\begin{aligned}(3.18)\quad \frac{4}{n^4} \sum_{j<k} \sum_{j'<k'} \iint r_{Sn}^{jk}(w_1,z_1,z) r_{Sn}^{jk}(w_2,z_2,z) a(z) f(z)\, \mathrm{d}z \\ \times \int r_{Sn}^{j'k'}(w_1,z_1,z) r_{n}^{j'k'}(w_2,z_2,z) a(z) f(z)\, \mathrm{d}z f(w_1,z_1) f(w_2,z_2)\, \mathrm{d}w_1\, \mathrm{d}z_1)\, \mathrm{d}w_2\, \mathrm{d}z_2\end{aligned}$$

changing variables[2] to $\bar{z} = (z - z_1)/h$ in both integrals and expanding $\alpha(\cdot), a(\cdot)$ and $f(\cdot)$ gives (we only consider the leading terms from equation (3.15))

$$\begin{aligned}= \frac{4}{n^4 h^{2(M+1)} h^2} \sum_{j<k} \sum_{j'<k'} \iint \big(\alpha^j(w_1,z_1) K^k(z_1,\bar{z}) - \alpha^k(w_1,z_1) K^j(z_1,\bar{z})\big) \\ \big(\alpha^j(w_2,z_1) K^k(z_1,\bar{z}+(z_1-z_2)/h) - \alpha^k(w_2,z_1) K^j(z_1,\bar{z}+(z_1-z_2)/h)\big) a(z_1) f(z_1)\, \mathrm{d}\bar{z} \\ \times \int \big(\alpha^{j'}(w_1,z_1) K^{k'}(z_1,\bar{z}) - \alpha^{k'}(w_1,z_1) K^{j'}(z_1,\bar{z})\big) \\ \big(\alpha^{j'}(w_2,z_1) K^{k'}(z_1,\bar{z}+(z_1-z_2)/h) - \alpha^{k'}(w_2,z_1) K^{j'}(z_1,\bar{z}+(z_1-z_2)/h)\big) a(z_1) f(z_1)\, \mathrm{d}\bar{z} \\ f(w_1,z_1) f(w_2,z_2)\, \mathrm{d}w_1\, \mathrm{d}w_2\, \mathrm{d}z_1\, \mathrm{d}z_2 (1 + O(h)).\end{aligned}$$

[2] Here the notation is simplified. As z_1 is $M + 1$-dimensional one has to apply $M + 1$ substitutions.

Now substitute $\bar{\bar{z}} = (\widetilde{z}_1 - \widetilde{z}_2)/h$ and expand $f(\cdot)$ to obtain

$$
\begin{aligned}
= \frac{4}{n^4 h^{M+5}} \sum_{j<k} \sum_{j'<k'} \iint & \big(\alpha^j(w_1, z_1) K^k(z_1, \bar{z}) - \alpha^k(w_1, z_1) K^j(z_1, \bar{z})\big) \\
& \big(\alpha^j(w_2, z_1) K^k(z_1, \bar{z} + \bar{\bar{z}}) - \alpha^k(w_2, z_1) K^j(z_1, \bar{z} + \bar{\bar{z}})\big) a(z_1) f(z_1)\,\mathrm{d}\bar{z} \\
& \times \int \big(\alpha^{j'}(w_1, z_1) K^{k'}(z_1, \bar{z}) - \alpha^{k'}(w_1, z_1) K^{j'}(z_1, \bar{z})\big) \\
& \big(\alpha^{j'}(w_2, z_1) K^{k'}(z_1, \bar{z} + \bar{\bar{z}}) - \alpha^{k'}(w_2, z_1) K^{j'}(z_1, \bar{z} + \bar{\bar{z}})\big) a(z_1) f(z_1)\,\mathrm{d}\bar{z} \\
& f(w_1, z_1) f(w_2, z_1)\,\mathrm{d}w_1\,\mathrm{d}w_2\,\mathrm{d}z_1\,\mathrm{d}\bar{\bar{z}}(1 + O(h)).
\end{aligned}
$$

Multiplying out the remaining brackets results in 16 terms of the kind (remember the definition of $K^{jkj'k'}(z)$

$$
\begin{aligned}
\iint & \alpha^j(w_1, z_1) \alpha^{j'}(w_1, z_1) f(z_1) f(w_1, z_1)\,\mathrm{d}w_1 \\
& \int \alpha^k(w_2, z_1) \alpha^{k'}(w_2, z_1) f(z_1) f(w_2, z_1)\,\mathrm{d}w_2 a(z_1)^2 K^{jkj'k'}(z_1)\,\mathrm{d}z_1
\end{aligned}
$$

Now, by the definition of α^j one concludes

$$
= \int \sigma^{jj'}(z) \sigma^{kk'}(z) a(z)^2 K^{jkj'k'}(z)\,\mathrm{d}z
$$

. Taking care of the summation we have in total that

$$
\sigma_n^2 = \frac{4}{n^4 h^{M+5}} \sigma_S^2 (1 + O(h)).
$$

In the other terms arising from equation (3.17) one has a product of two expectations. This allows to change variables once more and these terms are of total order of $O(n^{-4} h^{-4})$.

Similar calculations show that

$$
\begin{aligned}
\mathbf{E}\, h_n(Y_1, Y_2)^4 &= O(n^{-8} h^{-3M-7}) \\
\mathbf{E}\, h_n(Y_1, Y_2)^2 h_n(Y_1, Y_3)^2 &= O(n^{-8} h^{-2M-6}) \\
\mathbf{E}\, h_n(Y_1, Y_2)^2 h_n(Y_1, Y_3) h_n(Y_2, Y_3) &= O(n^{-8} h^{-2M-6}) \\
\mathbf{E}\, h_n(Y_1, Y_2) h_n(Y_2, Y_3) h_n(Y_3, Y_4) h_n(Y_1, Y_4) &= O(n^{-8} h^{-M-5}).
\end{aligned}
$$

Using some combinatorics one sees from equation (3.16) that the total contribution of terms of these kinds to $\mathbf{E}\, I_{Sn1}^4$ is at most $O(n^{-4} h^{-(M+1)})$. So $\mathbf{E}\, I_{n1}^4$ is asymptotically dominated by terms with $\mathbf{E}\, h_n(Y_1, Y_2)^2 h_n(Y_3, Y_4)^2 = (\mathbf{E}\, h_n(Y_1, Y_2^2))^2$. Therefore the second condition from Lemma 3.1 is fulfilled because

$$
\frac{\mathbf{E}\, I_{Sn1}^4}{(\mathbf{var}\, I_{Sn1})^2} = \frac{12 n^{-4} h^{-2(M+5)} \sigma_H^4 (1 + o(1))}{(2 n^{-2} h^{-(M+5)} \sigma_H^2 (1 + o(1)))^2} \longrightarrow 3
$$

and the asymptotic normality of I_{n1} is established.

Convergence in Probability of I_{n2} For the rest of the proof the lower order terms in equation (3.15) are omitted. The expected value of the test statistic is given by

$$\begin{aligned}\mathbf{E}\, I_{n2} = \frac{1}{n}\sum_{j<k}\iint \big(&\alpha^j(w_1,z)h^{-(M+2)}K^k(z,(z-z_1)/h)\\ &- \alpha^k(w_1,z)h^{-(M+2)}K^j(z,(z-z_1)/h)\big)^2 a(z)f(z)\,\mathrm{d}z f(w_1,z_1)\,\mathrm{d}w_1\,\mathrm{d}z_1\end{aligned}$$

Substitution of $\bar{z}=(z-z_1)/h$ leads by a Taylor expansion to

$$\begin{aligned}&= \frac{1}{nh^{M+3}}\sum_{j<k}\iint\big(\alpha^j(w_1,z_1)K^k(z_1,\bar{z}) - \alpha^k(w_1,z_1)K^j(z_1,\bar{z})\big)^2\\ &\qquad \times a(z_1)f(z_1)\,\mathrm{d}\bar{z}f(w_1,z_1)\,\mathrm{d}w_1\,\mathrm{d}z_1 + o(n^{-1}h^{-(M+5)/2})\\ &= n^{-1}h^{-(M+3)}B_S + o(n^{-1}h^{-(M+5)/2}),\end{aligned}$$

where the brackets are resolved before integrating.
To establish convergence in probabilty, Markov's inequality with second moments is applied, which requires to calculate

$$\begin{aligned}\mathbf{E}\, I_{n2}^2 = \frac{1}{n^3}\int\Big(\sum_{j<k}\int\big(&\alpha^j(w_1,z)h^{-(M+2)}K^k(z,(z-z_1)/h)\\ &- \alpha^k(w_1,z)h^{-(M+2)}K^j(z,(z-z_1)/h)\big)a(z)f(z)\,\mathrm{d}z\Big)^2\mathrm{d}f(w_1,z_1)\,\mathrm{d}w_1\,\mathrm{d}z_1\end{aligned}$$

Changing variables as before results in

$$\begin{aligned}&= \frac{1}{n^3h^{M+3}}\int\Big(\sum_{j<k}\int\big(\alpha^j(w_1,z)K^k(z,\bar{z})\\ &\qquad - \alpha^k(w_1,z)K^j(z,\bar{z})a(z_1)f(z_1)\big)\,\mathrm{d}\bar{z}\Big)^2 f(w_1,z_1)\,\mathrm{d}w_1\,\mathrm{d}z_1.\end{aligned}$$

This gives the second statement of the lemma.

Convergence in Probability of I_{Sn3} Because $\widetilde{r}_n^{jk}(W_i,Z_i;z)$ are centered functions, we have that $\mathbf{E}\, I_{Sn3}=0$. Substituting $\bar{z}=(z-z_1)/h$ for z_1 gives

(3.19)

$$\begin{aligned}&\mathbf{E}\, r_{Sn}^{jk}(W_1,Z_1;z)\\ &\quad = h^{-1}\int\big(\alpha^j(w_1,z)K^k(z,\bar{z}) - \alpha^k(w_1,z)K^j(z,\bar{z})\big)f(w_1,z)dw_1 d\bar{z} + O(h^{r-1})\\ &\quad = O(h^{r-1})\end{aligned}$$

for every $z \in \mathcal{A}$ because of equation (3.11) and therefore

$$\begin{aligned}\mathbf{E}\, I_{Sn3}^2 &= \frac{4(n-1)^2}{n^3}\, \mathbf{E}\Big(\sum_{j<k} \int \widetilde{r}_{Sn}^{jk}(W_i, Z_i; z)\, \mathbf{E}\, r_{Sn}^{jk}(W_1, Z_1; z) a(z) f(z)\, \mathrm{d}z\Big)^2 \\ &= O(n^{-1} h^{2(r-1)}),\end{aligned}$$

which is of $o(n^{-1}h^{-(M+5)/2})$.

Convergence of I_{Sn4} The convergence of the deterministic part follows from (3.19) and the upper bound of the bandwidth sequence

$$I_{Sn4} = O(h^{2(r-1)}) = o(n^{-1}h^{-(M+5S)/2}),$$

which completes the proof of the lemma. □

Proof of Lemma 3.6 It follows from equation (3.6) that only the first part has to be investigated, because the second part in equation (3.6), concerning the density estimator, is unchanged in the bootstrap sample and has the desired rate.

Uniform convergence of the function estimator For the norm $\|\cdot\|_f$ we have to show uniform convergence of

$$\widehat{e}_h^{k*}(z) = \frac{\widehat{g}_h^{k*}(z)}{\widehat{f}_h(z)} = \frac{n^{-1}\sum_{i=1}^n K_h(z - Z_i)\varepsilon_i^{k*}}{n^{-1}\sum_{i=1}^n K_h(z - Z_i)}$$

to $\mathbf{E}(\varepsilon \mid Z) = 0$. Because $\widehat{f}_h(z)$ is bounded from below on $\mathcal{A}$ almost surely for n large enough, the rate of convergence follows from the numerator.
First a truncation has to be applied. Define $\widetilde{\varepsilon}_i^{k,*} = \mathbf{1}_{\{\varepsilon_i^{k*} \le nh^{M+1}\}}$ and then decompose

$$\begin{aligned}(3.20)\quad \frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\varepsilon_i^{k*} &= \frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*} \\ &\qquad + \frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\varepsilon_i^{k*}\mathbf{1}_{\{\varepsilon_i^{k*} > nh^{M+1}\}}.\end{aligned}$$

The second term on the right hand side can be bounded using Markov's inequality with the first moment and $\mathbf{E}\,|\varepsilon_i^{k*}\mathbf{1}_{\{\varepsilon_i^{k*}>nh^{M+1}\}}| = O(n^{-2}h^{-2(M+1)})$, because the forth moment of $\varepsilon_i^{k,*}$ is finite. Changing variables once it follows that

$$\mathbf{E}\Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\varepsilon_i^{k*}\mathbf{1}_{\{\varepsilon_i^{k,*}>nh^{M+1}\}}\Big| = O(n^{-2}h^{-2(M+1)}),$$

from which the desired rate for the second term in (3.20) follows.
To bound the first term in 3.20 the compact set $\mathcal{A}$ is covered with N cubes $\mathcal{A}_l = \{z \mid \|z - z_l\| < \eta_N\}, l = 1, \dots, N$. Then it holds that

$$\mathbb{P}\Big(\sup_{z \in \mathcal{A}} \Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*}\Big| > c\Big) \leq \mathbb{P}\Big(\sup_{z_l} \Big|\frac{1}{n}\sum_{i=1}^n K_h(z_l - Z_i)\widetilde{\varepsilon}_i^{k*}\Big| > c/2\Big)$$
$$+ \sup_l \mathbb{P}\Big(\sup_{z \in \mathcal{A}_l} \Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*} - \frac{1}{n}\sum_{i=1}^n K_h(z_l - Z_i)\widetilde{\varepsilon}_i^{k*}\Big| > c/2\Big).$$

If N becomes large, the second terms becomes negligible compared to the first. Applying Bonferroni's inequality the first can be bounded by

$$N \sup_l \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\varepsilon_i^{k*}\Big| > c/2\Big).$$

And this probability is bounded using Bernstein's inequality

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*}\Big| > \Big(\frac{\log n}{nh^{M+1}}\Big)^{1/2}\frac{c}{2}\Big)$$
$$\leq 2\exp\Big(-\frac{c^2(\log n)/(4nh^{M+1})}{4\sum_{i=1}^n \mathbf{E}(\frac{1}{n}K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*})^2 + \widetilde{c}(\frac{\log n}{n^3 h^{3(M+1)}})^{1/2}}\Big),$$

where $\widetilde{c}$ is the constant arising from Cramer's conditions on the distribution of $\widetilde{\varepsilon}^{k*}$. It follows from standard arguments that

$$\sum_{i=1}^n \mathbf{E}\Big(\frac{1}{n}K_h(z - Z_i)\widetilde{\varepsilon}_i^{k*}\Big)^2 = O(n^{-1}h^{-(M+1)}),$$

and so we get that

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n K_h(z - Z_i)\widetilde{\varepsilon}_i^{*}\Big| > \Big(\frac{\log n}{nh^{M+1}}\Big)^{1/2}\frac{c}{2}\Big) = O(n^{-1}).$$

Then, for $N = o(n)$ the desired rate of convergence is obtained.

Uniform convergence of the derivative estimator Applying the quotient rule we obtain

$$\partial_p \widehat{e}_h^{k*}(z) = \frac{\partial_p \widehat{g}_h^{k*}(z)}{\widehat{f}_h(z)} - \frac{\widehat{g}_h^{k*}(z)}{\widehat{f}_h(z)}\frac{\partial_p \widehat{f}_h(z)}{\widehat{f}_h(z)}$$

The second term converges faster by Lemma 3.2 and the first part of this lemma. For the first term, the proof of concerning the norm $\|\cdot\|_f$ has to be repeated with an extra h^{-1} from the inner derivative. Note that except for this only the kernel changes. Then, the statement follows. □

Proof of Lemma 3.7 Define $Y_i^* = (\varepsilon_i^*, Z_i)$. To derive the convergence result for

$$I_{Sn1}^* = \sum_{i_1 < 1_2} h_n(Y_{i_1}^*, Y_{i_2}^*),$$

again Lemma 3.1 has to be applied. This is done by showing that the conditions hold with probability tending to one, i. e.

$$\frac{\max_{1\le i\le n}\sum_{j=1}^n \mathbf{E}^*\, h_n(Y_i^*, Y_j^*)^2}{\mathbf{var}^*\, I_{Sn1}^*} \xrightarrow{P} 0$$
$$\frac{\mathbf{E}^*(I_{Sn1}^*)^4}{(\mathbf{var}^*\, I_{Sn1}^*)^2} \xrightarrow{P} 3.$$

Note, that by construction $\mathbf{E}^*\, h_n(Y_1^*, Y_2^*) = 0$ and $\mathbf{E}^*(h_n(Y_1^*, Y_2^*) \mid Y_1^*) = \mathbf{E}^*(h_n(Y_1^*, Y_2^*) \mid Y_2^*) = 0$ almost surely.
Calculating the derivatives in $r_{Sn}^{jk*}(\cdot)$ analogously to equation (3.15), one obtains

(3.21)
$$\begin{aligned} r_{Sn}^{jk*}(\varepsilon_i^*, Z_i; z) &= \varepsilon_i^{j*} h^{-(M+2)} \frac{K^k(z, (z-Z_i)/h)}{f(z)} - \varepsilon_i^{k*} h^{-(M+2)} \frac{K^j(z, (z-Z_i)/h)}{f(z)} \\ &\quad + \varepsilon_i^{j*} \frac{K_h(z-Z_i)}{f(z)} \Big(\frac{\partial_k f(z)}{f(z)} + m^k(z)\frac{\partial_{M+1} f(z)}{f(z)}\Big) \\ &\quad - \varepsilon_i^{k*} \frac{K_h(z-Z_i)}{f(z)} \Big(\frac{\partial_j f(z)}{f(z)} - m^j(z)\frac{\partial_{M+1} f(z)}{f(z)}\Big) \end{aligned}$$

Next $\mathbf{E}^*\, h_n(Y_1^*, Y_2^*)^2$ has to be calculated. Using the definition of $h_n(\cdot)$ and multiplying $\widetilde{r}_{Sn}^{jk*}(\cdot)$ gives four terms where the first is given as in (3.18) by replacing the distributions of Y_1, Y_2 with the distributions of Y_1^*, Y_2^* conditional on the data. Replacing (3.21) and omitting the last two terms as they are of lower order, the leading term is given by

$$\begin{aligned} = \frac{4}{n^4 h^{2(M+1)} h^2} \sum_{j<k} \sum_{j'<k'} \iint & \big(\varepsilon_1^{j*} K^k(z, (z-Z_1)/h) - \varepsilon_1^{k*} K^j(z, (z-Z_1)/h)\big) \\ & \big(\varepsilon_2^{j*} K^k(z, (z-Z_2)/h) - \varepsilon_2^{k*} K^j(z, (z-Z_2)/h)\big) a(z) f(z)^{-1}\, \mathrm{d}z \\ & \times \int \big(\varepsilon_1^{j'*} K^{k'}(z, (z-Z_1)/h) - \varepsilon_1^{k'*} K^{j'}(z, (z-Z_1)/h)\big) \\ & \big(\varepsilon_2^{j'*} K^{k'}(z, (z-Z_2)/h) - \varepsilon_2^{k'*} K^{j'}(z, (z-Z_2)/h)\big) a(z) f(z)^{-1}\, \mathrm{d}z \\ & \mathrm{d}\mathbb{P}^{Y_1^* | Y_1, \ldots, Y_n}(\varepsilon^{1*})\, \mathrm{d}\mathbb{P}^{Y_2^* | Y_1, \ldots, Y_n}(\varepsilon^{2*}). \end{aligned}$$

The bootstrap residuals are chosen such that they match the first moments of the empirical residuals. Multiplying out, replacing the conditional expectation of

ε_i^{j*} and rearranging inside the brackets it follows that

$$= \frac{4}{n^4 h^{2(M+1)} h^2} \sum_{j<k} \sum_{j'<k'} \iint \big((W_1^j - \widehat{m}_h^j(Z_1)) K^k(z,(z-Z_1)/h) \\ - (W_1^k - \widehat{m}_h^k(Z_1)) K^j(z,(z-Z_1)/h) \big) \\ \big((W_2^j - \widehat{m}_h^j(Z_2)) K^k(z,(z-Z_2)/h) - (W_2^k - \widehat{m}_h^k(Z_2)) K^j(z,(z-Z_2)/h) \big) \frac{a(z)}{f(z)} \,\mathrm{d}z \\ \times \int \big((W_1^{j'} - \widehat{m}_h^{j'}(Z_1)) K^{k'}(z,(z-Z_1)/h) - (W_1^{k'} - \widehat{m}_h^{k'}(Z_1)) K^{k'}(z,(z-Z_1)/h) \big) \\ \big((W_2^{j'} - \widehat{m}_h^{j'}(Z_2)) K^{k'}(z,(z-Z_2)/h) - (W_2^{k'} - \widehat{m}_h^{k'}(Z_2)) K^{j'}(z,(z-Z_2)/h) \big) \frac{a(z)}{f(z)} \,\mathrm{d}z$$

and now using the uniform convergence of the regression estimator

$$= \frac{4}{n^4 h^{2(M+1)} h^2} \sum_{j<k} \sum_{j'<k'} \iint \big(\alpha^j(W_1,Z_1) K^k(z,(z-Z_1)/h) \\ - \alpha^k(W_1,Z_1) K^j(z,(z-Z_1)/h) \big) \\ \big(\alpha^j(W_2,Z_2) K^k(z,(z-Z_2)/h) - \alpha^k(W_2,Z_2) K^j(z,(z-Z_2)/h) \big) \frac{a(z)}{f(z)} f(Z_1) f(Z_2) \,\mathrm{d}z \\ \times \int \big(\alpha^{j'}(W_1,Z_1) K^{k'}(z,(z-Z_1)/h) - \alpha^{k'}(W_1,Z_1) K^{k'}(z,(z-Z_1)/h) \big) \\ \big(\alpha^{j'}(W_2,Z_2) K^{k'}(z,(z-Z_2)/h) - \alpha^{k'}(W_2,Z_2) K^{j'}(z,(z-Z_2)/h) \big) \frac{a(z)}{f(z)} f(Z_1) f(Z_2) \,\mathrm{d}z \\ \times (1 + O_P(h^r + (\log n/(nh)^{M+1})^{1/2}),$$

which has a similar structure as $h_n(Y_1, Y_2)^2$. Therefore, taking expectations and applying the appropriate changes of variables, the same leading term can be derived (see the calculations of (3.17)). Next by the conditional independence of the bootstrap residuals, we get

$$\mathbf{var}^* I_{n1}^* = \sum_{i_1<i_2} \mathbf{E}^* h_n(Y_{i_1}^*, Y_{i_2}^*)^2,$$

because $h_n(Y_{i_1}^*, Y_{i_2}^*)$ is centered conditional on the data. To bound this in probability, use Markov's inequality with the first moment

$$\mathbf{E} \big| \sum_{i_1<i_2} \mathbf{E}^* h_n(Y_{i_1}^*, Y_{i_2}^*)^2 \big| = \sum_{i_1<i_2} \mathbf{E}\, h_n(Y_{i_!}^*, Y_{i_2}^*)^2 = n^{-2} h^{-(M+5)} \sigma_S^2 (1 + o(1)),$$

from which

$$\mathbf{var}^* I_{Sn1}^* \xrightarrow{P} \mathbf{var}\, I_{Sn1}$$

follows. This is now used to show the first condition. By the iid-assumption on the data sample, for the maximum it holds that

$$\mathbb{P}\left(\frac{\max_{i=1,\dots,n}\sum_{j=1,j\neq i}^{n}\mathbf{E}^{*}\,h_n(Y_i^*,Y_j^*)^2}{\mathbf{var}\,I_{Sn1}}>c\right)=n\mathbb{P}\left(\frac{\sum_{j=2}^{n}\mathbf{E}^{*}\,h_n(Y_1^*,Y_j^*)^2}{\mathbf{var}\,I_{Sn1}}>c\right).$$

And for the right hand side we use the Markov inequality with second moments and similar calculations as in Lemma 3.5 to obtain

$$\mathbb{P}\left(\frac{\sum_{j=2}^{n}\mathbf{E}^{*}\,h_n(Y_1^*,Y_j^*)^2}{\mathbf{var}\,I_{Sn1}}>c\right)=O(n^{-2}h^4)=o(n^{-1}).$$

For the second condition we again use the convergence of the denominator. Then bounding the numerator with Markov's inequality and the first moments leads to similar calculations as done in Lemma 3.5. Stochastic convergence of I_{Sn2}^* and I_{Sn3}^* consists of using iterated expectations and repeating there the same calculations as in this lemma. □

Chapter 4

Nonparametric Estimation of Additive Multivariate Diffusion Processes

4.1 Introduction

Motivated by the application of continuous-time stochastic processes in financial econometrics, nonparametric estimation methods for diffusion processes have become a broad area of statistical research. The review papers of Cai and Hong (2003) and Fan (2005) provide an overview over recent results. Since the estimation of the drift and diffusion function of a diffusion process can be regarded as a regression problem, kernel smoothing techniques arise naturally.

Beginning with Florens-Zmirou (1993) a large number of articles has been concerned with the application of nonparametric regression techniques to diffusion processes. Various modifications have been considered, among them higher order approximations (Stanton, 1997, Fan and Zhang, 2003), nonstationary processes (Bandi and Philips, 2003) or jump diffusions (Bandi and Nguyen, 2003). Usually, high frequency sampling is considered, where both the total observation time tends to infinity and the distance between consecutive observations shrinks to zero. But other sampling schemes have been used as well. The monograph by Kutoyants (2004) covers the case of continuous time observations. The issue of fixed time intervals between consecutive observations, so called low frequency sampling, has been addressed by Aït-Sahalia (1996a), Jiang and Knight (1997) and Gobet, Hoffmann and Reiß (2005) among others.

Most of the articles cited above only deal with the case of a scalar diffusion. Brugiere (1993) and Bandi and Moloche (2001) investigate kernel estimators for multivariate diffusion processes. Their results report the well-known curse of dimensionality in nonparametric regression. This means that the rate of convergence of the estimators becomes worse, if the number of dimensions increases and therefore larger samples have to be available. Because of the dependence structure even in the scalar case relatively large samples are required to obtain reliable estimators. This effect is thus enlarged for multivariate data and therefore the use of nonparametric regression techniques is restricted (curse of dependence). However, the curse of dimensionality can be circumvented by imposing more structure on the unknown functions.

A common approach is to use additive models, assuming that the unknown function is a sum of one-dimensional components. These models provide a powerful technique to overcome the dimensionality problem and maintain high flexibility. Estimation of such models requires iterative procedures and the asymptotic analysis is much more complex than in the classical setting. For the estimation of the additive components Mammen, Linton and Nielsen (1999) have introduced smooth backfitting, an iterative procedure that uses a projection interpretation of usual kernel estimators. For the classical nonparametric regression model it has been shown that smooth backfitting based on local linear estimators is oracle efficient, i. e. the estimator of a single component has the same bias and variance as an infeasible estimator based on the knowledge of all other components.

In this article a multivariate diffusion process is considered and (some or all) elements of the drift vector and the diffusion matrix are modelled as additive functions. Smooth backfitting based on local linear and Nadaraya-Watson estimators is used to estimate the components. For all estimators the asymptotic distributions under high frequency sampling are derived. The Nadaraya-Watson based estimators achieve the same variance as the oracle estimator, while the bias is not oracle. The local linear based estimators are shown to be fully oracle efficient.

The remainder of this chapter is organized as follows. First, the additive diffusion model is formally introduced. In section 3 the smooth backfitting estimators are defined. The asymptotic properties are presented in Section 4 and results of a finite sample study, investigating the performance of the estimators, are given in Section 5. An illustrative data example, using interest rate data is presented in Section 6. All proofs are deferred to the appendix.

4.2 Additive Multivariate Diffusion Processes

Let $(X_t)_{t\geq 0} = \big((X_t^1,\dots,X_t^d)'\big)_{t\geq 0}$ be a d-dimensional stochastic process on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F})_{t\geq 0}, \mathbb{P})$ which satisfies the time-homogenous stochastic differential equation

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \Sigma(X_t)\,\mathrm{d}W_t, \tag{4.1}$$

with some initial condition X_0 and a $\widetilde{d}$-dimensional Brownian motion $(W_t)_{t\geq 0} = \big((W_t^1,\dots,W_t^{\widetilde{d}})'\big)_{t\geq 0}$ with independent components adapted to the same filtration $\mathbb{F}$. The drift vector $\mu(x) = (\mu^1(x),\dots,\mu^d(x))'$ and the dispersion matrix $\Sigma(x) = \big(\sigma^{ij}(x)\big)_{1\leq i\leq d, 1\leq j\leq \widetilde{d}}$ are both Borel measurable. Since the dispersion matrix itself is not identified, the diffusion matrix $A(x) = \Sigma(x)\Sigma(x)'$ with elements $a^{ij}(x) = \sum_{k=1}^{\widetilde{d}} \sigma^{ik}(x)\sigma^{kj}(x)$ is defined.

Standard assumptions that guarantee the existence of a strong solution of the stochastic differential equation (4.1) are the so-called global Lipschitz and linear growth conditions. These conditions ensure that the process does not explode and is unique. To retain the length of the proofs, it will furthermore be assumed that the solution is stationary and strongly mixing. Intuitively this can only be the case if the drift pulls the process back to its mean whenever the Brownian motion creates a large deviation. There are different sets of assumptions on the drift and diffusion that ensure this. For instance Veretennikov (1997) provides the following condition: There exist constants $r > 0, C \geq 0$ such that

$$\Big\langle \mu(x), \frac{x}{\|x\|} \Big\rangle < -\frac{r}{\|x\|} \quad \text{for } \|x\| \geq C$$

and $(r - (d\Lambda - \lambda_-)/2)/\lambda_+ > 3/2$ where

$$\lambda_- = \inf_{x\neq 0} \Big\langle A(x)\frac{x}{\|x\|}, \frac{x}{\|x\|} \Big\rangle, \quad \lambda_+ = \sup_{x\neq 0} \Big\langle A(x)\frac{x}{\|x\|}, \frac{x}{\|x\|} \Big\rangle,$$
$$\Lambda = \sup_x trA(x)/d.$$

Of course the process can only be stationary if the initial random variable X_0 already follows the stationary distribution. The process is completely characterized by the drift and the diffusion functions and in particular for the stationary density $f(x)$ it holds

$$\mu^j(x)f(x) = \frac{1}{2}\sum_{i=1}^{d} \frac{\partial(a^{ij}(x)f(x))}{\partial x^i},$$

for all $j = 1, \dots, d$. Then it is no problem to assume that for given drift and diffusion functions the initial random variable X_0 is distributed with density $f(x)$.

The assumption of stationarity can be relaxed to assume that the process is recurrent. This guarantees that (X_t) returns to any state infinitely often, which enables the local estimation using almost uncorrelated observations. Extending the results of this chapter to this more general class of processes can be done by using the proofing techniques of Bandi and Moloche (2001) and Schienle (2006).

Assume that the process is observed at $nT + 1$ equispaced time points in the interval $[0, T]$. Defining the distance between subsequent observations with $\Delta = n^{-1}$, the observations are denoted with $X_{k\Delta}, k = 0, 1, \dots, nT$. This setting allows to study different sampling schemes. High frequency sampling is considered if the sampling interval shrinks to zero, i. e. $n \to \infty$ (or $\Delta \to 0$). Then, the nonparametric estimation can be based on the property of the conditional expectation operator

$$\lim_{\Delta \to 0} \mathbf{E}(\Delta^{-1}(X^j_{(k+1)\Delta} - X^j_{k\Delta}) \mid X_{k\Delta} = x) = \mu^j(x) \tag{4.2}$$

$$\lim_{\Delta \to 0} \mathbf{E}(\Delta^{-1}(X^i_{(k+1)\Delta} - X^i_{k\Delta})(X^j_{(k+1)\Delta} - X^j_{k\Delta}) \mid X_{k\Delta} = x) = a^{ij}(x). \tag{4.3}$$

Thus, estimators are given by regressing the increments of the process (resp. their products) onto the state. For example a classical Nadaraya-Watson estimator of the drift function is given by

$$\widehat{\mu}^{j,NW}_h(x) = \frac{\frac{1}{Tn}\sum_{k=0}^{nT-1} K_h(x, X_{k\Delta})\Delta^{-1}(X^j_{(k+1)\Delta} - X^j_{k\Delta})}{\frac{1}{Tn}\sum_{k=0}^{nT-1} K_h(x, X_{k\Delta})} \tag{4.4}$$

with a kernel weight $K_h(x, X_{k\Delta}) = \prod_{i=1}^d K_h(x^i, X_{k\Delta})$. For simplicity of notation it is assumed that the same bandwidth is used for all dimension and it will be denoted with h. The corresponding estimators of the entries of the diffusion matrix are obtained by replacing $\Delta^{-1}(X^j_{(k+1)\Delta} - X^j_{k\Delta})$ with $\Delta^{-1}(X^i_{(k+1)\Delta} - X^i_{k\Delta})(X^j_{(k+1)\Delta} - X^j_{k\Delta})$. In the scalar case ($d = 1$) this estimator of the diffusion was first considered by Florens-Zmirou (1993) for a fixed time horizon. A bivariate extension was proposed by Brugiere (1993). In that case the diffusion function can be estimated with mixed asymptotic ($n \to \infty, T = \bar{T}$ fixed) normality. If the time horizon tends to infinity as well, asymptotic normality of the estimator holds and the rate is given by $\sqrt{nTh^d}$, which can be found in Bandi and Moloche (2001).

The drift functions, in contrast, are not estimable over a fixed time horizon. If both n and T tend to infinity, the estimator is asymptotically normal and the rate of convergence is given by $\sqrt{Th^d}$ (see Bandi and Moloche, 2001).

In finite samples the estimation based on the conditional expectation operator suffers from a bias of order Δ associated with the sampling frequency. Using the infinitesimal generator of the process, Stanton (1997) and Fan and Zhang (2003) introduce different approximation schemes, that result in a bias of order Δ^k. As they point out, the increasing precision has to be paid with a larger variance of the resulting estimators.

Various articles have extended the basic framework to nonstationary but recurrent processes (Bandi and Philips, 1998), low frequency sampling (Gobet, Hoffmann and Reiß, 2004), local polynomial estimators (Fan and Zhang, 2003, Moloche, 2001) or jump diffusions (Bandi and Nguyen, 2003).

The results of Bandi and Moloche (2001) indicate the presence of the well known curse of dimensionality, which means that for the estimation of higher dimensional processes the rate of convergence becomes slower and the sample sizes have to be larger. One possibility to circumvent this problem is to impose more structure on the unknown functions but to keep them still more flexible than in parametric specifications.

Additivity of the drift functions means that one or all elements of the drift vector are assumed to be fully additive, i. e.

$$\mu^j(x) = \mu^{j,1}(x^1) + \cdots + \mu^{j,d}(x^d).$$

Analogously, additivity of the diffusion functions means that some or all elements of the diffusion matrix are fully additive in its arguments. Using an appropriate estimation technique, it can be possible to estimate the one-dimensional components $\mu^{i,j}(x^j)$ with the one-dimensional rate of convergence.

4.3 The Smooth Backfitting Algorithm

For the nonparametric estimation of additive functions in a classical regression setting, different estimators have been proposed. The most prominent smoothing based techniques are the classical backfitting algorithm by Buja, Hastie and Tibshirani (1989), marginal integration by Linton and Nielsen (1995) and Tjøstheim and Auestad (1994), smooth backfitting by Mammen, Linton and Nielsen (1999) and local partitioned regression by Christopeit and Hoderlein (2006). Marginal integration and local partitioned regression use a full-dimensional estimator in a first stage and therefore these methods suffer from the curse of dimensionality in the sense that the sample size has to increase with d (but the rate of convergence is one-dimensional). Opsomer and Ruppert (1997) have investigated the asymptotic properties of classical backfitting and found out that the algorithm is not oracle-efficent. Furthermore, the correlation between the covariates is restricted in their analysis, which is an important drawback in the present application of diffusion processes. In contrast, smooth backfitting was shown to be fully oracle efficient in a standard regression problem. Therefore this algorithm is chosen for the estimation of diffusion processes in this chapter.

The algorithm will be presented for an estimator of $\mu^1(x)$. Other components of the drift and the diffusion follow by appropriately changing the response variable according to equations (4.2) and (4.3). The additive model for $\mu^1(x)$ is given by

$$\lim_{\Delta\to 0} \mathbf{E}(\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) \mid X_{k\Delta} = x) = \mu^{1,0} + \mu^{1,1}(x^1) + \cdots + \mu^{1,d}(x^d). \tag{4.5}$$

Without an additional constraint constants could be interchanged between the additive components and they would not be identified. Therefore

$$\int \mu^{1,j}(x^j) f(x^j)\, \mathrm{d}x^j = 0, \qquad j = 1, \ldots, d \tag{4.6}$$

is imposed where $f(x^j) = \int f(x)\, \mathrm{d}x^{-j}$ denotes[1] the marginal density of X^j. To estimate an unknown regression function by kernel smoothing, local polynomials of different order can be used. Usual considerations are local constant (Nadaraya-Watson) or local linear estimators. In this section smooth backfitting estimators of the additive model based on these popular estimators are described.

4.3.1 Smooth Backfitting Based on Local Constant Estimation

The classical Nadaraya-Watson estimator as explicitly given in equation (4.4) can be obtained as solution of the minimization problem

$$\widehat{\mu}^{1,NW}_h = \arg\min_{\bar{\mu}^1 \in \mathcal{M}} \int \frac{1}{nT} \sum_{k=0}^{nT-1} \big(\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) - \bar{\mu}^1(x)\big)^2 \prod_{j=1}^{d} K_h(x^j, X^j_{k\Delta})\, \mathrm{d}x, \tag{4.7}$$

[1] Here the notation $\mathrm{d}x^{-j} = \mathrm{d}x^1 \ldots \mathrm{d}x^{j-1}\, \mathrm{d}x^j \ldots \mathrm{d}x^d$ is introduced.

where the minimization runs over an appropriate function class $\mathcal{M}$ in the space of square integrable functions. A natural way to obtain an estimator of the additive model would be to restrict the minimization to the class of additive functions $\mathcal{M}^{\text{add}} = \{\bar{\mu}^1 \in \mathcal{M} : \bar{\mu}^1(x) = \bar{\mu}^{1,0} + \bar{\mu}^{1,1}(x^1) + \cdots + \bar{\mu}^{1,d}(x^d)\}$. By a simple projection argument it holds that

$$\begin{aligned}\int \frac{1}{nT}\sum_{k=1}^{nT}\big(\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) - \bar{\mu}^1(x)\big)^2 \prod_{j=1}^{d} K_h(x^j, X^j_{k\Delta})\,\mathrm{d}x \\ = \int \frac{1}{nT}\sum_{k=1}^{nT}\big(\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) - \widehat{\mu}^{1,NW}(x)\big)^2 \prod_{j=1}^{d} K_h(x^j, X^j_{k\Delta})\,\mathrm{d}x \\ + \int \frac{1}{nT}\sum_{k=1}^{nT}\big(\widehat{\mu}^{1,NW}(x) - \bar{\mu}^1(x)\big)^2 \prod_{j=1}^{d} K_h(x^j, X^j_{k\Delta})\,\mathrm{d}x.\end{aligned}$$

Obviously, minimizing the left hand side over $\bar{\mu}^1$ is equivalent to minimizing the second term on the right hand side over $\bar{\mu}^1$. This second term can be written as

$$\|\widehat{\mu}_h^{NW} - \bar{\mu}\|_{\widehat{f}} = \int \Big(\widehat{\mu}_h(x) - \bar{\mu}^{1,0} - \bar{\mu}^{1,1}(x^1) - \cdots - \bar{\mu}^{1,d}(x^d)\Big)^2 \widehat{f}_h(x)\,\mathrm{d}x \tag{4.8}$$

where $\widehat{f}_h(x) = \sum_{k=0}^{nT-1}\prod_{i=1}^{d} K_h(x^i, X^i_{k\Delta})$ is a kernel density estimator. Interpreting this equation, the smooth backfitting estimators can be regarded as the projection of the full-dimensional Nadaraya-Watson estimator onto the space of additive functions under the semi-norm induced by $\widehat{f}_h(x)$. The projection interpretation of the estimators is discussed in more detail in Mammen, Linton and Nielsen (1999) and more generally in Mammen et al. (2001).

To ensure identifiability, the minimization is restricted to the empirical version of equation (4.6)

$$\int \widehat{\mu}^{1,j,NW}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j = 0, \tag{4.9}$$

for $j = 1, \dots, d$, using marginal density estimators $\widehat{f}_h(x^j) = \sum_{k=0}^{nT-1} K_h(x^j, X^j_{k\Delta})$. Solving the minimization (4.8) with respect to (4.9) the minimum $(\widetilde{\mu}^{0,1,NW}, \widetilde{\mu}^{1,1,NW}(x^1), \dots, \widetilde{\mu}^{1,d,NW}(x^d))$ is not given explicitly but as solution of the set of equations

$$\widetilde{\mu}_h^{1,j,NW}(x^j) = \int \widehat{\mu}_h^{1,NW}(x)\frac{\widehat{f}_h(x)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^{-j} - \sum_{i\neq j}\int \widetilde{\mu}_h^{1,i,NW}(x^i)\frac{\widehat{f}_h(x)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^{-j} - \widetilde{\mu}^{1,0}, \tag{4.10}$$

for $j = 1, \dots, d$ together with (4.9). The two integrals in equation (4.10) can be simplified to

$$\begin{aligned}&\int \widehat{\mu}_h^{1,NW}(x)\frac{\widehat{f}_h(x)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^{-j}\\ &\quad= \frac{\frac{1}{nT}\int\sum_{k=0}^{nT-1}\prod_{i=1}^d K_h(x^i, X^i_{k\Delta})\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta})\,\mathrm{d}x^{-j}}{\widehat{f}_h(x^j)}\\ &\quad= \widehat{\mu}_h^{1,j,NW}(x^j),\end{aligned}$$

which are subsequently called marginal Nadaraya-Watson estimators and

$$\int \widetilde{\mu}_h^{1,i,NW}(x^i)\frac{\widehat{f}_h(x)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^{-j} = \int \widetilde{\mu}_h^{1,i,NW}(x^i)\frac{\widehat{f}_h(x^i, x^j)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^i.$$

Using these transformations the equations (4.10) can be rewritten to

$$\widetilde{\mu}_h^{1,j,NW}(x^j) = \widehat{\mu}_h^{1,j,NW}(x^j) - \sum_{i\neq j}\int \widetilde{\mu}_h^{1,i,NW}(x^i)\frac{\widehat{f}_h(x^i, x^j)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^i - \widetilde{\mu}_j^{1,0,NW}. \tag{4.11}$$

Then, the Nadaraya-Watson smooth backfitting estimators as solutions to the equations (4.11) together with the normalizations (4.9). These equations can directly be motivated by noting that the marginal Nadaraya-Watson estimators will converge to

$$\mathbf{E}(\mu^1(X) \mid X^j = x^j) = \mu^{1,0} + \mu^{1,j}(x^j) + \sum_{i\neq j}\mathbf{E}(\mu^{1,i}(X^i) \mid X^j = x^j).$$

To obtain the additive components, the conditional expectation operator has to be inverted and equations (4.11) provide an empirical version of the integral equation. In that sense smooth backfitting estimation is a (well-posed) statistical inverse problem.

In the algorithm only marginal Nadaraya-Watson estimators $\widehat{\mu}_h^{1,j,NW}(x^j)$ and one- and two-dimensional kernel density estimators $\widehat{f}_h(x^j)$ and $\widehat{f}_h(x^i, x^j)$ are used. Since no higher-dimensional kernel regression or density estimators are calculated, the estimation procedure does not suffer from the curse of dimensionality.

To compute the estimators, marginal Nadaraya-Watson estimators and the one- and two-dimensional kernel density estimators have to be calculated for a number of grid points that allow to evaluate the integrals in equation (4.11) numerically. Using as starting values the marginal Nadaraya-Watson estimators the smooth backfitting estimators are derived as the iterative solution of (4.11) and (4.9).

More practical details about the implementation can be found in Nielsen and Sperlich (2005).

If it holds for the kernel density estimators that

$$\int_{\mathcal{G}^i} \widehat{f}_h(x^i, x^j)\,\mathrm{d}x^i = \widehat{f}_h(x^j), \tag{4.12}$$

for all combinations of i and j and where $\mathcal{G}^i$ is the bounded support of X^i then the normalizations (4.9) are automatically fulfilled by choosing $\widetilde{\mu}_j^{1,0,NW} = \int \widehat{\mu}_h^{1,j,NW}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j = (nT)^{-1}\sum_{k=0}^{nT-1}\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) = T^{-1}(X^1_T - X^1_0)$. Thus, the normalization can be omitted from the algorithm if centered data is used.

One possibility to ensure (4.12) is to use modified kernels

$$K_h(u,v) = \frac{K_h(u-v)}{\int_{\mathcal{G}^j} K_h(w-v)\,\mathrm{d}w}, \tag{4.13}$$

where and K is a usual kernel function with support $[-1,1]$, say. However, restriction (4.12) is always violated if unmodified kernels are used or if one is interested in estimating the function $\mu^1(x)$ over a compact set $\mathcal{G}$ that is not rectangular. In that case it is still possible to use $\widetilde{\mu}^{1,0,NW} = T^{-1}(X^1_T - X^1_0)$. But now in each iteration step the centering condition (4.9) has to be updated. That means in the r-th iteration cycle, the update $\widetilde{\widetilde{\mu}}_h^{1,j,NW,[r]}(x^j)$ of equation (4.11) has to be recentered to obtain

$$\widetilde{\mu}_h^{1,j,NW,[r]} = \widetilde{\widetilde{\mu}}_h^{1,j,NW,[r]}(x^j) - \frac{\int \widetilde{\widetilde{\mu}}_h^{1,j,NW,[r]}(x^j)\widehat{f}(x^j)\,\mathrm{d}x^j}{\int \widehat{f}_h(x^j)\,\mathrm{d}x^j}.$$

But for this more general algorithm, the asymptotic theory is still not completely solved. While convergence of the algorithm can be shown, the bias behavior of the estimator has not been fully understood. This chapter concentrates therefore on the standard setting, where $f(x) > 0$ for $x \in \mathcal{G}$, which is the cross product of compact sets $\mathcal{G}^j$, on which the marginal distribution of X^j is bounded from below.

Mammen, Linton and Nielsen (1999) provide general conditions under which a unique solution of the algorithm exists and show that it converges with geometric rate with probability tending to one. These conditions are very general and in particular neither assume the additive model to hold nor restrict the underlying data generating process. The convergence of the algorithm will be established as a by-product of the asymptotic normality of the estimators in Section 4.4.

4.3.2 Smooth Backfitting Based on Local Linear Estimation

The minimization problem (4.7) can be extended by approximating the unknown function locally by a Taylor polynomial of higher order. This leads to the well known definition of local polynomial estimators. The local linear estimators are thus defined as

$$\begin{aligned}(4.14)\quad \widehat{\boldsymbol{\mu}}_h^{1,LL} = \arg \min_{(\bar{\mu}^1,\bar{\mu}_1,\dots,\bar{\mu}_d)\in\boldsymbol{\mathcal{M}}} \int \frac{1}{n}\sum_{k=0}^{nT-1}\big(\Delta^{-1}(X^1_{(k+1)\Delta}-X^1_{k\Delta})\\ -\bar{\mu}^1(x)-\sum_{j=1}^d \bar{\mu}^1_j(x)\frac{X^j_{k\Delta}-x^j}{h}\big)^2\prod_{j=1}^d K_h(x^j,X^j_{k\Delta})\,\mathrm{d}x,\end{aligned}$$

where $\boldsymbol{\mathcal{M}} = \mathcal{M}^{d+1}$. The quantity $h\mu^1_j(x)$ can be interpreted as the partial derivative of $\mu^1(x)$ with respect to x^j. The vector $\widehat{\boldsymbol{\mu}}_h^{1,LL}(x)$ then consists of an estimator of the function and of estimators of all partial derivatives. Introducing the matrices

$$\begin{aligned}\mathbf{X}(x) &= \begin{pmatrix} 1 & (X^1_0-x^1)/h & \dots & (X^d_0-x^d)/h \\ \vdots & & \ddots & \vdots \\ 1 & (X^1_{T-\Delta}-x^1)/h & \dots & (X^d_{T-\Delta}-x^d)/h \end{pmatrix}\\ \mathbf{K}(x) &= \frac{1}{NT}\big(\mathrm{diag}\Big(\prod_{j=1}^d K_h(X^j_\Delta,x^j),\dots,\prod_{j=1}^d K_h(X^j_{T-\Delta},x^j)\Big)\big)\\ \mathbf{Y} &= \big(\Delta^{-1}(X_\Delta-X_0),\dots,\Delta^{-1}(X_T-X_{T-\Delta})\big),\end{aligned}$$

the estimator can be written as

$$\widehat{\boldsymbol{\mu}}_h^{1,LL}(x) = \widehat{\mathbf{S}}^{-1}(x)\widehat{\mathbf{T}}(x),$$

where $\widehat{\mathbf{S}}(x) = \mathbf{X}^T(x)\mathbf{K}(x)\mathbf{X}(x)$ and $\widehat{\mathbf{T}}(x) = \mathbf{X}^T(x)\mathbf{K}(x)\mathbf{Y}$. The matrix $\widehat{\mathbf{S}}(x)$ contains kernel density estimators and for further reference their one- and two-dimensional marginals are introduced

$$\begin{aligned}\widehat{f}_{h,l}(x^j) &= \frac{1}{nT}\sum_{k=0}^{nT-1}K_h(x^j,X^j_{k\Delta})(X^j_{k\Delta}-x^j)^l \qquad \text{for } l=1,2\\ \widehat{f}_{h,\mathbf{l}}(x^i,x^j) &= \frac{1}{nT}\sum_{k=0}^{nT-1}K_h(x^i,X^i_{k\Delta})K_h(x^j,X^j_{k\Delta})(X^i_{k\Delta}-x^i)^{l_1}(X^j_{k\Delta}-x^j)^{l_2}\\ &\qquad \text{for } \mathbf{l}=(l_1,l_2)\in\{(1,0),(0,1),(1,1)\}.\end{aligned}$$

Following the same way as in the local constant case, the minimization problem (4.14) is restricted to the subset of additive functions $\boldsymbol{\mathcal{M}}^{\text{add}} = \{\bar{\boldsymbol{\mu}}^1(x) \in \boldsymbol{\mathcal{M}} \mid \bar{\mu}^1 \in \mathcal{M}^{\text{add}}, \bar{\mu}_j \colon \mathbb{R} \to \mathbb{R}$ does only depend on $x^j\}$ in order to obtain an estimator. Obviously $\boldsymbol{\mathcal{M}}^{\text{add}} \subset \boldsymbol{\mathcal{M}}$ and note that the j-th partial derivative only depends on x^j. By projection arguments it follows that this is equivalent to minimizing

$$\|\widehat{\boldsymbol{\mu}}_h^{1,LL}(x) - \bar{\boldsymbol{\mu}}^1(x)\|_{\widehat{\mathbf{S}}} = \int (\widehat{\boldsymbol{\mu}}_h^{LL}(x) - \bar{\boldsymbol{\mu}}^1(x))^T \widehat{\mathbf{S}}(x)(\widehat{\boldsymbol{\mu}}_h^{LL}(x) - \bar{\boldsymbol{\mu}}^1(x))\,\mathrm{d}x, \tag{4.15}$$

where $\bar{\boldsymbol{\mu}}^1(x) = (\bar{\mu}^1(x), \bar{\mu}_1^1(x^1), \ldots, \bar{\mu}_d^1(x^d))^T$ is an element of $\boldsymbol{\mathcal{M}}^{\text{add}}$. Again, the estimator can be regarded as a projection of the full-dimensional local linear estimator $\widehat{\boldsymbol{\mu}}_h^{1,LL}$ onto the space of additive functions with respect to the inner product induced by $\widehat{\mathbf{S}}(x)$. This is the analogous interpretation of the smooth backfitting estimator as a projection, as in the local constant case of this last subsection but this time with a different space and a different norm. To derive the solution of the minimization problem (4.15) the argumentation becomes slightly more complex than in the Nadaraya-Watson case. It is skipped here and the interested reader is referred to Mammen, Linton and Nielsen (1999). Finally the local linear smooth backfitting estimator $(\widetilde{\mu}_h^{1,1,LL}(x^1), \widetilde{\mu}_{1,h}^{1,1,LL}(x^1), \ldots, \widetilde{\mu}_h^{1,d,LL}(x^d), \widetilde{\mu}_{d,h}^{1,d,LL}(x^d))$ is defined as the solution of the following set of equations

$$\begin{aligned}\begin{pmatrix}\widetilde{\mu}_h^{1,j,LL}(x^j)\\ \widetilde{\mu}_{j,h}^{1,j,LL}(x^j)\end{pmatrix} &= \begin{pmatrix}\widehat{\mu}_h^{1,j,LL}(x^j)\\ \widehat{\mu}_{j,h}^{1,j,LL}(x^j)\end{pmatrix} - \begin{pmatrix}\widetilde{\mu}_j^{1,0,LL}\\ 0\end{pmatrix}\\ &\quad - \widehat{V}^j(x^j)^{-1}\sum_{i\neq j}\int \widehat{U}^{ij}(x^i,x^j)\begin{pmatrix}\widetilde{\mu}_h^{1,i,LL}(x^i)\\ \widetilde{\mu}_{i,h}^{1,i,LL}(x^i)\end{pmatrix}\mathrm{d}x^i,\end{aligned} \tag{4.16}$$

for $j = 1, \ldots, d$ together with the normalizations

$$\int \widehat{\mu}_h^{1,j,LL}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j + \int \bar{\mu}_{j,h}^{1,j,LL}(x^j)\widehat{f}_{h,1}(x^j)\,\mathrm{d}x^j = 0. \tag{4.17}$$

which ensure identification. Note that this is asymptotically also a consistent version of the original restriction (4.6). The matrices in (4.16) are defined as

$$\widehat{V}^j(x^j) = \begin{pmatrix}\widehat{f}_h(x^j) & \widehat{f}_{h,1}(x^j)\\ \widehat{f}_{h,1}(x^j) & \widehat{f}_{h,2}(x^j)\end{pmatrix}$$

$$\widehat{U}^{ij}(x^i,x^j) = \begin{pmatrix}\widehat{f}_h(x^i,x^j) & \widehat{f}_{h,(1,0)}(x^i,x^j)\\ \widehat{f}_{h,(0,1)}(x^i,x^j) & \widehat{f}_{h,(1,1)}(x^i,x^j)\end{pmatrix},$$

and $(\widehat{\mu}_h^{1,J,LL}(x^j), \widehat{\mu}_{j,h}^{1,j,LL}(x^j))$ are the marginal (one-dimensional) local linear estimators of the regression of $(X_{(k+1)\Delta}^1 - X_{k\Delta}^1)$ on $X_{k\Delta}^j$. As in the local constant case only one- and two-dimensional smoothers are used and therefore the estimator does not suffer from the curse of dimensionality. Using the marginal local linear estimators as starting values the smooth backfitting estimators are obtained as the iterative solution of (4.16). For computational purposes the representation (4.16) is not very convenient. Nielsen and Sperlich (2005) describe the implementation in detail if modified kernels are used (i. e. equation (4.12) holds). In that case direct calculations show that the normalization is directly achieved for $\widetilde{\mu}_j^{1,0,LL} = T^{-1}(X_T^1 - X_0^1)$ for all $j = 1, \dots, d$. If (4.12) does not hold, normalization is achieved by choosing

$$\widetilde{\mu}_j^{1,0} = \Big(\int \widetilde{\mu}_h^{1,j,LL}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j + \int \widetilde{\mu}_{j,h}^{1,j,LL}(x^j)\widehat{f}_{h,1}(x^j)\,\mathrm{d}x^j\Big)\Big(\int \widehat{f}_h(x^j)\,\mathrm{d}x^j\Big)^{-1}$$

and implementing the normalization as in the local constant case. As pointed out, the algorithm can be regarded as a projection method and Mammen, Linton and Nielsen (1999) provide general conditions under which the iterative procedure converges as well as properties of the limit. In the next section limit results for the estimation of the components of a diffusion process will be derived.

4.4 Asymptotic Results

After the presentation of the basic algorithms the asymptotic behavior of the estimators will be derived. First, estimation of the drift vector is considered and the two backfitting methods (Nadaraya-Watson and local linear) are compared via their oracle properties. First, the required assumptions are stated.

Assumption 4.1. *1. The elements of the drift vector $\mu(x)$ and the diffusion matrix $A(x)$ are twice continuously differentiable.*

2. There exists a solution to the stochastic differential equation (4.1) *and the process (X_t) is stationary, has compact support $\mathcal{G} = \mathcal{G}^1 \times \cdots \times \mathcal{G}^d$ and is strongly mixing with α-mixing coefficients satisfying $\sum_{i=1}^{\infty} \alpha(i)^{1/2} < \infty$. The stationary density $f(x)$ is twice continuously differentiable and the marginal densities $f(x^j)$ are bounded from below on $\mathcal{G}^j$.*

It is not very natural to assume a process that lives on a compact support. But this has to be done for technical reasons only. Consider an arbitrary stationary strongly mixing process $(\widetilde{X})_t$, it can be transformed into a process satisfying the assumption. For this purpose select sets $\mathcal{G}^j$ that fulfil Assumption 4.1.

Then, a new process can be defined by excluding all observations when $(\widetilde{X})_t$ is outside $\mathcal{G}$ and taper the remaining parts together. Formally this is done by defining the time-changed process $X_t = \widetilde{X}_{\tau^{-1}(t)}$ where τ^{-1} is the inverse of $\tau(t) = \int_0^t \mathbb{1}_{\{\widetilde{X}_s \in \mathcal{G}\}} \,\mathrm{d}s$. Then, the new process satisfies Assumption 4.1.

Finally, standard assumptions on the kernel function have to be employed.

Assumption 4.2. *For a bandwidth sequence $h \to 0$ the kernel weights are given by equation* (4.13). *The kernel function $K(u)$ is a positive symmetric (around 0) and bounded function with compact support that integrates to one. $u^k K(u)$ is Lipschitz continuous for $k = 0, 1, 2$. The kernel is of order 2, i. e. $\int u^2 K(u)\,\mathrm{d}u < \infty$.*

Note that for the use of modified kernels, extra attention has to be paid to the kernel moments. For simplicity assume $\mathcal{G}^j = [0, 1]$. The numerator of $K_h(u, v)$ is equal to one if $v \in [h, 1-h]$ and depends on v (but not on h) otherwise. Kernel constants are defined as

$$\kappa_l(u) = \int_0^1 (u-v)^l \frac{K_h(u-v)}{\int_0^1 K_h(w-v)\,\mathrm{d}w} \,\mathrm{d}v.$$

Easy calculations show that three cases have to be distinguished

$$\kappa_l(u) = \begin{cases} \int_{-1}^1 v^l K(v)\,\mathrm{d}v & \text{for } u \in [2h, 1-2h] \\ \int_{-1}^1 v^l K(v)\,\mathrm{d}v + O(h^{l+1}) & \text{for } u \in [h, 2h] \cup [1-2h, 1-h] \\ \int_0^1 (u-v)^l K_h(u-v)\,\mathrm{d}v + O(h^{l+1}) & \text{for } u \in [0, h] \cup [1-h, 1] \end{cases}.$$

The modified kernels only have an influence at boundary points $u \in [0, 2h] \cup [1-2h, 1]$, where they differ from usual kernel constants. Analogously, kernel constants $\kappa_l^2 = \int_0^1 (u-v)^l (K_h(u,v))^2 \,\mathrm{d}v$ are defined.

4.4.1 Estimation of the Drift Function

Without loss of generality the exposition is restricted to the case of estimating the first component of the drift vector $\mu^1(x)$. The Nadaraya-Watson smooth backfitting estimators $\widetilde{\mu}_h^{1,j}(x^j), j = 1, \ldots, d$ are defined as the iterative solution of the set of equations (4.11) and the normalization (4.9). Their asymptotic properties are given in the following

Theorem 4.1. *Let Assumptions 4.1 and 4.2 be fulfilled and the additive model* (4.5) *with centering* (4.6) *hold. For the bandwidth sequence it holds that $h^2 =$*

$O((Th)^{-1/2})$ and $nh^3 \to \infty$. Then the algorithm (4.11) *converges with geometric rate and for the estimators $\widetilde{\mu}_h^{1,j,NW}(x^j), j = 1, \dots, d$ it holds that*

$$\sqrt{Th}\frac{\widetilde{\mu}_h^{1,j,NW}(x^j) - \mu^{1,j}(x^j) - b_h^{1,j} - \beta_\mu^{1,j,NW}(x^j)}{\sqrt{v^1(x^j)\kappa^2(x^j)/\kappa_0(x^j)^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\begin{aligned} b_h^{1,j} &= -\iint \mu^{1,j}(x^j)K_h(x^j,u^j)f(u^j)\,\mathrm{d}u^j\,\mathrm{d}x^j \\ &\quad - h\iint \frac{\partial}{\partial x^j}\mu^{1,j}(x^j)\frac{\kappa_1(x^j)}{\kappa_0(x^j)}K_h(x^j,u^j)f(u^j)\,\mathrm{d}x^j \\ \beta_\mu^{1,j,NW}(x^j) &= h\frac{\kappa_1(x^j)}{\kappa_0(x^j)}\frac{\partial}{\partial x^j}\mu^{1,j}(x^j) + h^2\widetilde{\beta}_\mu^{1,j}(x^j) \end{aligned}$$

and

$$v^1(x^j) = (f^j(x^j))^{-1}\,\mathbf{E}(a^{11}(X) \mid X^j = x^j).$$

Note that the first part of the bias $\beta_\mu^{1,j,NW}(x^j)$ is zero for $x^j \in [h, 1-h]$ and therefore only present at the boundary. The second part is not given in explicit form, it is only defined as

$$\begin{aligned} &(\widetilde{\beta}_\mu^{1,0}, \widetilde{\beta}_\mu^{1,j}(x^1), \dots, \widetilde{\beta}_\mu^{d,j}(x^d)) \\ &\qquad = \arg\min_{\beta_\mu^{1,0},\dots,\beta_\mu^{1,d}} \int (\beta_\mu^1(x) - \beta_\mu^{1,0} - \beta_\mu^{1,1}(x^1) - \dots - \beta_\mu^{1,d}(x^d))^2 f(x)\,\mathrm{d}x, \end{aligned}$$

with

$$\beta_\mu(x) = \frac{\kappa_2(x^j)}{\kappa_0(x^j)}\sum_{j=1}^d (f(x))^{-1}\frac{\partial}{\partial x^j}(\mu^{1,j}(x^j))\frac{\partial}{\partial x^j}(f(x)) + \frac{1}{2}\frac{\partial}{\partial (x^j)^2}\mu^{1,j}(x^j).$$

Therefore the bias can be interpreted as the projection of $\beta_\mu(x)$ on the space of additive functions with respect to the $L_2(f)$-norm.

The term $b_h^{1,j}$ converges to zero asymptotically since it holds that

$$\begin{aligned} &\iint \mu^{1,j}(x^j)f(u^j)K_h(x^j,u^j)\,\mathrm{d}u^j\,\mathrm{d}x^j \\ &\qquad = \int_{u^j \notin [h,1-h]} \int \mu^1(x^j)f^j(x^j)(K_h(x^j,u^j) - K_h(x^j - u^j))\,\mathrm{d}x^j\,\mathrm{d}u^j + O(h^2) \\ &\qquad = O(h). \end{aligned}$$

and the second term is of order $O(h^2)$ because $\kappa_1(x^j)$ is zero at interior points x^j. However, this term is constant over x^j and does therefore only affect the normalization of $\mu^{1,j}(x^j)$. It is generated by the difference between the empirical normalization (4.9) used in the algorithm and the theoretical normalization (4.6) and does not influence the shape of the estimator.

Convergence of the algorithm follows from consistency of the (one and two-dimensional) kernel density estimators. In particular, the unknown function do not have to be additive. If the additive model does not hold, the estimators will converge to a projection of the high-dimensional function onto the space of additive functions. In Chapter 5 this case is investigated for independent and identically distributed data.

The limit distribution of the vector $(\widetilde{\mu}_h^{1,1,NW}(x^1), \dots, \widetilde{\mu}_h^{1,d,NW}(x^d))$ is a multivariate (d-dimensional) normal distribution where the covariances are zero asymptotically. Considering the joint estimation of the additive components of $\mu^i(x)$ and $\mu^{i'}(x)$ there are asymptotically non-vanishing covariances, given by

$$\mathbf{cov}(\sqrt{hT}\widetilde{\mu}_h^{i,j,NW}(x^j), \sqrt{hT}\widetilde{\mu}_h^{i',j,NW}(x^j)) = \frac{\kappa^2(x^j)}{f(x^j)}\,\mathbf{E}(a^{ii'}(X) \mid X^j = x^j).$$

To judge the efficiency of the Smooth backfitting estimator it has to be compared to the oracle estimator, which is based on knowledge of all other $\mu^{1,i}(x^i), i \neq j$. With this knowledge, the response variables could be modified to

$$\begin{aligned} Y^{\star}_{k\Delta} &= \Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta}) - \sum_{i\neq j} \mu^{1,i}(X^i_{k\Delta}) \\ &= \int_{k\Delta}^{(k+1)\Delta} \mu^{1,j}(X^j_s)\,\mathrm{d}s + \sum_{i=1}^{d} \int_{k\Delta}^{(k+1)\Delta} \sigma^{1i}(X_s)\,\mathrm{d}W^l_s \\ &\quad + \sum_{i\neq j} \int_{k\Delta}^{(k+1)\Delta} (\mu^{1,i}(X^i_s) - \mu^{1,i}(X^i_{k\Delta}))\,\mathrm{d}s. \end{aligned}$$

Then, the infeasible oracle estimator is given by

$$\breve{\mu}_h^{1,j,NW}(x^j) = \frac{\sum_{k=0}^{nT-1} K_h(x^j, X^i_{k\Delta}) Y^{\star}_{k\Delta}}{\sum_{k=0}^{nT-1} K_h(x^j, X^i_{k\Delta})}.$$

The knowledge of the other components allows to estimate $\mu^{1,j}(x^j)$ from discrete data only. The discretization errors are of order $O_P(n^{-3/2})$ and therefore do not affect the estimation asymptotically. Therefore it holds that (even under weaker assumptions than Theorem 4.1)

$$\sqrt{Th}\frac{\breve{\mu}_h^{1,j,NW}(x^j) - \mu^{1,j}(x^j) - \breve{\beta}^{1,j}(x^j)}{\sqrt{\kappa^2(x^j)v^1(x^j)/\kappa_0(x^j)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\begin{aligned}\check{\beta}^{1,j}(x^j) = h\frac{\kappa_1(x^j)}{\kappa_0(x^j)}\frac{\partial}{\partial x^j}(\mu^{1,j}(x^j)) \\ + h^2\left(\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\frac{\frac{\partial}{\partial x^j}(\mu^{1,j}(x^j))\frac{\partial}{\partial x^j}(f(x^j))}{f(x^j)} + \frac{1}{2}\frac{\partial^2}{\partial(x^j)^2}\mu^{1,j}(x^j)\right).\end{aligned}$$

This follows from Lemmata 4.2 and 4.4 in the appendix and therefore, the smooth backfitting estimator $\widetilde{\mu}_h^{1,j,NW}(x^j)$ achieves the same variance as the oracle estimator, but has a different bias. This is the same efficiency result as in the classical regression setting, which was shown by Mammen, Linton and Nielsen (1999). To understand the bias behavior recall that the smooth backfitting estimator can be regarded as a projection of the full dimensional Nadaraya-Watson estimator onto the space of additive functions. Theorem 4.1 shows that the bias of the smooth backfitting estimator is the additive projection of the bias of the full-dimensional estimator. But this is not additive because the stationary density of the process $f(x)$ is in general not additive. In contrast, the bias of a full-dimensional local linear estimator is additive and consists of the sum of the second derivatives of the additive components (times a constant). Smooth backfitting based on the local linear estimator can again be regarded as a projection of the full-dimensional local linear estimator. In the next theorem it will be shown that the bias of the local linear smooth backfitting estimator is again the projection of the bias of the full-dimensional estimator and therefore local linear backfitting is fully oracle efficient. The design independence of local linear estimation, which means that the bias is independent of the density of the regressors, carries over to the projected estimators and drives the efficiency result.

The next theorem states the asymptotic properties of the local linear smooth backfitting estimators, defined in equations (4.16) and (4.17).

Theorem 4.2. *Let Assumptions 4.1 and 4.2 be fulfilled and the additive model* (4.5) *with centering* (4.6) *hold. For the bandwidth sequence it holds that* $h^2 = O((Th)^{-1/2})$ *and* $nh^3 \to \infty$. *Then the algorithm* (4.16) *converges with geometric rate and for the estimators* $\widetilde{\mu}_h^{1,j,LL}(x^j), j = 1, \ldots, d$ *it holds that*

$$\sqrt{Th}\frac{\widetilde{\mu}_h^{1,j,LL}(x^j) - \mu^{1,j}(x^j) - b_h^{1,j} - \beta_\mu^{LL}(x^j)}{\sqrt{v^1(x^j)\widetilde{\kappa}(x^j)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

with $b_h^{1,j}$ and $v^1(x^j)$ as given in Theorem 4.1 and where

$$\beta_\mu^{LL}(x^j) = h^2 \frac{1}{2}\frac{\kappa_2(x^j)}{\kappa_0(x^j)} \frac{\partial^2}{\partial(x^j)^2}\mu^1(x^j)$$
$$\widetilde{\kappa}(x^j) = \frac{\kappa_0^2(x^j)\kappa_2(x^j) - \kappa_1(x^j)\kappa_1^2(x^j)}{\kappa_0(x^j)\kappa_2(x^j) - (\kappa_1(x^j))^2}.$$

For interior points, the variance reduces to $\widetilde{\kappa}(x^j) = \kappa_0^2(x^j)$ because all other kernel constants are zero or one. In contrast to local constant smooth backfitting, the bias is given in explicit form. To derive the oracle efficiency, consider the unfeasible local linear estimator based on the data $Y_{k\Delta}^*$. Applying Lemmata 4.2 and 4.4, the asymptotic properties of the oracle estimator (under Assumptions 4.1 and 4.2) are given by

$$\sqrt{Th}\frac{\breve{\mu}_h^{1,j,LL}(x^j) - \mu^{1,j}(x^j) - \beta_\mu^{LL}(x^j)}{\sqrt{v^1(x^j)\widetilde{\kappa}(x^j)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

From this it can be seen that both bias and variance are identical to the expressions in Theorem 4.2. Therefore the local linear estimators are fully oracle efficient.

4.4.2 Estimation of the Diffusion Function

Now, the estimation of the elements of the diffusion matrix $A(x)$ is considered. To avoid confusion with the increasing number of indices, the exposition is restricted to the case to estimate $a^{12}(x)$, which is assumed to be fully additive, i. e.

$$a^{12}(x) = a^{12,0} + a^{12,1}(x^1) + \cdots + a^{12,d}(x^d). \tag{4.18}$$

For identifiability it is imposed that $\int a^{12,j}(x^j)f(x^j)\,\mathrm{d}x^j = 0$ for all $j = 1, \ldots, d$. Based on equation (4.3) the marginal Nadaraya-Watson estimators are given by

$$\widehat{a}_h^{12,j,NW}(x^j) = \frac{\sum_{k=0}^{nT-1} K_h(x^j, X_{k\Delta}^j)\Delta^{-1}(X_{(k+1)\Delta}^1 - X_{k\Delta}^1)(X_{(k+1)\Delta}^2 - X_{k\Delta}^2)}{\sum_{k=0}^{nT-1} K_h(x^j, X_{k\Delta}^j)}$$

The local constant smooth backfitting estimators are defined by plugging these estimators into equation (4.11). Explicitly, the backfitting estimators $\widetilde{a}_h^{12,j,NW}(x^j)$, $j = 1, \ldots, d$ of the additive components of $a^{12}(x)$ are defined as the iterative solutions to

$$\widetilde{a}_h^{12,j,NW}(x^j) = \widehat{a}_h^{12,j,NW}(x^j) - \sum_{i \neq j} \int \widetilde{a}_h^{12,i,NW}(x^i)\frac{\widehat{f}_h(x^i, x^j)}{\widehat{f}_h(x^j)}\,\mathrm{d}x^i - \widetilde{a}_j^{1,0,NW} \tag{4.19}$$

together with the norming $\int \widetilde{a}_h^{12,j,NW}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j = 0$. Then, the asymptotic behavior is given by the following

Theorem 4.3. *Let Assumptions 4.1 and 4.2 be fulfilled and assume that the additive model* (4.18) *with centering holds. For the bandwidth sequence it holds that* $h^2 = O((Tnh)^{-1/2})$, $nh^3 \to \infty$ *and* $(Th)^{-1/2} = o(h^2)$*. Then the algorithm* (4.19) *converges with geometric rate and for the estimators* $\widetilde{a}_h^{12,j,NW}(x^j), j = 1, \dots, d$ *it holds that*

$$\sqrt{Tnh}\frac{\widetilde{a}_h^{12,j,NW}(x^j) - a^{12,j}(x^j) - b_h^{12,j}(x^j) - h^2\beta_\sigma^{12,j,NW}(x^j)}{v^{12}(x^j)\kappa^2(x^j)/\kappa_0(x^j)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where

$$\begin{aligned} b_h^{12,j} &= -\iint a^{12,j}(x^j)K_h(x^j,u^j)f(u^j)\,\mathrm{d}u^j\,\mathrm{d}x^j \\ &\quad - h\iint \frac{\partial}{\partial x^j}a^{12,j}(x^j)\frac{\kappa_1(x^j)}{\kappa_0(x^j)}K_h(x^j,u^j)f(u^j)\,\mathrm{d}x^j \\ \beta_\sigma^{12,j,NW}(x^j) &= h\frac{\partial}{\partial x^j}(a^{12}(x^j))\frac{\kappa_1(x^j)}{\kappa_0(x^j)} + h^2\widetilde{\beta}_\sigma^{12,j}(x^j) \end{aligned}$$

and

$$v^{12}(x^j) = (f(x^j))^{-1}\,\mathbf{E}\big((a^{12}(X))^2 \mid X^j = x^j\big).$$

Again the bias is given only in implicit form by

$$\begin{aligned} &(\widetilde{\beta}_\sigma^{12,0}, \widetilde{\beta}_\sigma^{12,1}(x^1), \dots, \widetilde{\beta}_\sigma^{12,d}(x^d)) \\ &\qquad = \arg\min_{\beta_\sigma^{12,0},\dots,\beta_\sigma^{12,d}} \int (\beta_\sigma(x) - \beta_\sigma^{12,0} - \beta_\sigma^{12,1}(x^1) - \dots - \beta_\sigma^{12,d}(x^d))^2 f(x)\,\mathrm{d}x \end{aligned}$$

with

$$\beta_\sigma(x) = \frac{\kappa_2(x^j)}{\kappa_0(x^j)}\sum_{j=1}^d (f(x))^{-1}\frac{\partial}{\partial x^j}(a^{12}(x^j))\frac{\partial}{\partial x^j}(f(x)) + \frac{1}{2}\frac{\partial^2}{\partial(x^j)^2}a^{12}(x^j).$$

The rate of convergence is given by $\sqrt{Tnh}$ and is faster than in the drift case. This is consistent with the general finding that the diffusion function of diffusion processes can always be estimated with a faster rate.

The joint distribution of the vector of the $\widetilde{a}_h^{12,j}(x^j)$ is a multivariate (d-dimensional) normal distribution and all covariances are asymptotically zero. In the joint distribution of components of different elements of the diffusion matrix, say $a^{ij}(x)$ and $a^{i'j'}(x)$, non-vanishing covariances are present. These are given by

$$\mathbf{cov}(\sqrt{hnT}\widetilde{a}_h^{ij,k,NW}(x^k), \sqrt{hnT}\widetilde{a}_h^{i'j',k,NW}(x^k)) \\ = \frac{\kappa^2(x^k)}{\kappa_0(x^k)f(x^k)}\,\mathbf{E}(a^{ij}(X)a^{i'j'}(X) \mid X^k = x^k)$$

and zero otherwise.

For the estimators of the diffusion function, the efficiency results of the estimators of the drift function carry over. To see this, consider the oracle estimator for $a^{12,j}(x^j)$, which is based on infeasible data

$$Y_{k\Delta}^{**} = \Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta})(X^2_{(k+1)\Delta} - X^2_{k\Delta}) - \sum_{i\neq j} a^{12,i}(X^i_{k\Delta}).$$

The Nadaraya-Watson oracle estimator $\check{a}_h^{12,j,NW}(x^j)$ is then obtained by regressing $Y_{k\Delta}^{**}$ on $X_{k\Delta}$. Using Lemmata 4.5 and 4.7 the asymptotic properties of the oracle estimator can be derived as

$$\sqrt{Tnh}\frac{\check{a}_h^{12,j,NW}(x^j) - a^{12,j}(x^j) - h^2\check{\beta}^{12,j}(x^j)}{v^{12}(x^j)\kappa^2(x^j)/\kappa_0(x^j)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

with

$$\check{\beta}^{12,j}(x^j) = h\frac{\kappa_1(x^j)}{\kappa_0(x^j)}\frac{\partial}{\partial x^j}(a^{12,j}(x^j)) \\ + h^2\Big(\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\frac{\frac{\partial}{\partial x^j}(a^{12,j}(x^j))\frac{\partial}{\partial x^j}(f(x^j))}{f(x^j)} + \frac{1}{2}\frac{\partial^2}{\partial(x^j)^2}a^{12,j}(x^j)\Big).$$

As in the estimation of the drift vector, the local constant smooth backfitting estimator is not fully oracle efficient. While the variance is the same, the two estimators have different bias. A fully oracle efficient estimator is obtained using local linear smooth backfitting.

First, define the marginal local linear estimators as

$$(\widehat{a}_h^{12,j,LL}(x^j), \widehat{a}_{j,h}^{12,j,LL}(x^j)) \\ = \arg\min_{(\bar{a}^{12,j},\bar{a}_j^{12,j})}\int\frac{1}{n}\sum_{k=0}^{nT-1}\big(\Delta^{-1}(X^1_{(k+1)\Delta} - X^1_{k\Delta})(X^2_{(k+1)\Delta} - X^2_{k\Delta}) \\ - \bar{a}^{12,j}(x^j) - \bar{a}_j^{12,j}(x^j)\frac{X^j_{k\Delta} - x^j}{h}\big)^2\prod_{j=1}^{d}K_h(x^j, X^j_{k\Delta})\,\mathrm{d}x$$

and then the local linear smooth backfitting estimators as the solution of

$$\begin{aligned}(4.20)\quad \begin{pmatrix}\widetilde{a}_h^{12,j,LL}(x^j)\\ \widetilde{a}_{j,h}^{12,j,LL}(x^j)\end{pmatrix} &= \begin{pmatrix}\widehat{a}_h^{12,j,LL}(x^j)\\ \widehat{a}_{j,h}^{12,j,LL}(x^j)\end{pmatrix} - \begin{pmatrix}\widetilde{a}_j^{12,0}\\ 0\end{pmatrix}\\ &\quad - \widehat{V}^j(x^j)^{-1}\sum_{i\neq j}\int \widehat{U}^{ij}(x^i,x^j)\begin{pmatrix}\widetilde{a}_h^{12,i,LL}(x^i)\\ \widetilde{a}_{i,h}^{12,i,LL}(x^i)\end{pmatrix}\mathrm{d}x^i\end{aligned}$$

with normalization

$$\int \widetilde{a}_h^{12,j,LL}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j + \int \widetilde{a}_{j,h}^{12,j,LL}(x^j)\widehat{f}_{h,1}(x^j)\,\mathrm{d}x^j = 0.$$

The asymptotic behavior of these estimators is given in the following

Theorem 4.4. *Let Assumptions 4.1 and 4.2 be fulfilled and assume that the additive model* (4.18) *with centering holds. For the bandwidth sequence it holds that* $h^2 = O((Tnh)^{-1/2})$, $nh^3 \to \infty$ *and* $(Th)^{-1/2} = o(h^2)$. *Then the algorithm* (4.19) *converges with geometric rate and for the estimators* $\widetilde{a}^{12,j,LL}(x^j), j = 1,\dots,d$ *it holds that*

$$\sqrt{Tnh}\frac{\widetilde{a}_h^{12,j,LL}(x^j) - a^{12,j}(x^j) + b_h^{12,j}(x^j) - \beta_\sigma^{12,j,LL}(x^j)}{v^{12}(x^j)\widetilde{\kappa}(x^j)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

with $b_h^{12,j}$ *and* $v^{12}(x^j)$ *as given in Theorem 4.3,* $\widetilde{\kappa}(x^j)$ *as in Theorem 4.2 and where*

$$\beta_\sigma^{12,j,LL}(x^j) = h^2\frac{1}{2}\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\frac{\partial^2}{\partial(x^j)^2}a^{12,j}(x^j).$$

This estimator has to be compared to the local linear oracle estimator, which is the local linear estimator of the regression of $Y_{k\Delta}^{**}$ on $X_{k\Delta}$. Then, the result is analogous to the drift estimation. The smooth backfitting estimator $\widetilde{a}_h^{12,j,LL}(x^j)$ has the same bias and the same variance as the oracle estimator.

4.4.3 Bandwidth Choice

An important issue in the application of kernel regression techniques is the selection of the smoothing parameters $h_1,\dots,h_d$. While the derivation of formal results for data-driven bandwidths is beyond the scope of the present work, two possibilities will briefly be described. There are two proposals to select the bandwidth for smooth backfitting estimation. Nielsen and Sperlich (2005) introduce a cross-validation procedure and give evidence by simulations that the method works. However, they do not provide theoretical results. Mammen and Park (2006) investigate the use of penalizing functions and theoretically derive the validity of their procedure.

Presumably the results by Mammen and Park (2006) for independent data could be generalized to the present setting of diffusion processes. However for finite samples, the correlation structure of the data can produce misleading results. Cross-validation can be adjusted to dependent data more easily. The cross validated bandwidths are given as minimizers of the criterion

$$CV(h_1,\dots,h_d) = \sum_{k=0}^{nT-1} (X_{(k+1)\Delta} - X_{k\Delta} - \widetilde{\mu}^{1,0} - \sum_{j=1}^{d} \widetilde{\mu}^{1,j,NW}_{h_j,-\mathcal{J}(k)}(X^j_{k\Delta}))^2.$$

The estimators at the point $X_{k\Delta}$ should be independent from $X_{(k+1)\Delta} - X_{k\Delta}$ to obtain a reliable fit of the unknown function. For independent data this is achieved by using a leave-one-out estimator. For time series data this can be generalized by excluding more data points in the (time-) neighborhood $\mathcal{J}(k)$ of $X_{k\Delta}$ from the prediction.

The estimator $\widetilde{\mu}^{1,j,NW}_h(x^j)$ does not depend asymptotically on the bandwidths of all other components. Therefore, the cross-validation function can be minimized by minimizing over $h_1,\dots,h_d$ subsequently. Usually the optimization is performed via a grid search procedure. The subsequent minimization requires only one-dimensional grid searches and reduces computation time. More time can be saved by optimizing the bandwidth in each iteration step of the calculation of the smooth backfitting estimator. This is described in detail in Nielsen and Sperlich (2005).

4.5 Simulation

In this section the finite sample performance of the smooth backfitting estimators is investigated via a simulation study. Two three-dimensional data generating processes of the form

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \Sigma(X_t)\,\mathrm{d}W_t$$

are considered. For the first process all entries of the drift vector and of the diffusion matrix are not only additive but also linear. For the second process some of these functions are nonlinear. The exact specification of the processes will be given below.

The paths of the process are simulated using the Euler-scheme with $N = 10$ intermediate points to approximate the stochastic integrals. The simulation study will concentrate on the estimation of the components of $\mu^{1,1}(x^1)$ and $a^{11,1}(x^1)$. The results for other functions are comparable. Two sample sizes are considered, a small sample with $n = 35, T = 30$ and a large sample with $n = 50, T = 100$. In the estimation, the Epanechnikov-kernel is implemented and the smooth backfitting estimators are evaluated on a grid of 51 equidistant points in each direction. Ten different combinations of bandwidth constants are considered and they will be adjusted according to the relevant sample size. To judge the performance of the estimators, they are compared via the mean integrated square error given as $\text{MISE}\,(\widetilde{\mu}_h^{1,1}) = \mathbf{E}\big(\int (\widetilde{\mu}_h^{1,1}(x^1) - \mu^{1,1}(x^1))^2 \,\mathrm{d}x^1\big)$. The ISE for one estimator is approximated by evaluating the integral over the interior gridpoints 6–41, not to be affected too severely by boundary problems. The MISE is then estimated as mean or median over 201 simulation runs.

4.5.1 Linear Model

For the linear model an affine diffusion process is considered. The process is specified as

$$\mu(x) = \begin{pmatrix} 0.75 - x^1 + 0.5x^2 + x^3 \\ 0.75 + 0.5x^1 - 2x^2 + 0.25x^3 \\ 1.5 + 0.25x^1 + x^2 - 3x^3 \end{pmatrix},$$

$$\Sigma(x) = \begin{pmatrix} \sqrt{0.3x^1 + 0.3x^2} & 0 & 0 \\ 0 & \sqrt{0.3x^2} & 0 \\ 0 & 0 & \sqrt{0.3x^3} \end{pmatrix}.$$

The estimation is performed on the cube $\mathcal{G} = [0.95, 4.25] \times [0.50, 1.85] \times [0.60, 1.65]$. By simulations it was found that $\mathbb{P}(X_t \in \mathcal{G}) = 0.86$. This means that the estimation was on average based on 900 observations for the small sample and on 4300 observations for the large sample.

First, the results for the drift function are presented. The normalized function is given by

$$\mu^{1,1}(x^1) = 2.32 - x^1,$$

where the mean was calculated by simulations. The bandwidth is given by $\mathbf{h} = (h_0^1, h_0^2, h_0^3)' \times T^{-1/5}$. The results in Table 4.1 indicate that the bandwidths h_2 and h_3 have an influence on the estimator of $\mu^{1,1}(x^1)$. As expected by the asymptotic

Table 4.1: MISE for estimating $\mu^{1,1}(x^1)$ in the linear specification

	$n = 35, T = 30$				$n = 50, T = 100$			
	Nadaraya-Watson		Local Linear		Nadaraya-Watson		Local Linear	
(h_0^1, h_0^2, h_0^3)	Mean	Median	Mean	Median	Mean	Median	Mean	Median
(1.2, 0.38, 0.42)	0.512	0.340	1.478	0.506	0.143	0.118	0.180	0.145
(1.2, 0.54, 0.42)	0.434	0.305	0.791	0.454	0.146	0.113	0.179	0.139
(1.2, 0.54, 0.55)	0.406	0.314	0.863	0.438	0.140	0.114	0.171	0.144
(1.7, 0.54, 0.55)	0.278	0.183	1.106	0.350	0.108	0.090	0.138	0.111
(1.7, 0.54, 0.68)	0.313	0.218	1.234	0.410	0.099	0.086	0.137	0.114
(1.7, 0.70, 0.55)	0.254	0.174	0.729	0.389	0.096	0.068	0.136	0.097
(1.7, 0.70, 0.68)	0.285	0.208	0.634	0.366	0.101	0.083	0.132	0.105
(2.2, 0.70, 0.68)	0.204	0.158	0.714	0.302	0.099	0.074	0.118	0.078
(2.2, 0.86, 0.68)	0.212	0.149	0.625	0.325	0.097	0.080	0.115	0.086
(2.2, 0.70, 0.80)	0.199	0.154	0.508	0.243	0.102	0.079	0.123	0.082

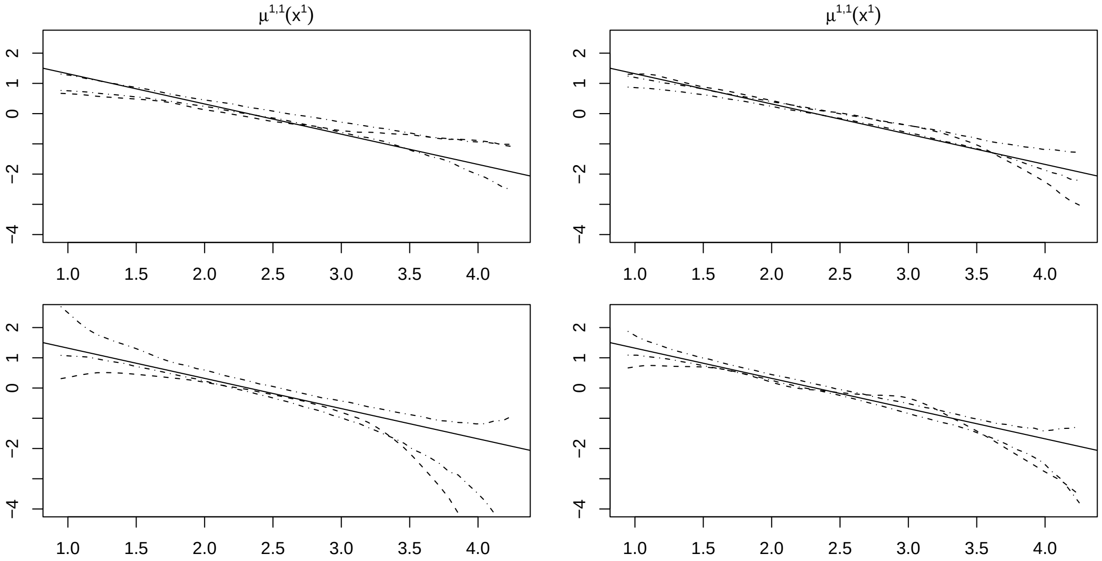

Figure 4.1: Estimators of the drift part $\mu^{1,1}(x^1)$ for $n = 35, T = 30$ (left column) and $n = 50, T = 100$ (right c.). Nadaraya Watson estimators are in first row, local linear in second. In each panel there are given the true function (solid), pointwise 0.25 and 0.75 quantiles of the estimates over 201 simulations (dashed-dotted) and the MISE-median estimator over the simulations (dashed). The bandwidth is given by $\mathbf{h} = (1.7, 0.85, 0.66)' \times T^{-1/5}$

results, this effect is smaller for the large sample. However, the influence of h_1 onto the MISE of $\widetilde{\mu}_h^{1,1}(x^1)$ is stronger than the influence of h_2 and h_3. This finding gives evidence that the recommended bandwidth selection procedure leads to reliable results even in small samples.

The values of the mean and the median differ considerably large, indicating a number of outliers in the simulations. This results from the fact, that in some simulations not the whole cube $\mathcal{G}$ is filled with observations. In that case the density estimators can be very close to zero[2]. This causes problems for the marginal estimators and the integration over the estimated conditional densities in the algorithm. Then, the backfitting estimators can be dominated by some extreme values, based on too few observations in a local neighborhood. However this simulation effect decreases with an increasing number of observations. In practice, the cube $\mathcal{G}$ would be selected such that there are enough observations to avoid this problem. Thus, the result for the median is more reliable to judge the performance of the estimator.

[2] The convention $0/0 = 0$ is used in the implementation

In all settings the Nadaraya-Watson estimator outperforms the local linear estimator. This effect can only partly be attributed to the increasing variance of the local linear estimator at boundary points, because in the calculation of the MISE some boundary points are excluded. In particular in the large sample, no boundary points are used to estimate the MISE, but still the Nadaraya-Watson estimator performs better.

Figure 4.1 underlines these findings. In the upper row, the Nadaraya-Watson estimator is displayed and in the lower row the local linear estimator. The left column shows the results from the small sample and the right column the results from the large sample. The local linear estimators show a large variance near the boundary, which is smaller for the large sample, but still their performance is worse than the Nadaraya-Watson estimators.

The second important finding is that the Nadaraya-Watson estimator seems to exhibit a larger bias, because the interquartile range of the estimators seems to follow a different slope than the true function. Theoretically this can be explained by the difference in the bias behavior of the two estimators. Recall from Theorem 4.1 that the bias of the Nadaraya-Watson estimator is given only implicitly as an additive projection of the first derivative of the component function and of the density. In contrast the bias of the local linear estimator is zero for this data generating process, because the second derivative of $\mu^{1,1}(x^1)$ is zero. This effect is reduced for the large sample. Recall however that the MISE of the Nadaraya-Watson estimator is always smaller. Therefore its variance must be much smaller in finite samples.

The results for estimating the diffusion function $a^{11,1}(x^j) = 0.3(x^1 - 2.32)$ are given in Table 4.2 and Figure 4.2. The bandwidth is now given by $\mathbf{h} = (h_0^1, h_0^2, h_0^3)' \times (nT)^{-1/5}$ because of the faster rate of convergence of the diffusion estimator. The findings from Table 4.2 are similar to those for the drift estimator. For all settings the Nadaraya-Watson estimator has a smaller MISE than the local linear estimator, however the difference decreases with an increasing sample size. The effect of the bandwidth constant on the MISE is much smaller than for estimating the drift function, which should be due to the faster rate of convergence.

Next, compare the results of Figure 4.2 to the estimation of the drift function. One can see that the bias of the Nadarya-Watson estimator is still present, but the effect is much smaller. The interquartile range of the Nadaraya-Watson estimator is still smaller than that of the local linear estimator and the magnitude of this

Table 4.2: MISE for estimating $a^{11,1}(x^1)$ in the linear specification

	$n = 35, T = 30$				$n = 50, T = 100$			
	Nadaraya-Watson		Local Linear		Nadaraya-Watson		Local Linear	
(h_0^1, h_0^2, h_0^3)	Mean	Median	Mean	Median	Mean	Median	Mean	Median
(1.2, 0.38, 0.42)	0.453	0.403	0.506	0.421	0.388	0.378	0.394	0.384
(1.2, 0.54, 0.42)	0.446	0.400	0.475	0.421	0.386	0.382	0.391	0.386
(1.2, 0.54, 0.55)	0.433	0.389	0.456	0.406	0.390	0.385	0.395	0.393
(1.7, 0.54, 0.55)	0.383	0.365	0.434	0.393	0.370	0.367	0.380	0.380
(1.7, 0.54, 0.68)	0.405	0.366	0.512	0.387	0.374	0.374	0.384	0.381
(1.7, 0.70, 0.55)	0.421	0.365	0.458	0.397	0.376	0.373	0.385	0.379
(1.7, 0.70, 0.68)	0.445	0.381	0.487	0.406	0.373	0.368	0.383	0.381
(2.2, 0.70, 0.68)	0.372	0.358	0.413	0.399	0.361	0.355	0.375	0.369
(2.2, 0.86, 0.68)	0.406	0.378	0.505	0.405	0.369	0.363	0.383	0.371
(2.2, 0.70, 0.80)	0.392	0.362	0.448	0.404	0.371	0.364	0.387	0.383

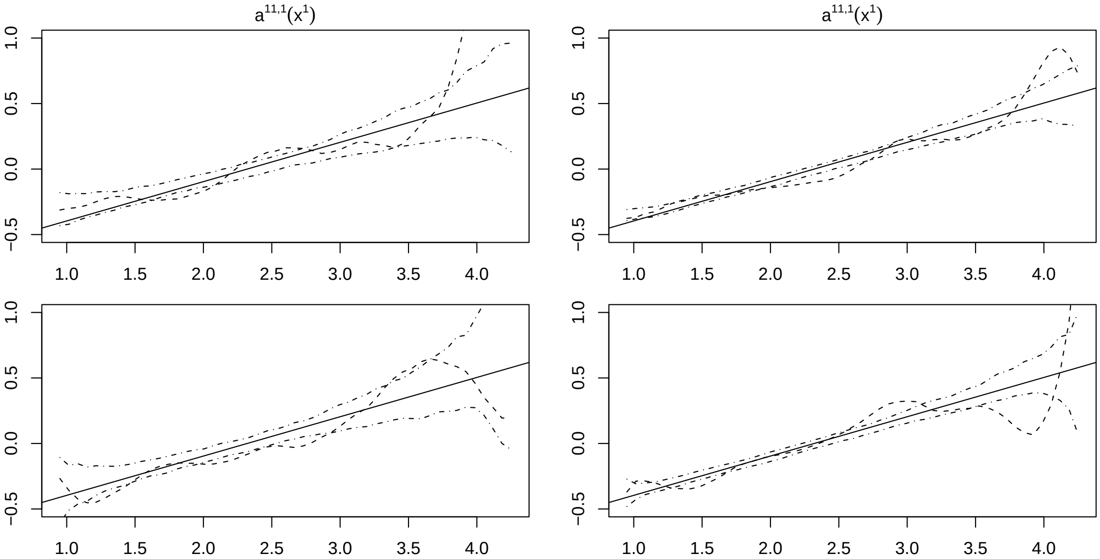

Figure 4.2: Estimators of the diffusion part $a^{11,1}(x^1)$ for $n = 35, T = 30$ (left column) and $n = 50, T = 100$ (right c.). Nadaraya Watson estimators are in first row, local linear in second. In each panel there are given the true function (solid), pointwise 0.25 and 0.75 quantiles of the estimates over 201 simulations (dashed-dotted) and the MISE-median estimator over the simulations (dashed). The bandwidth is given by $\mathbf{h} = (1.7, 0.70, 0.66)' \times (nT)^{-1/5}$

distance is much smaller than for the drift estimation. All these findings highlight the increased rate of convergence for the estimation of the diffusion function.

4.5.2 Nonlinear Model

In the second specification, a process with nonlinear elements of the drift and diffusion is simulated. The concrete specification is given by

$$\mu(x) = \begin{pmatrix} 0.4x^1 - 1.1(x^1)^2 + .01/x^1 + 0.25x^2 + x^3 \\ 0.75 + 0.5x^1 - 2x^2 + 0.25x^3 \\ 1.5 + 0.25x^1 + x^2 - 3x^3 \end{pmatrix},$$

$$\Sigma(x) = \begin{pmatrix} \sqrt{0.3(x^1)^2 + 0.3(x^2)^2} & 0 & 0 \\ 0 & \sqrt{0.3x^2} & 0 \\ 0 & 0 & \sqrt{0.3x^3} \end{pmatrix}.$$

The specification of $\mu^{1,1}(x^1)$ was considered by Aït-Sahalia (1996b) for a scalar diffusion to model interest rates. The estimation is restricted to the cube $\mathcal{G} =$

Table 4.3: MISE for estimating $\mu^{1,1}(x^1)$ in the nonlinear specification ($\times 10^{-2}$)

	$n = 35, T = 30$				$n = 50, T = 100$			
	Nadaraya-Watson		Local Linear		Nadaraya-Watson		Local Linear	
(h_0^1, h_0^2, h_0^3)	Mean	Median	Mean	Median	Mean	Median	Mean	Median
(0.40, 0.28, 0.36)	8.114	5.995	10.853	8.002	3.644	2.789	4.104	3.176
(0.40, 0.39, 0.36)	8.738	6.091	11.878	7.090	3.809	3.001	4.199	3.435
(0.40, 0.39, 0.47)	8.848	6.965	11.677	8.872	3.711	2.652	4.227	3.133
(0.57, 0.39, 0.47)	5.816	3.972	9.520	5.941	2.938	2.001	3.334	2.341
(0.57, 0.39, 0.58)	5.144	3.959	9.351	6.317	2.156	1.845	2.781	2.149
(0.57, 0.50, 0.47)	5.185	3.704	8.911	5.896	2.937	1.765	3.470	2.134
(0.57, 0.50, 0.58)	5.583	4.207	8.907	6.691	2.561	1.666	3.084	1.934
(0.74, 0.50, 0.58)	4.549	3.198	7.959	5.217	2.217	1.721	2.404	1.832
(0.74, 0.61, 0.58)	4.315	2.990	7.370	5.019	2.032	1.478	2.195	1.611
(0.74, 0.50, 0.69)	4.055	3.196	8.455	4.287	2.607	1.556	2.827	1.697

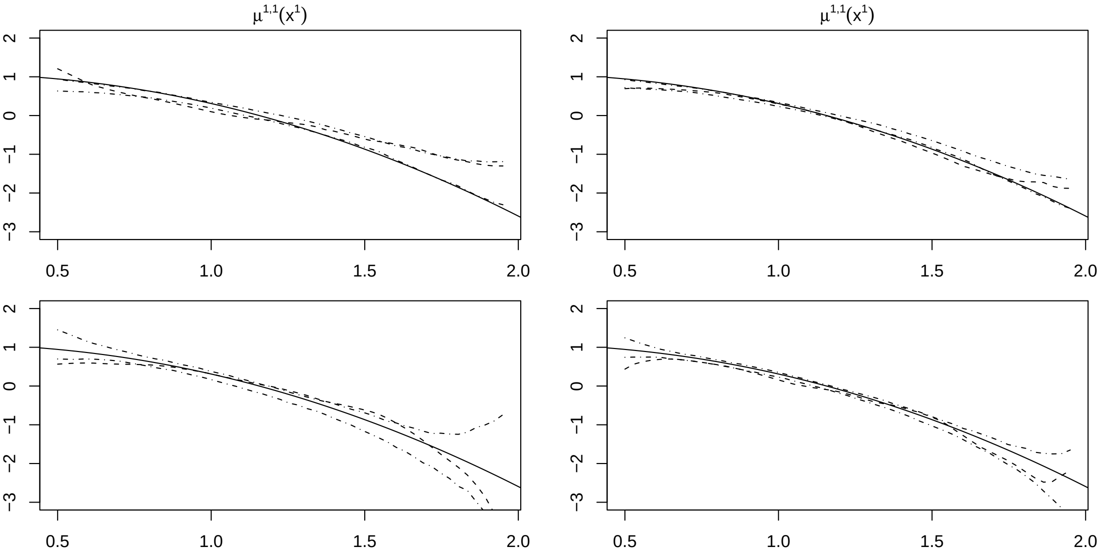

Figure 4.3: Estimators of the drift part $\mu^{1,1}(x^1)$ for $n = 35, T = 30$ (left column) and $n = 50, T = 100$ (right c.). Nadaraya Watson estimators are in first row, local linear in second. In each panel there are given the true function (solid), pointwise 0.25 and 0.75 quantiles of the estimates over 201 simulations (dashed-dotted) and the MISE-median estimator over the simulations (dashed). The bandwidth is given by $\mathbf{h} = (0.74, 0.60, 0.57)' \times T^{-1/5}$

$[0.50, 1.95] \times [0.35, 1.30] \times [0.45, 1.35]$ with $\mathbb{P}(X_t \in \mathcal{G}) = .85$. For estimation in this model the bandwidth constants are changed by a factor corresponding to the different range of the cube $\mathcal{G}$.

The results for estimating the first component of the drift function

$$\mu^{1,1}(x^1) = 0.4x^1 - 1.1(x^1)^2 + .01/x^1 + 1$$

are presented in Table 4.3 and Figure 4.3. Note that the MISE in Table 4.3 is multiplied by 100. As in the linear case, the Nadaraya-Watson estimator outperforms the local linear estimator. Furthermore the local linear estimator suffers from severe outlier problems. This is less evident for the large sample, but still present. It is also observed that the magnitude of the MISE is mainly affected by its own bandwidth, but some finite sample effects are present.

From Figure 4.3 it can be seen that the local linear estimator exhibits a larger variance, especially at the right boundary. On the other hand, the bias of the Nadaraya-Watson estimator is well visible for the small sample and in particular the estimator seems not to capture the nonlinearity very well. For the large

Table 4.4: MISE for estimating $a^{11,1}(x^1)$ in the nonlinear specification ($\times 10^{-2}$)

	$n = 35, T = 30$				$n = 50, T = 100$			
	Nadaraya-Watson		Local Linear		Nadaraya-Watson		Local Linear	
(h_0^1, h_0^2, h_0^3)	Mean	Median	Mean	Median	Mean	Median	Mean	Median
(0.40, 0.28, 0.36)	0.954	0.823	1.036	0.860	0.300	0.243	0.302	0.245
(0.40, 0.39, 0.36)	1.065	0.806	1.151	0.875	0.346	0.277	0.349	0.271
(0.40, 0.39, 0.47)	0.964	0.833	1.026	0.892	0.347	0.279	0.351	0.289
(0.57, 0.39, 0.47)	0.667	0.535	0.703	0.568	0.222	0.199	0.224	0.200
(0.57, 0.39, 0.58)	0.686	0.543	0.735	0.566	0.223	0.199	0.222	0.196
(0.57, 0.50, 0.47)	0.672	0.544	0.728	0.581	0.260	0.212	0.258	0.214
(0.57, 0.50, 0.58)	0.664	0.557	0.709	0.589	0.240	0.205	0.238	0.200
(0.74, 0.50, 0.58)	0.559	0.465	0.627	0.472	0.213	0.151	0.206	0.143
(0.74, 0.61, 0.58)	0.585	0.451	0.623	0.510	0.172	0.153	0.168	0.147
(0.74, 0.50, 0.69)	0.674	0.461	0.787	0.501	0.204	0.159	0.198	0.150

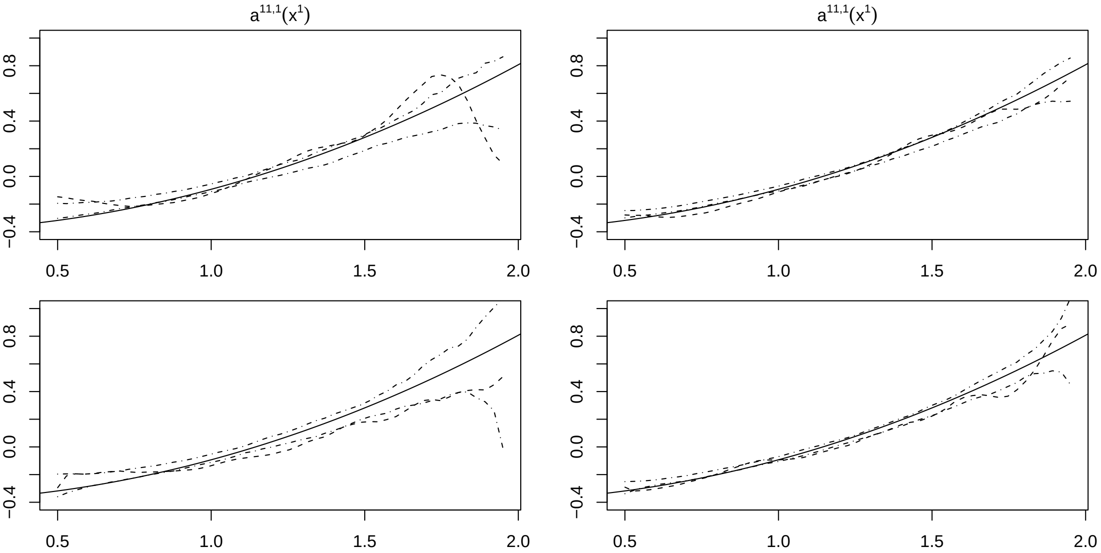

Figure 4.4: Estimators of the drift part $a^{11,1}(x^1)$ for $n = 35, T = 30$ (left column) and $n = 50, T = 100$ (right c.). Nadaraya Watson estimators are in first row, local linear in second. In each panel there are given the true function (solid), pointwise 0.25 and 0.75 quantiles of the estimates over 201 simulations (dashed-dotted) and the MISE-median estimator over the simulations (dashed). The bandwidth is given by $\mathbf{h} = (0.74, 0.60, 0.57)' \times (nT)^{-1/5}$

sample, the performance of the estimators seems to be comparable, but recall from Table 4.3, that the Nadaraya-Watson estimator has a smaller MISE.

Finally, turn to the diffusion estimator in the nonlinear setting. Here, the function under investigation is

$$a^{11,1}(x^1) = 0.3((x^1)^2 - 1.31).$$

The simulated MISE is presented in Table 4.4 and for the first time in the simulations the local linear estimator outperforms the Nadaraya-Watson estimator in the large sample for some bandwidth settings. From Figure 4.4 it can be seen that both estimators capture the shape of the unknown function well.

The simulation study performed in this section gives evidence that the theoretical properties of the smooth backfitting estimators hold in finite samples. In comparable studies, Chapman and Pearson (2000) and Fan and Zhang (2003) have investigated univariate Nadaraya-Watson and local linear estimators with similar (or even larger) sample sizes. The results for the smooth backfitting estimators are comparable to the scalar estimators and underline the theoretical finding that the curse of dimensionality can be circumvented by the structured model. In contrast to the general theory, the local constant estimators outperform the local linear estimator in almost all cases. Therefore it seems advisable to use the Nadaraya-Watson estimator if the sample size is not very large (≤ 5000) and not to rely on the asymptotic advantages of the local linear estimator in applications.

4.6 Estimating Interest Yields

As an illustrative example, the estimators are applied to estimating the dynamics of bond yields. Beginning with Aït-Sahalia (1996a,b) a large number of authors have applied kernel regression techniques to estimate (univariate) short term interest rates. As an extension the smooth backfitting estimators are applied to a multivariate model of interest yields, using different maturities.

The data consist of daily interest rates for selected U.S. Treasury securities at different fixed maturities. The series is constructed by the U.S. Federal Reserve Board and can be downloaded from its homepage[3]. The three variables under consideration are the three-month interest rate, the spread between the two-year rate and the three-month rate and the spread between the ten-year rate and the three-month rate. The sample consists of daily data from January 1, 1991 to December 29, 2000, which results in a total of 2 504 observations. The data is displayed in Figure 4.5.

To apply the smooth backfitting estimators a rectangular subset of the original data must be chosen, over which the estimation procedure has to be carried out. The density of the process has to be bounded from below over this subset and therefore $\mathcal{G} = [3, 6.25] \times [-0.25, 1.75] \times [-0.5, 3.5]$ (short rate × spread 3m/2y × spread 3m/10y) was selected. This resulted in a final sample of 2 147 observations. The frequency was set to $n = 20$, leading to roughly one month between two time units. Then, the entries of the drift vector $\mu(x)$ and of the diffusion matrix $A(x)$ are estimated using Nadaraya-Watson smooth backfitting. The choice of the local constant fitting is based on the simulation results of the last subsection. The exposition of the estimation in this section will concentrate on $\mu^1(x)$, which is the drift function of the short rate.

The bandwidth was selected via a cross-validation procedure as described in

[3] `www.federalreserve.gov/Releases/H15/data.htm`

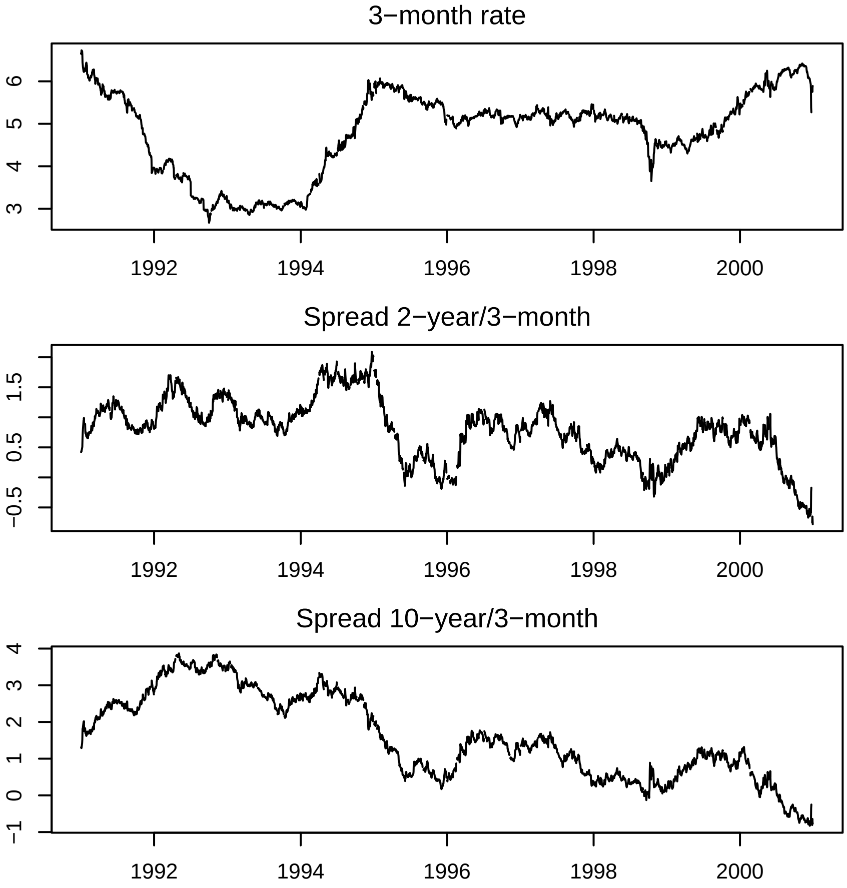

Figure 4.5: U.S. treasury data for 3-month yields (upper), spread between 2-year and 3-month yields (middle), spread between 10-year and 3-month yields (lower) from Jan. 1, 1991 to Dec. 29, 2000.

section 4.4.3. The leave-out estimator was constructed by omitting the 250 observations closest in time to $X_{i\Delta}$. Notationally, a subset $\mathcal{J}(i) = \{X_{(i-125)\Delta}, \dots, X_{i\Delta}, \dots, X_{(i+125)\Delta}\}$ is left out. To save computation time, the cross-validation function was not evaluated at all data points, but only at a subset of 250 randomly chosen observations away from the boundary. The cross-validated bandwidths are given by $\widehat{h}^{CV} = (0.26, 0.26, 0.44)'$.

Using the asymptotic theory, pointwise confidence bands can be obtained by estimating the asymptotic variances, given by

$$(Th)^{-1/2} v^1(x^j) \kappa^2 / (\kappa_0)^2.$$

Kernel density estimators $\widehat{f}_{\widetilde{h}}(x^j)$ as defined above are used to estimate the mar-

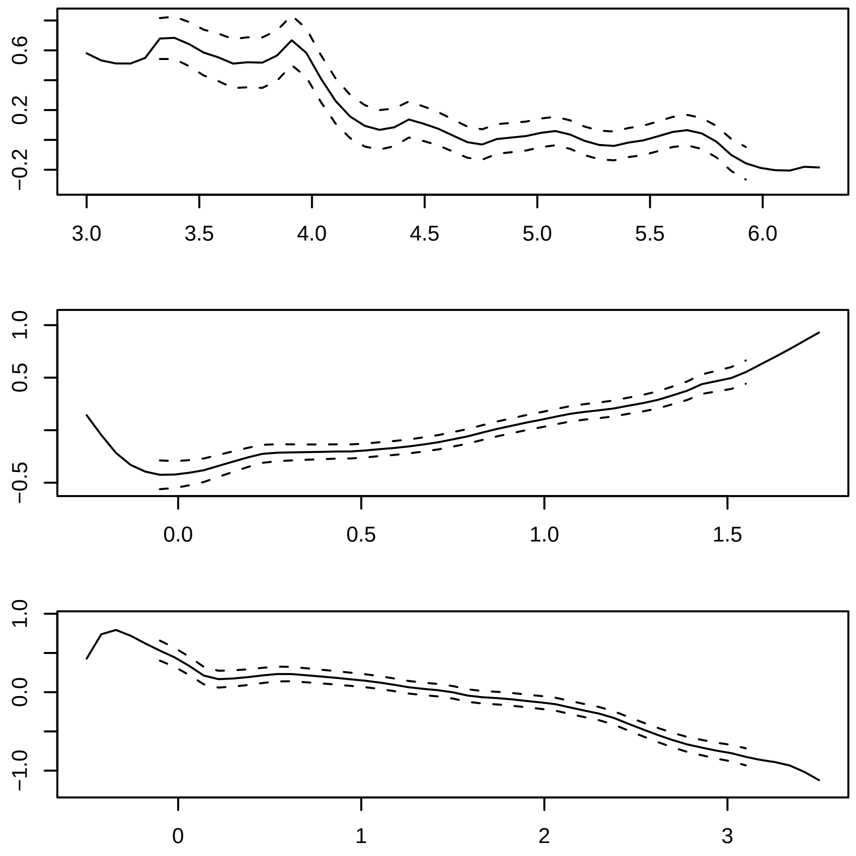

Figure 4.6: Additive components of the drift function of the short rate process. Component function of the short rate (upper), of the 2-y/3-m spread (middle) and of the 10y/3m spread (lower). Smooth backfitting estimators (solid line) together with 90 % confidence bands (dashed).

ginal densities and $\mathbf{E}(a^{11}(X) \mid X^j = x^j)$ is estimated by

$$(\widehat{f}_{\widetilde{h}}(x^j))^{-1}\frac{1}{nT}\sum_{i=0}^{nT-1} K_{\widetilde{h}}(x^j, X^j_{i\Delta})\Delta^{-1}(X_{(i+1)\Delta} - X_{i\Delta})^2.$$

To construct the confidence bands a larger bandwidth $\widetilde{h}_i = 1.5h_i$ was implemented and the bands are only computed for interior points, where the kernel constants do not depend on x^j.

In Figure 4.6 the additive components of the drift function in the short rate process are displayed. The first component clearly has a (large) positive influence for small values of the interest rate and a (small) negative influence for large values. This is in line with the mean reverting property of the short rate, i. e. the

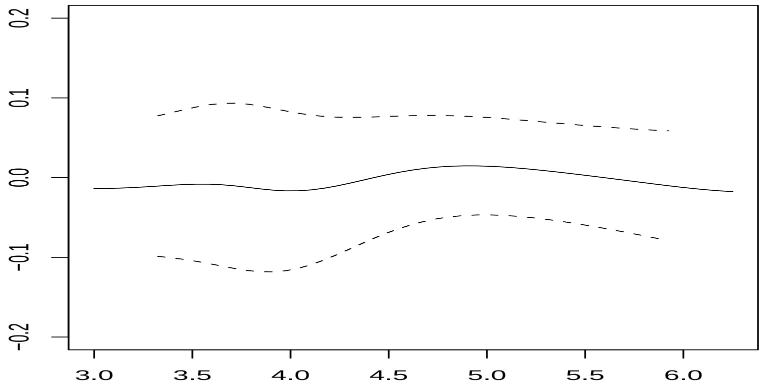

Figure 4.7: Nadarya-Watson estimator (solid) of the drift function of the short rate together with 90 % confidence bands (dashed).

process is always pulled back to its long term mean. From the other two pictures in Figure 4.6 it can be seen that the spreads seem to have a significant influence on the changes in the short rate. The second component of the drift function seems to be a linearly increasing function of the 2-year/3-month spread and the third component is linearly decreasing in the 10-year/3-month spread.

Figure 4.7 shows the estimator of regressing the increments of the short rate on the short rate. This corresponds to modeling the 3-month rate as a scalar diffusion process. The drift function is nearly constant. For this range of estimation this was also observed by Aït-Sahalia (1996b) and Stanton (1997) among others. However the multivariate analysis changes this impression. The spreads seem to have an influence on the evolution of the short rate. This phenomenon could be investigated further by adapting the testing procedure of Chapter 5 to test affine term structure models.

4.7 Conclusion

The theoretical results and the simulation study show evidence for the applicability of smooth backfitting estimators to the estimation of diffusion models. In these models the curse of dimensionality is augmented by the dependence structure of the data and multivariate kernel regression is therefore not applicable even in relatively small dimensions. The simulation results show that the estimators behave like one-dimensional estimators in similar data samples.

The estimators converge even if the additive model does not hold. In that case the additive projection of the unknown multivariate function will be estimated (see Chapter 5 for the case of independent and identically distributed data). Therefore smooth backfitting provides a powerful data analytic tool, even if the additive model is not assumed to hold. For future research it is desirable to obtain testing procedures for the diffusion model. To test the hypothesis that the model is fully additive the procedure by Mammen and Sperlich (2006) can be extended. Secondly, the additive estimators can be used to test for parametric form as in Chapter 5.

Appendix I: Preliminary Lemmata

To establish the asymptotic properties of the smooth backfitting estimators one-dimensional marginal estimators have to be investigated, since the smooth backfitting estimators inherits their behavior. In the proofs the dimension index of the estimators will be suppressed, i. e. $\widehat{\mu}_h^{NW}(x^j) = \widehat{\mu}_h^{1,j,NW}(x^j), \widetilde{\mu}_h^{NW}(x^j) = \widetilde{\mu}_h^{1,j,NW}(x^j)$ and for the local linear estimators analogously. Using the integral form of the stochastic differential equaltion for X^1, the local constant and the local linear estimator will be decomposed into a bias and a variance part. For this define for $l = 0, 1$

$$\widehat{t}_{h,l}^B(x^j) = \frac{1}{T}\sum_{k=1}^{nT} K_h(x^j, X^j_{(k-1)\Delta})\Big(\frac{X^j_{(k-1)\Delta} - x^j}{h}\Big)^l \int_{(k-1)\Delta}^{k\Delta} \mu^1(X_s)\,\mathrm{d}s$$

$$\widehat{t}_{h,l}^V(x^j) = \frac{1}{T}\sum_{k=1}^{nT} K_h(x^j, X^j_{(k-1)\Delta})\Big(\frac{X^j_{(k-1)\Delta} - x^j}{h}\Big)^l \sum_{i=1}^{d} \int_{(k-1)\Delta}^{k\Delta} \sigma^{1i}(X_s)\,\mathrm{d}W_s^i.$$

For abbreviation

$$\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta}) = K_h(x^j, X^j_{(k-1)\Delta})\Big(\frac{X^j_{(k-1)\Delta} - x^j}{h}\Big)^l$$

is defined and for a more compact notation $\widehat{f}_{h,0}(x^j) = \widehat{f}_h(x^j)$ and $\widehat{f}_{h,(0,0)}(x^i, x^j) = \widehat{f}_h(x^i, x^j)$ is introduced. This enables to write the marginal Nadaraya-Watson estimator as

$$(4.21)\quad \widehat{\mu}_h^{NW}(x^j) = \widehat{f}_{h,0}(x^j)^{-1}\widehat{t}_{h,l}^B(x^j) + \widehat{f}_{h,0}(x^j)^{-1}\widehat{t}_{h,l}^V(x^j) = \widehat{\mu}_h^{NW,B}(x^j) + \widehat{\mu}_h^{NW,V}(x^j)$$

and the marginal local linear estimator

$$
\begin{aligned}
\begin{pmatrix} \widehat{\mu}_h^{LL}(x^j) \\ \widehat{\mu}_{j,h}^{LL}(x^j) \end{pmatrix} &= \begin{pmatrix} \widehat{f}_{h,0}(x^j) & \widehat{f}_{h,1}(x^j) \\ \widehat{f}_{h,1}(x^j) & \widehat{f}_{h,2}(x^j) \end{pmatrix}^{-1} \begin{pmatrix} \widehat{t}_{h,0}^B(x^j) \\ \widehat{t}_{h,1}^B(x^j) \end{pmatrix} \\
&\quad + \begin{pmatrix} \widehat{f}_{h,0}(x^j) & \widehat{f}_{h,1}(x^j) \\ \widehat{f}_{h,1}(x^j) & \widehat{f}_{h,2}(x^j) \end{pmatrix}^{-1} \begin{pmatrix} \widehat{t}_{h,0}^V(x^j) \\ \widehat{t}_{h,1}^V(x^j) \end{pmatrix} \\
&= \begin{pmatrix} \widehat{\mu}_h^{LL,B}(x^j) \\ \widehat{\mu}_{j,h}^{LL,B}(x^j) \end{pmatrix} + \begin{pmatrix} \widehat{\mu}_h^{LL,V}(x^j) \\ \widehat{\mu}_{j,h}^{LL,V}(x^j) \end{pmatrix}.
\end{aligned} \tag{4.22}
$$

As building blocks for the asymptotic distribution for the smooth backfitting estimator serve the uniform convergence rates of both parts and the asymptotic distribution of the variance part. To establish this, the one-and two-dimensional marginal density estimates, will be investigated first, beacuse they arise as well in the algorithms of the smooth backfitting estimators.

Lemma 4.1. *Under Assumptions 4.1 and 4.2 it holds that*

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{f}_{h,k}(x^j) - \kappa_k(x^j) f(x^j)| = O_P(h^2 + \Big(\frac{\log T}{hT}\Big)^{1/2})$$

$$\sup_{(x^i,x^j) \in \mathcal{G}^i \times \mathcal{G}^j} |\widehat{f}_{h,(k_1,k_2)}(x^i,x^j) - \kappa_{k_1}(x^i)\kappa_{k_2}(x^j) f(x^i,x^j)| = O_P(h^2 + \Big(\frac{\log T}{h^2 T}\Big)^{1/2})$$

for $k = 0, 1, 2$ and $0 \leq k_1, k_2 \leq 1$ if the rates on the right hand side converge to zero for $T \to \infty$.

The result is here stated for arbitrary n. In general there are cases, where it is possible to obtain the superoptimal rate $T^{-1/2}$ for $n \to \infty$ (see e. g. Bosq, 1998). However for multivariate diffusion processes (in contrast to the scalar case) these conditions are not satisfied and more than the standard nonparametric rate cannot be achieved.

Proof. Consider the one-dimensional case first and decompose the estimator into bias and variance

$$\widehat{f}_{h,k}(x^j) - \kappa_k(x^j) f(x^j) = \widehat{f}_{h,k}(x^j) - \mathbf{E}\,\widehat{f}_{h,k}(x^j) + \mathbf{E}\,\widehat{f}_{h,k}(x^j) - \kappa_k(x^j) f(x^j).$$

Standard kernel calculations show that the bias is of order h^2 uniformly over the interior of $\mathcal{G}^j$.

Let the variance part be denoted by $J_V(x^j)$. Cover the compact set $\mathcal{G}^j$ by N intervals $\mathcal{G}^j_l = \{x : |x - x^j_l| < N^{-1}\}, l = 1, \dots, N$ and choose $N = O((T/(h^3 \log T))^{1/2})$. Then bound

$$\sup_{x^j \in \mathcal{G}^j} |J_V(x^j)| \leq \max_{l=1,\dots,N} \sup_{x^j \in \mathcal{G}^j_l} |J_V(x^j) - J_V(x^j_l)| + \max_{l=1,\dots,N} |J_V(x^j_l)|.$$

By the Lipschitz continuity of the kernel it holds for the first maximum that

$$\max_{l=1,\dots,N} \sup_{x^j \in \mathcal{G}^j_l} |J_V(x^j) - J_V(x^j_l)| = O(N^{-1}h^{-2}) = o\Big(\Big(\frac{\log T}{hT}\Big)^{1/2}\Big).$$

Now regard $J_V(x^j_l)$ as a sum of α-mixing random variables

$$S_{t,n}(x^j_l) = \frac{1}{nT} \sum_{i=1}^{n} \widetilde{K}_{h,k}(x^j_l, X^j_{t+i\Delta}) - \mathbf{E}\,\widetilde{K}_{h,k}(x^j_l, X^j_{t+i\Delta}).$$

Trivially $\mathbf{E}\,S_{i,n}(x^j_k) = 0$ and furthermore

$$\begin{aligned}\mathbf{E}(S_{t,n}(x^j_l))^2 &= \frac{1}{n^2T^2} \sum_{i=1}^{n} \mathbf{E}\big(\widetilde{K}_{h,k}(x^j_l, X_{t+i\Delta})\big)^2 \\ &\quad + \frac{2}{n^2T^2} \sum_{i=1}^{n-2} \sum_{i'=i+1}^{n-1} \mathbf{E}\big(\widetilde{K}_{h,k}(x^j_l, X^j_{t+i\Delta})\widetilde{K}_{h,k}(x^j_l, X^j_{t+i'\Delta})\big) \\ &= O(h^{-1}T^{-2}),\end{aligned}$$

because of the Cauchy-Schwarz inequality and $\mathbf{E}\,\widetilde{K}_{h,k}(x^j_l, X_{t+i\Delta})\big)^2 = O(h^{-1})$. Cramer's conditions are are easily verified with a constant $(Th)^{-1}$.
With these results a Hoeffding-type inequality (see Theorem 1.3 in Bosq, 1998) can be applied to obtain

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{T} S_{t,n}(x^j_k)\Big| > \varepsilon\lambda_n\Big) \leq const\ T^{-2}$$

where $\lambda_n = (\log T^2/(Th))^{1/2}$. Because $NT^{-2} = o(1)$ the desired result follows.
For the two-dimensional case, it can be decomposed as above into bias and variance and proceeded as before. The only difference is that the variance of the kernel is then of order h^{-2}. □

Next, the uniform convergence of the estimators will be investigated, beginning with the bias parts

Lemma 4.2. *Under Assumptions 4.1 and 4.2 it holds that*

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{NW,B}(x^j) - \mathbf{E}(\mu^1(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log T}{(nTh)^{1/2}}\Big)$$
$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{LL,B}(x^j) - \mathbf{E}(\mu^1(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log T}{(nTh)^{1/2}}\Big)$$
$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_{j,h}^{LL,B}(x^j) - h\partial_j \mathbf{E}(\mu^1(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log T}{(nTh)^{1/2}}\Big)$$

if the terms on the right hand side converge to zero as $T \to \infty$ and $n \to \infty$.

Proof. Define $m^1(x^j) = \mathbf{E}(\mu^1(X) \mid X^j = x^j)$. Standard kernel calculations show that $\mathbf{E}(\widehat{\mu}_h^{B,NW}(x^j)) = m^1(x^j) + O(h^2)$, $\mathbf{E}(\widehat{\mu}_h^{B,LL}(x^j)) = m^1(x^j) + O(h^2)$ and $\mathbf{E}(\widehat{\mu}_{j,h}^{B,LL}(x^j)) = h\partial_j m^1(x^j) + O(h^2)$ uniformly in x^j.

For the proof uniform bounds for $\widehat{t}_{h,l}^B(x^j)$ or more precisely for a centered version have to be established

$$\begin{aligned}
\widehat{t}_{h,l}^{B*}(x^j) &= \widehat{t}_{h,l}^B(x^j) - \widehat{f}_{h,l}(x^j)\,\mathbf{E}(\widehat{\mu}_h^B(x^j)) \\
&= \frac{1}{nT}\sum_{k=1}^{nT} \widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\Big(\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} \mu^1(X_s)\,\mathrm{d}s - \mathbf{E}(\widehat{\mu}_h^B(x^j))\Big) \\
&= \frac{1}{nT}\sum_{k=1}^{nT} \widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\Big(\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} \mu^1(X_s)\,\mathrm{d}s - \mu^1(X^j_{(k-1)\Delta})\Big) \\
&\quad + \frac{1}{nT}\sum_{k=1}^{nT} \widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\big(\mu^1(X^j_{(k-1)\Delta}) - \mathbf{E}(\widehat{\mu}_h^B(x^j))(x^j)\big) \\
&= A(x^j) + B(x^j),
\end{aligned}$$

where $\widehat{\mu}_h^B(x^j)$ is the estimator under investigation. To apply a Hoeffding-type inequality (as in the proof of Lemma 4.1) treat the two terms separately and regard them as sums of T α-mixing random variables $S_{t,n}^A(x^j)$ and $S_{t,n}^B(x^j), t = 0, \dots, T-1$. Next it will be shown that $\mathbf{E}(S_{t,n}^A(x^j))^2 = O(T^{-2}n^{-1}h^{-1})$ and $\mathbf{E}(S_{t,n}^B(x^j))^2 = O(T^{-2}n^{-1}h^{-1})$.

Apply the mean value theorem to obtain (setting $t = 0$ wlog)

$$\begin{aligned}
&\mathbf{E}(S_{0,n}^A(x^j))^2 \\
&\quad = \frac{1}{n^2T^2}\sum_{k=1}^{n} \mathbf{E}\Big(\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} (X_s - X_{(k-1)\Delta})\partial_j\mu^1(\xi_{k,s})\,\mathrm{d}s\Big)^2 \\
&\quad + \frac{2}{n^2T^2}\sum_{k=1}^{n-1}\sum_{k'=k+1}^{n} \mathbf{E}\Big(\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} (X_s - X_{(k-1)\Delta})\partial_j\mu^1(\xi_{k,s})\,\mathrm{d}s
\end{aligned}$$

$$\times\, \widetilde{K}_{h,l}(x^j, X^j_{(k'-1)\Delta})\frac{1}{\Delta}\int_{(k'-1)\Delta}^{k'\Delta} (X_s - X_{(k'-1)\Delta})\partial_j\mu^1(\xi_{k',s})\,\mathrm{d}s\Big)$$

where $\xi_{k,s} \in (X_{(k-1)\Delta}, X_s)$ (wlog $X_{(k-1)\Delta} < X_s$). Beginning with the variance parts, it holds that

$$\begin{aligned}
&\mathbf{E}\Big(\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} (X_s - X_{(k-1)\Delta})\partial_j\mu^1(\xi_{k,s})\,\mathrm{d}s\Big)^2\\
&\quad\le \sup_{x\in\mathcal{G}} |\partial_j\mu^1(x)|^2\, \mathbf{E}\big(\widetilde{K}_{h,l}(x^j, X_{(k-1)\Delta})^2\big(\max_{(k-1)\Delta\le s\le k\Delta} |X_s - X_{(k-1)\Delta}|)^2\big)\\
&\quad= O(h^{-1}n^{-1}).
\end{aligned}$$

The last line follows by iterated expectations and an application of the Burkholder-Davis-Grundy inequality. The covariance terms can be bounded by Cauchy-Schwarz, and the verification of Cramer's conditions goes along the same lines. This yields the stated rate for $\mathbf{E}(S^A_{0,n}(x^j)^2$.
Secondly, turn to $S^B_{0,n}(x^j)$.

$$\begin{aligned}
\mathbf{E}(S^B_{0,n}(x^j))^2 &= \frac{1}{n^2T^2}\sum_{k=1}^n \mathbf{E}\big(\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\big(\mu^1(X^j_{(k-1)\Delta}) - \mathbf{E}(\widehat{\mu}_h^{1,B}(x^j))\big)\big)^2\\
&\quad+ \frac{2}{n^2T^2}\sum_{k=1}^{n-1}\sum_{k'=k+1}^n \mathbf{E}\Big(\widetilde{K}_{h,l}(x^j, X^j_{(k-1)\Delta})\big(\mu^1(X^j_{(k-1)\Delta}) - \mathbf{E}(\widehat{\mu}_h^{1,B}(x^j)\big)\\
&\quad\quad\times \widetilde{K}_{h,l}(x^j, X^j_{(k'-1)\Delta})\big(\mu^1(X^j_{(k'-1)\Delta}) - \mathbf{E}(\widehat{\mu}_h^{1,B}(x^j))\big)\Big)\\
&= O(T^{-2}n^{-1}h^{-1}) + O(T^{-2}).
\end{aligned}$$

Because μ^1 and the expected value of the estimator are bounded the differences can be taken out of the expectations. Then the rates follow from standard kernel calculations. This completes the proof of

$$\sup_{x^j\in\mathcal{G}^j} |\widehat{t}^{B,*}_{h,l}(x^j)| = O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big).$$

To show the lemma, consider the following decomposition

$$\widehat{\mu}_h^{B,NW}(x^j) - m^1(x^j) = \widehat{f}_{h,0}(x^j)^{-1}\widehat{t}^{B*}_{h,0}(x^j) + \mathbf{E}(\widehat{\mu}_h^{1,B}(x^j)) - m^1(x^j).$$

For the investigation of the first term it suffices to concentrate on the numerator, because the density is bounded from below on $\mathcal{G}^j$. Then the statement follows. For the local linear case, use the analogous expansion and the same arguments hold. □

Lemma 4.3. *Under Assumptions 4.1 and 4.2 it holds that*

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{V,NW}(x^j)| = O_P\Big(\frac{\log T}{(hT)^{1/2}}\Big)$$

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{V,LL}(x^j)| = O_P\Big(\frac{\log T}{(hT)^{1/2}}\Big) \qquad \sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_{j,h}^{V,LL}(x^j)| = O_P\Big(\frac{\log T}{(hT)^{1/2}}\Big),$$

if the terms on the right hand side converge to zero as $T \to \infty$ and $n \to \infty$.

Proof. Because the density is bounded from below, it suffices to consider the numerator parts of the estimators. Defining random variables

$$R_{i,l}(x^j) = \frac{1}{T} \sum_{k=0}^{n-1} \widetilde{K}_{h,l}(x^j, X^j_{i+k\Delta}) \sum_{m=1}^{d} \int_{i+k\Delta}^{i+(k+1)\Delta} \sigma^{1m}(X_s)\, \mathrm{d}W_s^m$$

it can be written (for $l = 0, 1$)

$$\widehat{t}^V_{h,l}(x^j) = \sum_{i=1}^{T} R_{i,l}(x^j).$$

Obviously $\mathbf{E}\, R_i(x^j) = 0$ and by Itô's lemma it holds that

$$\begin{aligned}
\mathbf{E}(R_i(x^j))^2 &= \frac{1}{T^2} \sum_{k=0}^{n-1} \mathbf{E}\big(\widetilde{K}_{h,l}(x^j, X^j_{i+k\Delta})\big)^2 \int_{i+k\Delta}^{i+(k+1)\Delta} a^{11}(X_s)\, \mathrm{d}s \\
&= h^{-1} T^{-2} \kappa_l^2(x^j) f(x^j)\, \mathbf{E}(a^{11}(X_0) \mid X^j = x^j)(1 + o(1)).
\end{aligned}$$

Utilizing Itô's lemma, Cramer's conditions can be proved with a constant $(Th)^{-1}$. Then as in the proof of Lemmata 4.1 and 4.2 (using an exponential inequality and covering arguments) it follow that

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{t}^V_{h,l}(x^j)| = O_P\Big(\frac{\log T}{(hT)^{1/2}}\Big)$$

and both parts of the lemma follow. □

Beside the uniform convergence results, the asymptotic distribution of the variance parts of the two estimators have to be derived. This is given in the following

Lemma 4.4. *Under Assumptions 4.1 and 4.2 and if $Th \to \infty$ and $nh^3 \to \infty$ for $T \to \infty$ and $n \to \infty$, it holds that*

$$\sqrt{Th}\widehat{\mu}_h^{V,NW}(x^j) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2} v^1(x^j)\big)$$

$$\sqrt{Th}\widehat{\mu}_h^{V,LL}(x^j) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2} v^1(x^j)\big),$$

where $v^1(x^j) = (f(x^j))^{-1}\, \mathbf{E}(a^{11}(X) \mid X^j = x^j)$.

The joint distribution of the vector of the $\widehat{\mu}_h^{V,i,NW}(x^j)$ is a multivariate (d^2-dimensional) normal distribution with covariances given by

$$\mathbf{cov}(\sqrt{Th}\widehat{\mu}_h^{i,V}(x^j), \sqrt{Th}\widehat{\mu}_h^{i',V}(x^j)) = \frac{\kappa^2(x^j)}{\kappa_0(x^j)f(x^j)}\mathbf{E}(a^{ii'}(X) \mid X^j = x^j)$$

and zero otherwise. The same holds in the local linear case.

Proof. To derive asymptotic normality, the distribution of $\widehat{t}_{h,l}^V$ has to be considered. To do so, decompose it into a discretization error and a stochastic error

$$\begin{aligned}\widehat{t}_{h,l}^V(x^j) =& \frac{1}{T}\sum_{k=1}^{nT}\widetilde{K}_{h,l}(x^j, X_{(k-1)\Delta}^j)\sum_{m=1}^{d}\int_{(k-1)\Delta}^{k\Delta}\sigma^{1m}(X_s)\,\mathrm{d}W_s^m \\ &= \frac{1}{T}\sum_{k=1}^{nT}\int_{(k-1)\Delta}^{k\Delta}\big(\widetilde{K}_{h,l}(x^j, X_s^j) - \widetilde{K}_{h,l}(x^j, X_{(k-1)\Delta}^j)\big)\sum_{m=1}^{d}\sigma^{1m}(X_s)\,\mathrm{d}W_s^m \\ &+ \frac{1}{T}\sum_{m=1}^{d}\int_0^T\widetilde{K}_{h,l}(x^j, X_s^j)\sigma^{1m}(X_s)\,\mathrm{d}W_s^m \\ &= J_{D,l}(x^j) + J_{T,l}(x^j).\end{aligned}$$

To bound the discretization error, calculate

$$\begin{aligned}\mathbf{E}(J_{D,l}(x^j))^2 &= nT^{-1}\,\mathbf{E}\Big(\int_0^{\Delta}\big(\widetilde{K}_{h,l}(x^j, X_s^j) - \widetilde{K}_{h,l}(x^j, X_0^j)\big)\sum_{m=1}^{d}\sigma^{1m}(X_s)\,\mathrm{d}W_s^m\Big)^2 \\ &= nT^{-1}\,\mathbf{E}\Big(\int_0^{\Delta}\big(\widetilde{K}_{h,l}(x^j, X_s^j) - \widetilde{K}_{h,l}(x^j, X_0^j)\big)^2 a^{11}(X_s)\,\mathrm{d}s\Big) \\ &\leq cnT^{-1}h^{-4}\,\mathbf{E}\Big((\max_{0\leq s\leq\Delta}|X_s - X_0|)^2\int_0^{\Delta}a^{11}(X_s)\,\mathrm{d}s\Big) \\ &\leq cT^{-1}h^{-4}\sup_{x\in\mathcal{G}}|a^{11}(x)|\,\mathbf{E}(\max_{0\leq s\leq\Delta}|X_s - X_0|)^2 \\ &= O(T^{-1}n^{-1}h^{-4}),\end{aligned}$$

where it is used that all covariances vanish. The last bound follows from the Burkholder-Davis-Grundy inequality. This yields in total

$$J_{D,l}(x^j) = O_P(T^{-1/2}n^{-1/2}h^{-2}) = o_P(T^{-1/2}h^{-1/2})$$

by assumption.
Next derive the asymptotic distribution of $\sqrt{hT}J_{T,l}(x^j)$. Note that for every T and x^j the functions $K_h(x^j, X_s^j)\sigma^{1l}(X_s)$ are progessively measurable and

$$h^{1/2}T^{-1/2}\int_0^T\widetilde{K}_{h,l}(x^j, X_s^j)\sigma^{1l}(X_s)\,\mathrm{d}s < \infty$$

with probability one. And for $T \to \infty$ it holds that

$$\begin{aligned} hT\langle J_{T,l}(x^j), J_{T,l}(x^j)\rangle &= hT^{-1}\int_0^T \widetilde{K}_{h,l}(x^j, X_s^j)^2 a^{11}(X_s)\,\mathrm{d}s \\ &\xrightarrow{P} \kappa_l^2(x^j) f(x^j)\,\mathbf{E}(a^{11}(X) \mid X^j = x^j) \\ hT\langle J_{T,l_1}(x^j), J_{T,l_2}(x^j)\rangle &= hT^{-1}\int_0^T \widetilde{K}_{h,l_1}(x^j, X_s^j)\widetilde{K}_{h,l_2}(x^j, X_s^j) a^{11}(X_s)\,\mathrm{d}s \\ &\xrightarrow{P} \kappa_{(l_1+l_2)/2}^2(x^j) f(x^j)\,\mathbf{E}(a^{11}(X) \mid X^j = x^j). \end{aligned}$$

Applying Proposition 1.21 in Kutoyants (2004) the following asymptotic distribution is obtained

$$\sqrt{hT}\begin{pmatrix} \widehat{t}_{h,0}^V \\ \widehat{t}_{h,1}^V \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, Vf(x^j)^2 v(x^j)) \qquad \text{where } V = \begin{pmatrix} \kappa_0^2 & \kappa_1^2 \\ \kappa_1^2 & \kappa_2^2 \end{pmatrix}.$$

Using the convergence results of Lemma 4.1, the statement of the Lemma follows for both estimators, recalling their definition. □

Next, some preliminary results for estimating the diffusion matrix are presented. Again the dimension index of the estimators will be omitted. To decompose the marginal kernel estimators, recall that by applying Itô's lemma

$$\begin{aligned} (X_{k\Delta}^1 - X_{(k-1)\Delta}^1)(X_{k\Delta}^2 - X_{(k-1)\Delta}^2) = \int_{(k-1)\Delta}^{k\Delta} a^{12}(X_s)\,\mathrm{d}s + \\ \int_{(k-1)\Delta}^{k\Delta} (X_s^1 - X_{(k-1)\Delta}^1)\,\mathrm{d}X_s^2 + \int_{(k-1)\Delta}^{k\Delta} (X_s^2 - X_{(k-1)\Delta}^2)\,\mathrm{d}X_s^1. \end{aligned}$$

Based on this decompose for $l = 0, 1$

$$\begin{aligned} \widehat{r}_{h,l}^B(x^j) &= \frac{1}{T}\sum_{k=1}^{nT} \widetilde{K}_{h,l}(x^j, X_{(k-1)\Delta}^j)\int_{(k-1)\Delta}^{k\Delta} a^{12}(X_s)\,\mathrm{d}s \\ \widehat{r}_{h,l}^V(x^j) &= \frac{1}{T}\sum_{k=1}^{nT} \widetilde{K}_{h,l}(x^j, X_{(k-1)\Delta}^j)\int_{(k-1)\Delta}^{k\Delta} (X_s^1 - X_{(k-1)\Delta}^1)\,\mathrm{d}X_s^2 \\ &\quad + \int_{(k-1)\Delta}^{k\Delta} (X_s^2 - X_{(k-1)\Delta}^2)\,\mathrm{d}X_s^1. \end{aligned}$$

Replacing $\widehat{t}_{h,l}^*(x^j)$ with $\widehat{r}_{h,l}^*(x^j)$ in equations (4.21) and (4.22) gives the decomposition of the marginal estimators of the diffusion matrix $\widehat{a}_h^{NW}(x^j)$ and $\widehat{a}_h^{LL}(x^j)$. Next, convergence results for these estimators will be derived.

Lemma 4.5. *Under Assumptions 4.1 and 4.2 it holds that*

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_h^{NW,B}(x^j) - \mathbf{E}(a^{12}(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log n}{(nTh)^{1/2}}\Big)$$

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_h^{LL,B}(x^j) - \mathbf{E}(a^{12}(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log n}{(nTh)^{1/2}}\Big)$$

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_{j,h}^{LL,B}(x^j) - h\partial_j \mathbf{E}(a^{12}(X) \mid X^j = x^j)| = O_P\Big(h^2 + \frac{\log n}{(nTh)^{1/2}}\Big),$$

if the terms on the right hand side converge to zero as $n \to \infty$ and $T \to \infty$.

Proof. Replacing $\mu^1(X_s)$ by $a^{12}(X_s)$, the structure of $\widehat{r}_{h,l}^B(x^j)$ and $\widehat{t}_{h,l}^B(x^j)$ are the same. Therefore, the proof of this lemma is analogous to the proof of Lemma 4.2 and therefore omitted. □

Lemma 4.6. *Under Assumptions 4.1 and 4.2 it holds that*

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_h^{NW,V}(x^j)| = O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big)$$

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_h^{LL,V}(x^j)| = O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big) \qquad \sup_{x^j \in \mathcal{G}^j} |\widehat{a}_{j,h}^{LL,V}(x^j)| = O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big),$$

if the terms on the right hand side converge to zero as $n \to \infty$ and $T \to \infty$.

Proof. Write the numerator parts of the estimators as sum of α-mixing random variables

$$\widehat{r}_{h,l}^V(x^j) = \frac{1}{T} \sum_{i=0}^{T-1} S_{t,n}(x^j),$$

where

$$S_{t,n}(x^j) = \sum_{k=0}^{n-1} \widetilde{K}_{h,l}(x^j, X_{t+k\Delta}^j) Z_{t+k\Delta},$$

with

(4.23)

$$Z_{t+k\Delta} = \big((X_{t+(k+1)\Delta}^1 - X_{t+k\Delta}^1)(X_{t+(k+1)\Delta}^2 - X_{t+k\Delta}^2) - \int_{t+k\Delta}^{t+(k+1)\Delta} a^{12}(X_s)\,\mathrm{d}s\big).$$

Clearly $\mathbf{E}\, S_{t,n}(x^j) = 0$. For the second moment it will be shown below that

$$\mathbf{E}(S_{0,n}(x^j))^2 = O\big(n^{-1}h^{-1}\big). \tag{4.24}$$

Cramer's conditions with a constant $(Th)^{-1}$ can then be verified by seeing that $\sup_k |Z_{t+k\Delta}| = O_P(n^{-1})$. This allows to apply an exponential inequality as above. Thus, it remains to show (4.24). First it holds that

$$\begin{aligned}(4.25)\quad \mathbf{E}(S_{0,n}(x^j))^2 = \sum_{k=0}^{n-1} \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})^2 Z^2_{0,k} \\ + \sum_{k=0}^{n-2} \sum_{k'=k+1}^{n-1} \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta}) Z_{0,k} \widetilde{K}_{h,l}(x^j, X^j_{k'\Delta}) Z_{0,k'}.\end{aligned}$$

Start with the first sum and resolve the square to obtain

$$\begin{aligned}\mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})^2 Z^2_{0,k} &= \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})^2 (X^1_{(k+1)\Delta} - X^1_{k\Delta})^2 (X^2_{(k+1)\Delta} - X^2_{k\Delta})^2 \\ &\quad - 2\,\mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})^2 (X^1_{(k+1)\Delta} - X^1_{k\Delta})(X^2_{(k+1)\Delta} - X^2_{k\Delta}) \int_{k\Delta}^{(k+1)\Delta} a^{12}(X_s)\,\mathrm{d}s \\ &\quad + \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{l\Delta})^2 \Big(\int_{k\Delta}^{(k+1)\Delta} a^{12}(X_s)\,\mathrm{d}s \Big)^2 \\ &= S_1 + S_2 + S_3.\end{aligned}$$

These three quantities are investigated separately. First recall that

$$(X^1_{(k+1)\Delta} - X^1_{k\Delta})^2 = \int_{k\Delta}^{(k+1)\Delta} a^{11}(X_s)\,\mathrm{d}s + O_P\big(\frac{(\log n)^{1/2}}{n^{3/2}}\big)$$

and then an application of the mean value theorem yields

$$\begin{aligned}S_1 &= \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})^2 \int_{k\Delta}^{(k+1)\Delta} a^{11}(X_s)\,\mathrm{d}s \int_{k\Delta}^{(k+1)\Delta} a^{22}(X_s)\,\mathrm{d}s + O\big(\frac{(\log n)^{1/2}}{n^{5/2}h}\big) \\ &= \frac{\kappa^2(x^j)}{n^2 h}\, \mathbf{E}(a^{11}(X_s) a^{22}(X_s) \mid X^j_s = x^j)(1 + o(1)).\end{aligned}$$

Because the drift is bounded, it holds that

$$S_2 = O(n^{-3}h^{-1}).$$

Finally, the last term satisfies

$$S_3 = \frac{\kappa^2(x^j)}{n^2 h}\, \mathbf{E}((a^{12}(X_s))^2 \mid X^j_s = x^j)(1 + o(1)).$$

In total the first term in equation (4.25) satisfies the desired rate. Because of the stationarity, the second term is bounded by $n \sum_{k'=1}^{n-1} \mathbf{E}\, \widetilde{K}_{h,l}(x^j, X^j_0) Z_{0,k}$

$\widetilde{K}_{h,l}(x^j, X_0^j)Z_{0,k'}$. This will be decomposed into three parts

$$\begin{aligned}
n\sum_{k'=1}^{n-1} &\mathbf{E}\,\widetilde{K}_{h,l}(x^j, X_0^j)(X_\Delta^1 - X_0^1)(X_\Delta^2 - X_0^2) \\
&\qquad\qquad \times \widetilde{K}_{h,l}(x^j, X_{k'\Delta}^j)(X_{(k'+1)\Delta}^1 - X_{k'\Delta}^1)(X_{(k'+1)\Delta}^2 - X_{k'\Delta}^2) \\
&= \frac{1}{n^3}\sum_{k'=1}^{n-1}\mathbf{E}\,\widetilde{K}_{h,l}(x^j, X_0^j)\widetilde{K}_{h,l}(x^j, X_{k'\Delta}^j) \\
&\qquad \times \mu^1(X_0)\mu^2(X_0)\mu^1(X_{k'\Delta})\mu^2(X_{k'\Delta})(1+o(1)) \\
&= O(n^{-2}),
\end{aligned}$$

because the density and the drift are bounded. For the second part, we get

$$\begin{aligned}
&n\sum_{k'=1}^{n-1}\mathbf{E}\,\widetilde{K}_{h,l}(x^j, X_0^j)(X_\Delta^1 - X_0^1)(X_\Delta^2 - X_0^2)\widetilde{K}_{h,l}(x^j, X_{k'\Delta}^j)\int_{k'\Delta}^{(k'+1)\Delta} a^{12}(X_s)\,\mathrm{d}s \\
&= \frac{1}{n^2}\sum_{k'=1}^{n-1}\mathbf{E}\,\widetilde{K}_{h,l}(x^j, X_0^j)\widetilde{K}_{h,l}(x^j, X_{k'\Delta}^j)\mu^1(X_0)\mu^2(X_0)a^{12}(X_{k'\Delta})(1+o(1)) \\
&= O(n^{-1})
\end{aligned}$$

and finally

$$n\sum_{k'=1}^{n-1}\mathbf{E}\,\widetilde{K}_{h,l}(x^j, X_0^j)\int_0^\Delta a^{12}(X_s)\,\mathrm{d}s\widetilde{K}_{h,l}(x^j, X_{k'\Delta}^j)\int_{k'\Delta}^{(k'+1)\Delta} a^{12}(X_s)\,\mathrm{d}s = O(n^{-1}).$$

Then, the covariances are of smaller order and in total equation (4.25) is established. □

Finally, the asymptotic distribution of the variance parts is derived.

Lemma 4.7. *Under Assumptions 4.1 and 4.2 and if $nTh \to \infty$ and $nh^3 \to \infty$ for $T \to \infty$ and $n \to \infty$ it holds that*

$$\begin{aligned}
\sqrt{nTh}\widehat{a}_h^{NW,V}(x^j) &\xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}v^{12}(x^j)\big) \\
\sqrt{nTh}\widehat{a}_h^{LL,V}(x^j) &\xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \frac{\kappa_0^2(x^j)}{\kappa_0(x^j)^2}v^{12}(x^j)\big),
\end{aligned}$$

where $v^{12}(x^j) = (f(x^j))^{-1}\,\mathbf{E}\big(a^{11}(X)a^{22}(X)(a^{12}(X))^2 \mid X^j = x^j\big)$.

The joint distribution of the vector of the $\widehat{a}_h^{ij,NW,V}(x^k)$ is a multivariate (d^3-dimensional) normal distribution with covariances given by

$$\mathbf{cov}(\sqrt{nTh}\widehat{a}_h^{ij,V}(x^k), \sqrt{nTh}\widehat{a}_h^{i'j',V}(x^k)) = \frac{\kappa^2(x^k)}{f(x^k)}\,\mathbf{E}(a^{ij}(X)a^{i'j'}(X) \mid X^k = x^k)$$

and zero otherwise (if $\{i,j\} \neq \{i',j'\}$). The same holds for the local linear estimator.

Proof. The distribution of $\widehat{r}_{h,l}(x^j)$ has to be derived. Write

$$\sqrt{nTh}\widehat{r}^V_{h,l}(x^j) = \sum_{k=0}^{nT-1} M_k(x^j),$$

with

$$M_k(x^j) = \sqrt{nTh}\widetilde{K}_{h,l}(x^j, X_{k\Delta})Z_k,$$

where Z_k is given in equation (4.23). Denote the σ-Algebra generated by $X_0, \ldots,$ X_k with $\mathcal{F}_k^{T,n} = \sigma(X_l, l = 0, \Delta, \ldots, k\Delta)$. From repeated application of the Burkholder-Davis-Grundy inequality (as in Florens-Zmirou, 1993) it follows that

$$\sum_{k=0}^{nT-1} \mathbf{E}(M_k(x^j) \mid \mathcal{F}_k^{T,n}) \xrightarrow{P} 0$$
$$\sum_{k=0}^{nT-1} \mathbf{E}((M_k(x^j))^2 \mid \mathcal{F}_k^{T,n}) \xrightarrow{P} \kappa_l^2(x^j)\frac{1}{T}\int_0^T a^{12}(X_s)\mathbf{1}_{\{X_s^j = x^j\}}\,\mathrm{d}s$$
$$\sum_{k=0}^{nT-1} \mathbf{E}(|M_k(x^j)|^3 \mid \mathcal{F}_k^{T,n}) \xrightarrow{P} 0.$$

Applying Lemma 2 in Florens-Zmirou (1993) it follows that $\sqrt{nTh}\widehat{r}^V_{h,l}(x^j)$ converges for $n \to \infty$ against a continuous martingale with increasing process given by $\kappa_l^2(x^j)\frac{1}{T}\int_0^T a^{12}(X_s)\mathbf{1}_{\{X_s^j = x^j\}}\,\mathrm{d}s$. By Knight's theorem mixed asymptotic normality for fixed T follows (analogously to Brugiére, 1991). Because for all $x^j \in \mathcal{G}^j$ the convergence $\frac{1}{T}\int_0^T a^{12}(X_s)\mathbf{1}_{\{X_s = x\}}\,\mathrm{d}s \xrightarrow{P} \mathbf{E}(a^{12}(X_s) \mid X_s^j = x^j)f(x^j)$ holds, the asymptotic normality for $T \to \infty$ follows by Knight's theorem. □

Appendix II: Proof of the Theorems

Proof of Theorem 4.1

The verification of Assumptions (A1)–(A6), (A8), (A9) in Mammen, Linton and Nielsen (1999) allows to apply their Theorems 1, 2 and 3, which will yield the statement of the theorem. These assumptions are shown using the lemmata of the last section. Subsequently all integrals are taken over $\mathcal{G}$ (resp. the corresponding projection).

(A1) It holds by Assumption 4.1.2 for all $i \neq j$

$$\int \frac{f^2(x^i, x^j)}{f(x^i)f(x^j)} \,\mathrm{d}x^i \,\mathrm{d}x^j < \infty.$$

(A2) Consists of three parts, all of them following by applications of Lemma 4.1.

(i)
$$\int \Big(\frac{\widehat{f}_{h,0}(x^j) - f(x^j)}{f(x^j)}\Big)^2 f(x^j)\,\mathrm{d}x^j \leq \sup_{x^j \in \mathcal{G}^j} |\widehat{f}_{h,0}(x^j) - f(x^j)|^2 \int (f(x^j))^{-1}\,\mathrm{d}x^j$$
$$= O_P\Big(h^4 + \frac{(\log T)^2}{Th}\Big) = o_P(1),$$

because the density is bounded from below.

(ii)
$$\int \Big(\frac{\widehat{f}_{h,(0,0)}(x^i, x^j) - f(x^i, x^j)}{f(x^i)f(x^j)}\Big)^2 f(x^i)f(x^j)\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$\leq \sup_{(x^i,x^j)\in \mathcal{G}^i \times \mathcal{G}^j} |\widehat{f}_{h,(0,0)}(x^i, x^j) - f(x^i, x^j)|^2 \int (f(x^i)f(x^j))^{-1}\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$= O_P\Big(h^4 + \frac{(\log T)^2}{Th^2}\Big) = o_P(1).$$

And the third part of the assumption

(iii)
$$\int \Big(\frac{\widehat{f}_{h,(0,0)}(x^i, x^j)}{f(x^i)\widehat{f}_{h,0}(x^j)} - \frac{f(x^i, x^j)}{f(x^i)f(x^j)}\Big)^2 f(x^i)f(x^j)\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$= \int (\widehat{f}_{h,(0,0)}(x^i, x^j))^2 \big((\widehat{f}_{h,0}(x^j))^{-1} - (f(x^j))^{-1}\big)^2 \frac{f(x^j)}{f(x^i)}\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$+ 2\int \widehat{f}_{h,(0,0)}(x^i, x^j)\big((\widehat{f}_{h,0}(x^j))^{-1} - (f(x^j))^{-1}\big)$$
$$\times \big(\widehat{f}_{h,(0,0)}(x^i, x^j) - f(x^i, x^j)\big)\frac{f(x^j)}{f(x^i)}\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$+ \int (\widehat{f}_{h,(0,0)}(x^i, x^j) - f(x^i, x^j))^2 (f(x^i)f(x^j))^{-1}\,\mathrm{d}x^i\,\mathrm{d}x^j$$
$$= o_P(1) + O_P\Big(h^4 + \frac{(\log T)^2}{Th^{3/2}}\Big) + O_P\Big(h^4 + \frac{(\log T)^2}{Th^2}\Big) = o_P(1)$$

by Lemma 4.1.

Note that at the boundary (of length h), the leading bias term is of order h and then the same results hold.

(A3) With probability tending to one it holds that

$$\int \big(\widehat{\mu}^{NW}(x^j)\big)^2 f(x^j)\,\mathrm{d}x^j \le C.$$

This follows directly from (A5) below, since $\widehat{\mu}_h^{NW}(x^j) = \widehat{\mu}_h^{NW,B}(x^j) + \widehat{\mu}_h^{NW,V}(x^j)$.

(A4) By similar arguments as (A3) and the last part of (A2) one sees that

$$\sup_{x^j \in \mathcal{G}^j} \int \frac{(\widehat{f}_{h,(0,0)}(x^i,x^j))^2}{(\widehat{f}_{h,0}(x^j))^2 f(x^i)}\,\mathrm{d}x^i \le \sup_{x^j \in \mathcal{G}^j} (f(x^j))^{-1} \int \frac{f(x^i,x^j))^2}{f(x^j)f(x^i)}\,\mathrm{d}x^i,$$

with probability tending to one, using Lemma 4.1. The right hand side is bounded by Assumption 4.1 (see A1).

(A5) Applying Lemmata 4.2 and 4.3 together with the quadratic integrability of the unknown function $\mu(\cdot)$. Starting with the variance part we have that

$$\int (\widehat{\mu}_h^{NW,V}(x^j))^2 f(x^j)\,\mathrm{d}x^j \le (\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{NW,V}(x^j)|)^2$$

which is bounded by an arbitrary positive constant with probability tending to one.

For the bias part

$$\int (\widehat{\mu}_h^{NW,B}(x^j))^2 f(x^j)\,\mathrm{d}x^j \le 2\int \mathbf{E}(\mu^1(X) \mid X^j = x^j) f(x^j)\,\mathrm{d}x^j \\ + 2\int (\widehat{\mu}_h^{NW,B}(x^j) - \mathbf{E}(\mu^1(X) \mid X^j = x^j))^2 f(x^j)\,\mathrm{d}x^j.$$

This is bounded by the first part of the right hand side plus a constant with probability tending to one.

(A6) First, decompose using the triangle inequality

$$\sup_{x^j \in \mathcal{G}^j} \Big| \int \frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^j)} \widehat{\mu}_h^{NW,V}(x^i)\,\mathrm{d}x^i \Big| \le \sup_{x^j \in \mathcal{G}^j} \Big| \int \frac{f(x^i,x^j)}{f(x^i)f(x^j)} \widehat{t}_{h,0}^V(x^i)\,\mathrm{d}x^i \Big| \\ + \sup_{x^j \in \mathcal{G}^j} \Big| \int \Big(\frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^i)\widehat{f}_{h,0}(x^j)} - \frac{f(x^i,x^j)}{f(x^i)f(x^j)} \Big) \widehat{t}_{h,0}^V(x^i)\,\mathrm{d}x^i \Big|. \tag{4.26}$$

The two terms are investigated separately. Linearizing in the usual way, it holds for the second supremum that

$$\begin{aligned}
&\sup_{x^j\in\mathcal{G}^j}\Big|\int\Big(\frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^i)\widehat{f}_{h,0}(x^j)}-\frac{f(x^i,x^j)}{f(x^i)f(x^j)}\Big)\widehat{t}^V_{h,0}(x^i)\,\mathrm{d}x^i\Big|\\
&\quad\le \sup_{(x^i,x^j)\in\mathcal{G}^j\times\mathcal{G}^i}|\widehat{f}_{h,(0,0)}(x^i,x^j)-f(x^i,x^j)|\sup_{x^i\in\mathcal{G}^i}|\widehat{t}^V_{h,0}(x^i)|\sup_{x^j\in\mathcal{G}^j}|(f(x^j))^{-1}|\\
&\quad\quad\times\int(f(x^i))^{-1}\,\mathrm{d}x^i(1+o_P(1))\\
&\quad= O_P\Big(h^2+\frac{\log T}{(Th^2)^{1/2}}\Big)O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big)\\
&\quad= o_P(h^2),
\end{aligned}$$

by Lemma 4.2 and Lemma 4.1.
The first part in equation (4.26) can be rewritten as

$$\sup_{x^j\in\mathcal{G}^j}\Big|\int\frac{f(x^i,x^j)}{f(x^i)f(x^j)}\widehat{t}^V_{h,0}(x^i)\,\mathrm{d}x^i\Big| = \sup_{x^j\in\mathcal{G}^j}\Big|\frac{1}{nT}\sum_{k=1}^{nT}\xi_k(x^j)\Big|,$$

where

$$\begin{aligned}
\xi_k(x^j)=\int\frac{f(X^i_{(k-1)\Delta}+hu,x^j)}{f(X^i_{(k-1)\Delta}+hu)f(x^j)}K(X^i_{(k-1)\Delta}+hu,X^i_{(k-1)\Delta})\,\mathrm{d}u&\\
\times\sum_{l=1}^{d}\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta}\sigma^{1l}(X_s)\,\mathrm{d}W^l_s.&
\end{aligned}$$

The first integral in $\xi_k(x^j)$ is bounded, because the density is bounded from above and below. Using the Itô-Isometry it follows that $T^{-2}\,\mathbf{E}(\int_0^T\sigma^{1l}(X_s)\,\mathrm{d}W^l_s)^2 = O(T^{-1})$. In total it holds that

$$\sup_{x^j\in\mathcal{G}^j}\Big|\frac{1}{nT}\sum_{k=1}^{nT}\xi_k(x^j)\Big| = O_P(T^{-1/2}) = o_P(h^2).$$

Combining the two results, it is obtained that

$$\sup_{x^j\in\mathcal{G}^j}\Big|\int\frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^j)}\widehat{\mu}^{NW,V}_h(x^i)\,\mathrm{d}x^i\Big| = o_P(h^2).$$

Because $\int\widehat{f}_{h,0}(x^j)=O_P(1)$ it follows directly that

$$\Big(\int\Big(\int\frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^j)}\widehat{\mu}^{NW,V}_h(x^i)\,\mathrm{d}x^i\Big)^2\widehat{f}_{h,0}(x^j)\,\mathrm{d}x^j\Big)^{1/2} = o_p(h^2).$$

This completes (A6).

(A8) By linearization, it is immediately clear that

$$\begin{aligned}
&\sup_{x^j \in \mathcal{G}^j} \int \Big| \frac{\widehat{f}_{h,(0,0)}(x^i, x^j)}{\widehat{f}_{h,0}(x^i)\widehat{f}_{h,0}(x^j)} - \frac{f(x^i, x^j)}{f(x^i)f(x^j)} \Big| f(x^i)\,\mathrm{d}x^i \\
&\quad \le \sup_{x^j \in \mathcal{G}^j} |\widehat{f}_{h,(0,0)}(x^i, x^j) - f(x^i, x^j)| \sup_{x^j \in \mathcal{G}^j} |(f(x^j))^{-1}|(1 + o_P(1)) \\
&\quad = O_P\Big(h^2 + \frac{\log T}{(Th^2)^{1/2}}\Big) = o_P(1),
\end{aligned}$$

by Lemma 4.1 and Assumption 4.1.2.

(A9) It has to be shown that

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{NW,B}(x^j) - \widehat{\nu}_{n,T,j}(x^j)| = o_P(h^2), \tag{4.27}$$

where

$$\begin{aligned}
\widehat{\nu}_{n,T,j}(x^j) = \alpha_{n,T,j}(x^j) + \sum_{i \neq j} \int \alpha_{n,T,i}(x^i) \frac{\widehat{f}_{h,(0,0)}(x^j, x^i)}{\widehat{f}_{h,0}(x^j)}\,\mathrm{d}x^i \\
+ h^2 \int \beta_\mu(x) \frac{f(x)}{f(x^j)}\,\mathrm{d}x^{-j}
\end{aligned} \tag{4.28}$$

and

$$\alpha_{n,T,j}(x^j) = \mu^{1,j}(x^j) + \partial_j \mu^{1,j}(x^j) \frac{h\kappa_1(x^j)}{\kappa_0(x^j)}.$$

Statement (4.27) is shown by decomposing $\widehat{\mu}_h^{NW,B}(x^j)$ appropriately. An application of the mean value theorem yields

$$\begin{aligned}
\widehat{\mu}_h^{NW,B}(x^j) &= \frac{1}{T\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} K_h(X^j_{k\Delta}, x^j) \sum_{i=1}^d \int_{k\Delta}^{(k+1)\Delta} \mu^{1,i}(X^i_s)\,\mathrm{d}s \\
&= \sum_{i=1}^d \frac{1}{T\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} K_h(X^j_{k\Delta}, x^j)(\Delta\mu^{1,i}(X^i_{k\Delta}) \\
&\qquad + \int_{k\Delta}^{(k+1)\Delta} \partial_i \mu^{1,i}(\xi^i_s)(X^i_s - X^i_{k\Delta})\,\mathrm{d}s) \\
&= \sum_{i=1}^d \frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} K_h(X^j_{k\Delta}, x^j)\mu^{1,i}(X^i_{k\Delta}) + O_P\big(n^{-1/2}\big).
\end{aligned} \tag{4.29}$$

This holds by the Burkholder-Davis-Grundy inequality and because the derivative is bounded. Next, lower order terms are omitted and the cases with $i = j$ and

$i \neq j$ are treated separately. Starting with $i = j$ we obtain from standard kernel calculations

$$
\begin{aligned}
&\frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} K_h(X^j_{k\Delta}, x^j)\mu^{1,j}(X^j_{k\Delta}) \\
&\quad = \mu^{1,j}(x^j) + \frac{\mathbf{E}\big(K_h(X^j_{k\Delta}, x^j)\mu^{1,j}(X^j_{k\Delta})\big) - \mu^{1,j}(x^j)\,\mathbf{E}\,K_h(X^j_{k\Delta}, x^j)}{\mathbf{E}\,K_h(X^j_{k\Delta}, x^j)} + R_{n,T,j}(x^j) \\
&\quad = \mu^{1,j}(x^j) + h\frac{\kappa_1(x^j)}{\kappa_0(x^j)}\partial_j\mu^{1,j}(x^j)\frac{\kappa_1(x^j)}{\kappa_0(x^j)} + h^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\frac{\partial_j\mu^{1,j}(x^j)}{f(x^j)}\int \partial_j f(x)\,\mathrm{d}x^{-j} \\
&\qquad\qquad + h^2\frac{1}{2}\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\partial_j^2\mu^{1,j}(x^j) + R_{n,T,j}(x^j) + o_P(h^2)
\end{aligned}
$$

$$
\begin{aligned}
&\quad = \alpha_{n,T,j}(x^j) + h^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\int\Big(\frac{\partial_j f(x)}{f(x)}\partial_j\mu^{1,j}(x^j) + \frac{1}{2}\partial_j^2\mu^{1,j}(x^j)\Big)\frac{f(x)}{f(x^j)}\,\mathrm{d}x^{-j} \\
&\qquad\qquad + o_P(h^2).
\end{aligned}
\tag{4.30}
$$

The last equation holds because of

$$
\begin{aligned}
&\sup_{x^j\in\mathcal{G}^j} |R_{n,T,j}(x^j)| \\
&\qquad = \sup_{x^j\in\mathcal{G}^j}\left|\frac{1}{nT}\sum_{k=0}^{nT-1}\frac{K_h(X^j_{k\Delta}, x^j)\mu^{1,j}(X^j_{k\Delta})}{\widehat{f}_{h,0}(x^j)} - \frac{\mathbf{E}\big(K_h(X^j_{k\Delta}, x^j)\mu^{1,j}(X^j_{k\Delta})\big)}{\mathbf{E}\,K_h(X^j_{k\Delta}, x^j)}\right| \\
&\qquad = O_P\Big(\frac{\log T}{(nTh)^{1/2}}\Big).
\end{aligned}
$$

This is shown as in the proof in Lemma 4.2.

Next, turn to the cases with $i \neq j$ in equation (4.29). Here, a Taylor expansion of $\mu^1(X^i_{k\Delta})$ around x^i and $\int K_h(X^i_{k\Delta}, x^i)\,\mathrm{d}x^i = 1$ are used to obtain

$$
\begin{aligned}
&\frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} K_h(X^j_{k\Delta}, x^j)\mu^{1,i}(X^i_{k\Delta}) \\
&\quad= \frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} \int K_h(X^j_{k\Delta}, x^j) K_h(X^i_{k\Delta}, x^i)\mu^{1,i}(X^i_{k\Delta})\,\mathrm{d}x^i \\
&\quad= \int \mu^{1,i}(x^i)\frac{\widehat{f}_{h,(0,0)}(x_i, x_j)}{\widehat{f}_{h,0}(x^j)}\,\mathrm{d}x^i \\
&\qquad+ \frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} \int K_h(X^j_{k\Delta}, x^j) K_h(X^i_{k\Delta}, x^i)(X^i_{k\Delta} - x^i)\partial_i\mu^{1,i}(x^i)\,\mathrm{d}x^i \\
&\qquad+ \frac{1}{2}\frac{1}{nT\widehat{f}_{h,0}(x^j)} \sum_{k=0}^{nT-1} \int K_h(X^j_{k\Delta}, x^j) K_h(X^i_{k\Delta}, x^i)(X^i_{k\Delta} - x^i)^2\partial_i^2\mu^{1,i}(x^i)\,\mathrm{d}x^i \\
&\qquad\quad+ o_P(h^2) \\
&\quad= \int \mu^{1,i}(x^i)\frac{\widehat{f}_{h,(0,0)}(x^i, x^j)}{\widehat{f}_{h,0}(x^j)}\,\mathrm{d}x^i + h\int \partial_i\mu^{1,i}(x^i)\frac{\kappa_1(x^j)}{\kappa_0(x^j)}\frac{\widehat{f}_{h,(0,0)}(x^i, x^j)}{\widehat{f}_{h,0}(x^j)}\,\mathrm{d}x^i \\
&\qquad+ h^2\frac{\kappa_2(x^j)}{\kappa_0(x^i)}\int \frac{\partial_i f(x^i)}{f(x^j)}\partial_i\mu^{1,i}(x^i)\,\mathrm{d}x^{-j} + h^2\frac{1}{2}\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\int \partial_i^2\mu^{1,i}(x^i)\frac{f(x)}{f(x^j)}\,\mathrm{d}x^{-j} \\
&\qquad+ R_{n,T,j,i}(x^j) + o_P(h^2)
\end{aligned}
$$

(4.31)

$$
\begin{aligned}
&\quad= \int \alpha_{n,T,i}(x^i)\frac{\widehat{f}_{h,(0,0)}(x^i, x^j)}{\widehat{f}_{h,0}(x^j)}\,\mathrm{d}x^i \\
&\qquad+ h^2\frac{\kappa_2(x^j)}{\kappa_0(x^j)}\int \Big(\frac{\partial_i f(x)}{f(x)}\partial_i\mu^{1,i}(x^i) + \frac{1}{2}\partial_i^2\mu^{1,j}(x^i)\Big)\frac{f(x)}{f(x^j)}\,\mathrm{d}x^{-j} + o_P(h^2).
\end{aligned}
$$

By the same arguments as above it follows that $\sup_{x^j\in\mathcal{G}^j} |R_{n,T,j,i}(x^j)| = o_P(h^2)$. The statement (4.27) follows from (4.29), (4.30) and (4.31) together.
From equation (4.29) it follows directly that

$$\int (\widehat{\mu}_h^{NW,B}(x^j) - \widehat{\nu}_{n,T,j}(x^j))^2 f(x^j)\,\mathrm{d}x^j = o_P(h^4). \tag{4.32}$$

Finally, it has to be shown that

$$
\begin{aligned}
\int \alpha_{n,T,j}(x^j)\widehat{f}_{h,0}(x^j)\,\mathrm{d}x^j &= \iint \mu^{1,j}(x^j)K_h(x^j, u)f(u)\,\mathrm{d}x^j\,\mathrm{d}u \\
&\quad+ h\iint \partial_j\mu^{1,j}(x^j)\frac{h\kappa_1(x^j)}{\kappa_0(x^j)}K_h(x^j, u)f(u)\,\mathrm{d}x^j\,\mathrm{d}u + O_P(T^{-1/2}) \\
&= b_h^{1,j} + o_P(h^2). \qquad (4.33)
\end{aligned}
$$

The convergence in probability follows from $\mathbf{E}(\int \alpha_{n,T,j}(x^j)\widehat{f}_h(x^j)\,\mathrm{d}x^j)^2 = O(T^{-1})$, which is shown by an application of Davydov's inequality.
Then (A9) is proofed by (4.27), (4.32) and (4.33).

Asymptotic distribution Now Theorems 1,2 and 3 of Mammen, Linton and Nielsen (1999) can be applied and the convergence of the algorithm follows. For the asymptotic distribution we write

$$\begin{aligned}\widetilde{\mu}_h^{NW}(x^j) = \widehat{\mu}_h^{NW,V}(x^j) + \nu_{n,T,j}(x^j)\\ + (\widetilde{\mu}_h^{NW,V}(x^j) - \widehat{\mu}_h^{NW,V}(x^j)) + (\widetilde{\mu}_h^{NW,B}(x^j) - \nu_{n,T,j}(x^j)).\end{aligned}$$

where $\nu_{n,T,j}(x^j) = \alpha_{n,T,j}(x^j) - \gamma_{n,T,j} + h^2\beta_\mu^j(x^j)$. The two terms in brackets are of lower stochastic order $(o_P(h^2))$ uniformly in x^j, which is shown by Theorem 2 and 3 in Mammen, Linton and Nielsen (1999). Then, the asymptotic result follows immediately. □

Proof of Theorem 4.3

As many calculation are analogous to the proof of theorem 4.1, only on the changes will be highlighted. Obviously, (A1), (A2), (A4) and (A8) are unchanged. (A5) is shown analogously, using Lemmata 4.5 and 4.6. (A3) follows from that.

(A6) From the the triangle inequality it follows that

$$\begin{aligned}\sup_{x^j\in\mathcal{G}^j}\Big|\int \frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^j)}\widehat{a}_h^{NW,V}(x^i)\,\mathrm{d}x^i\Big| \le \sup_{x^j\in\mathcal{G}^j}\Big|\int \frac{f(x^i,x^j)}{f(x^i)f(x^j)}\widehat{r}_h^{NW,V}(x^i)\,\mathrm{d}x^i\Big|\\ + \sup_{x^j\in\mathcal{G}^j}\Big|\int\Big(\frac{\widehat{f}_{h,(0,0)}(x^i,x^j)}{\widehat{f}_{h,0}(x^i)\widehat{f}_{h,0}(x^j)} - \frac{f(x^i,x^j)}{f(x^i)f(x^j)}\Big)\widehat{r}_h^{NW,V}(x^i)\,\mathrm{d}x^i\Big|.\end{aligned}$$

Usual linearization shows that the second supremum is of $o_P(h^2)$. The first part can now be written as

$$\sup_{x^j\in\mathcal{G}^j}\Big|\int \frac{f(x^i,x^j)}{f(x^i)f(x^j)}\widehat{r}_h^{NW,V}(x^i)\,\mathrm{d}x^i\Big| = \sup_{x^j\mathcal{G}^j}\Big|\frac{1}{nT}\sum_{k=1}^{nT}\xi_k(x^j)\Big|,$$

where

$$\begin{aligned}\xi_k(x^j) = \int \frac{f(X^i_{(k-1)\Delta} + hu, x^j)}{f(X^i_{(k-1)\Delta} + hu)f(x^j)}K(X^i_{(k-1)\Delta} + hu, X^i_{(k-1)\Delta})\,\mathrm{d}u\\ \times n\big((X^1_{k\Delta} - X^1_{(k-1)\Delta})(X^2_{k\Delta} - X^2_{(k-1)\Delta}) - \int_{(k-1)\Delta}^{k\Delta} a^{12}(X_s)\,\mathrm{d}s\big).\end{aligned}$$

Using Burkholder-Davis-Grundy-inequality it follows that $\mathbf{E}\big((X^1_{k\Delta} - X^1_{(k-1)\Delta})$ $(X^2_{k\Delta} - X^2_{(k-1)\Delta}) - \int_{(k-1)\Delta}^{k\Delta} a^{12}(X_s)\,\mathrm{d}s\big)^2 = O(n^{-2})$. Then it follows that

$$\sup_{x^j \mathcal{G}^j} \Big| \frac{1}{nT} \sum_{k=1}^{nT} \xi_k(x^j) \Big| = O_P(n^{-1/2}T^{-1/2}).$$

(A9) Analogously to the drift case define

$$\alpha_{n,T,j}(x^j) = a^{12}(x^j) + h\partial_j a^{12}(x^j) \frac{\kappa_1(x^j)}{\kappa_0(x^j)},$$

for $j = 1, \dots, d$ while $\alpha_{0,n,T} = 0$. Quadratic integrability follows from the assumptions. From this follows the choice of $\gamma_{n,T,j} = \int \alpha_{n,T,j}(x^j) f(x^j)\,\mathrm{d}x^j$. Define $\widehat{\nu}_{n,T,j}(x^j)$ analogously to equation (4.28). To show that

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{a}_h^{NW,B}(x^j) - \widehat{\nu}_{n,T,j}(x^j)| = o_P(h^2)$$

the bias part $\widehat{a}^{12,B}(x^j)$ has to be decomposed analogously to the drift case. Then the desired rate follows.
This completes the proof. □

Proof of Theorem 4.2

For the local linear case Theorems 1', 2' and 3' in Mammen, Linton and Nielsen (1999) have to be used. Therefore the validity of their Assumptions (A1')–(A6'), (A8') and (A9')has to be shown. Define

$$V(x^j) = \begin{pmatrix} \kappa_0(x^j) & \kappa_1(x^j) \\ \kappa_1(x^j) & \kappa_2(x^j) \end{pmatrix} f(x^j)$$

$$U(x^j, x^i) = \begin{pmatrix} \kappa_0(x^j)\kappa_0(x^i) & \kappa_1(x^j)\kappa_0(x^i) \\ \kappa_0(x^j)\kappa_1(x^i) & \kappa_2(x^j)\kappa_2(x^i) \end{pmatrix} f(x^j, x^i)$$

as the limits of the matrices $V(x^j)$ and $U(x^j, x^i)$. (A1') is identical to (A1) in the proof of Theorem 4.1 and fulfilled by Assumption 4.1.

(A2') The parts (i) and (ii) are identical to (A2) (i) and (ii). Consider (iii):

$$\int \big(\widehat{V}(x^j)^{-1}\widehat{U}(x^j, x^i) - V(x^j)^{-1}U(x^j, x^i)\big)^2 \frac{f(x^j)}{f(x^i)}\,\mathrm{d}x^i\,\mathrm{d}x^j = o_P(1)$$

Adding and subtracting $V(x^j)^{-1}\widehat{U}(x^j, x^i)$ this holds by the triangular inequality and Lemma 4.1.

(A3') The statement is immediately implied by (A5').

(A4') With probability tending to one (for $n \to \infty$)

$$\sup_{x^j \in \mathcal{G}^j} \int \operatorname{tr}\big(\widehat{U}(x^i, x^j)\widehat{V}(x^j)^{-2}\widehat{U}(x^i, x^j)\big) f(x^i)^{-1} \, \mathrm{d}x^i \le C,$$

for all $j = 1, \dots, d$, because all elements of the matrices are consistent estimators of densities, for which the statement holds.

(A5') is shown by applying Lemmata 4.2 and 4.3 and the boundedness of the function $\mu^1(x^j)$ that guarantees the quadratic integrability. In total we have that

$$\int (\widehat{\mu}_h^{LL,B}(x^j))^2 f(x^j) \, \mathrm{d}x^j \quad \text{and} \quad \int (\widehat{\mu}_h^{LL,V}(x^j))^2 f(x^j) \, \mathrm{d}x^j$$

are bounded with probability tending to. The same holds for

$$\int (\widehat{\mu}_{j,h}^{LL,B}(x^j))^2 f(x^j) \, \mathrm{d}x^j \quad \text{and} \quad \int (\widehat{\mu}_{j,h}^{LL,V}(x^j))^2 f(x^j) \, \mathrm{d}x^j.$$

(A6') Denote the L_2-norm in $\mathbb{R}^2$ with $\| \cdot \|_2$. Then, it is direct to show that

$$\begin{aligned}(4.34) \qquad & \sup_{x^i \in \mathcal{G}^i} \Big\| \int \widehat{V}(x^i)^{-1}\widehat{U}(x^i, x^j)\widehat{V}(x^j)^{-1} \begin{pmatrix} \widehat{t}_{h,0}^V(x^j) \\ \widehat{t}_{h,1}^V(x^j) \end{pmatrix} \mathrm{d}x^j \Big\|_2 \\ & = \sup_{x^i \in \mathcal{G}^i} \Big\| \int V(x^i)^{-1}U(x^i, x^j)V(x^j)^{-1} \begin{pmatrix} \widehat{t}_{h,0}^V(x^j) \\ \widehat{t}_{h,1}^V(x^j) \end{pmatrix} \mathrm{d}x^j \Big\|_2 + o_P(h^2),\end{aligned}$$

because $\sup_{x^j} |\widehat{t}_{h,l}^V(x^j)| = O_P(h^{-1/2}T^{-1/2} \log T)$ and all density estimators converge uniformly as given by Lemma 4.1. Next note that

$$\int \kappa(x^i, x^j) \frac{f(x^i, x^j)}{f(x^i)f(x^j)} \widehat{t}_{h,l}(x^j) \, \mathrm{d}x^j = \sum_{k=0}^{nT-1} \xi_{k,l}(x^i),$$

where $\kappa(x^i, x^j)$ is a kernel constant independent of h and

$$\begin{aligned}\xi_{k,l}(x^j) = \int \frac{f(X_{k\Delta}^i + hu, x^j)}{f(X_{k\Delta}^i + hu)f(x^j)} h\widetilde{K}_{h,l}(X_{k\Delta}^i + hu, X_{k\Delta}^i) \, \mathrm{d}u \\ \times \sum_{l=1}^{d} \frac{1}{\Delta} \int_{k\Delta}^{(k+1)\Delta} \sigma^{1l}(X_s) \, \mathrm{d}W_s^l.\end{aligned}$$

From a simple application of the Itô-isometry together with the boundedness of kernel and density it is obtained that $\sup_{x^j} |\sum_{k=0}^{nT-1} \xi_{k,l}(x^j)| = O_P(T^{-1/2})$ and then the quantity in (4.34) is of order $o_P(h^2)$. Using this result it follows directly that

$$\Big\| \int \widehat{V}(x^i)^{-1} \widehat{U}(x^i, x^j) \widehat{V}(x^j)^{-1} \begin{pmatrix} \widehat{t}^V_{h,0}(x^j) \\ \widehat{t}^V_{h,1}(x^j) \end{pmatrix} \, \mathrm{d}x^j \Big\|_{V(x^i)} = o_P(h^2).$$

(A8') The convergence results of Lemma 4.1 immediately imply that

$$\sup_{x^j \in \mathcal{G}^j} \int \big| \widehat{V}(x^j)^{-1} \widehat{U}(x^i, x^j) - V(x^j)^{-1} U(x^i, x^j) \frac{f(x^i, x^j)}{f(x^j)} \big| f(x^i) \, \mathrm{d}x^i = o_P(1),$$

where the supremum has to be taken elementwise in the matrix.

(A9') Define

$$\alpha_{n,T,j}(x^j) = \mu^1(x^j) + h^2 \kappa_2(x^j) \partial_j^2 \mu^1(x^j)/2 \quad \text{and} \quad \alpha^j_{n,T}(x^j) = h \partial_j \mu^1(x^j),$$

for $j = 0, \ldots, d$. Clearly $\int \alpha_{n,T,j}(x^j) f(x^j) \, \mathrm{d}x^j < \infty$ and $\int \alpha^j_{n,T}(x^j) f(x^j) \, \mathrm{d}x^j < \infty$. First, consider the constant. This is given by

$$\int \alpha_{n,T,j}(x^j) \widehat{f}_{h,0}(x^j) \, \mathrm{d}x^j + \int \alpha^j_{n,T}(x^j) \widehat{f}_{h,1}(x^j) \, \mathrm{d}x^j = b_h^{1,j} + o_p(h^2).$$

The limit of the first integral is as in (4.33) and the second integral is of $O(h^3)$, because the kernel constant $\kappa_1(x^j)$ is only at the boundary different from zero. The expression differs from that given in Mammen, Linton and Nielsen (1999), because a different normalization is used in the local linear case. This does not affect the statement of their Theorem 3'.
Define random variables

$$\begin{pmatrix} \widehat{\nu}_{n,T,j}(x^j) \\ \widehat{\nu}^j_{n,T}(x^j) \end{pmatrix} = \begin{pmatrix} \alpha_{n,T,0} + \alpha_{n,T,j}(x^j) \\ \alpha^j_{n,T}(x^j) \end{pmatrix} + \sum_{i \neq j} \int \widehat{V}(x^j)^{-1} \widehat{U}(x^i, x^j) \begin{pmatrix} \alpha_{n,T,i}(x^i) \\ \alpha^i_{n,T}(x^i) \end{pmatrix} \mathrm{d}x^i.$$

Then is has to be shown that

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_h^{LL,B}(x^j) - \widehat{\nu}_{n,T,0} - \widehat{\nu}_{n,T,j}(x^j)| = o_P(h^2), \tag{4.35}$$

$$\sup_{x^j \in \mathcal{G}^j} |\widehat{\mu}_{j,h}^{LL,B}(x^j) - \widehat{\nu}^0_{n,T} - \widehat{\nu}^j_{n,T}(x^j)| = o_P(h^2). \tag{4.36}$$

This is done by decomposing the numerator parts of the marginal estimator. Analogous to (4.30) it holds that

$$\widehat{t}^B_{h,l}(x^j) = \sum_{i=0}^{d} \frac{1}{Tn} \sum_{k=0}^{nT-1} \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})\mu^1(X^i_{k\Delta}) + O_P(n^{-1/2}).$$

Applying a Taylor expansion for $\mu^1(X^j_{k\Delta}) = \mu^1(x^j) + \partial_j\mu^1(x^j)(X^j_{k\Delta} - x^j) + \partial_j^2\mu^1(x^j)(X^j_{k\Delta} - x^j)^2/2 + R_k(x^j)$. Then, the cases $i = j$ and $i \neq j$ have to be treated separately. For the first it holds that

$$\begin{aligned}\frac{1}{Tn}\sum_{k=0}^{nT-1} \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})\mu^1(X^j_{k\Delta}) &= \mu^1(x^j)\widehat{f}_{h,0+l}(x^j) + h\partial_j\mu^1(x^j)\widehat{f}_{h,1+l}(x^j)\\ &\quad + h^2\partial_j^2\mu^1(x^j)\widehat{f}_{h,2+l}(x^j) + R_{n,T,j}(x^j).\end{aligned}$$

If $i \neq j$ we have to use that $1 = \int K_h(x^i, X_{k\Delta})\,\mathrm{d}x^i$ for all k to obtain

$$\begin{aligned}\frac{1}{Tn}\sum_{k=0}^{nT-1} &\widetilde{K}_{h,l}(x^j, X^j_{k\Delta})\mu^1(X^i_{k\Delta})\\ &= \frac{1}{Tn}\sum_{k=0}^{nT-1} \int \widetilde{K}_{h,l}(x^j, X^j_{k\Delta})K_h(x^i, X_{k\Delta})\mu^1(X^i_{k\Delta})\,\mathrm{d}x^i\\ &= \int \widehat{f}_{h,(0,l)}(x^i, x^j)\mu^1(x^i)\,\mathrm{d}x^i + h\int \widehat{f}_{h,(1,l)}(x^i, x^j)\partial_i\mu^1(x^i)\,\mathrm{d}x^i\\ &\quad + \frac{h^2}{2}\int \widehat{f}_{h,(2,l)}(x^i, x^j)\partial_j^2\mu^1(x^i)\,\mathrm{d}x^i + R_{n,T,i}(x^j).\end{aligned}$$

Now, plug these results into the representation of the local linear estimator

$$\begin{aligned}&\begin{pmatrix}\widehat{\mu}^{LL,B}(x^j)\\ \widehat{\partial_j\mu}^{LL,B}(x^j)\end{pmatrix} = \widehat{V}(x^j)^{-1}\begin{pmatrix}\widehat{t}^B_{h,0}(x^j)\\ \widehat{t}^B_{h,1}(x^j)\end{pmatrix}\\ &= \begin{pmatrix}\mu^1(x^j) + \frac{h^2}{2}\partial_j^2\mu^1(x^j)\frac{(\widehat{f}_{h,2}(x^j))^2 - \widehat{f}_{h,1}(x^j)\widehat{f}_{h,3}(x^j)}{\widehat{f}_{h,0}(x^j)\widehat{f}_{h,2}(x^j) - (\widehat{f}_{h,1}(x^j))^2}\\ h\partial_j\mu^1(x^j) + \frac{h^2}{2}\partial_j^2\mu^1(x^j)\frac{\widehat{f}_{h,0}(x^j)\widehat{f}_{h,3}(x^j) - \widehat{f}_{h,1}\widehat{f}_{h,2}(x^j)}{\widehat{f}_{h,0}(x^j)\widehat{f}_{h,2}(x^j) - (\widehat{f}_{h,1}(x^j))^2}\end{pmatrix}\\ &\quad + \sum_{i\neq j}\widehat{V}(x^j)^{-1}\int \widehat{U}(x^i, x^j)\begin{pmatrix}\mu^1(x^i) + h^2\kappa_2(x^i)\partial^2\mu^1(x^i)/2\\ h\partial_j\mu^1(x^i)\end{pmatrix}\mathrm{d}x^i\\ &\quad - \sum_{i\neq j}\widehat{V}(x^j)^{-1}\int \begin{pmatrix}h^2\partial_j^2\mu^1(x^i)\big(\widehat{f}_{h,(0,0)}(x^i)\kappa_2(x^i) - \widehat{f}_{h,(2,0)}(x^i)\big)/2\\ h^2\partial_j^2\mu^1(x^i)\big(\widehat{f}_{h,(0,1)}(x^i)\kappa_2(x^i) - \widehat{f}_{h,(2,1)}(x^i)\big)/2\end{pmatrix}\mathrm{d}x^i\\ &\quad + \sum_{i=1}^{d} R_{n,T,i}(x^j).\end{aligned}$$

By the uniform convergence rates of the density estimators the asymptotic expression of the bias is obtained. Because $\int \kappa_0(x^i) - 1 \, \mathrm{d}x^i = O(h)$, the third term is of $o_P(h^2)$ uniformly over x^j. Showing that $\sup_{x^j} R_{n,T,i}(x^j) = o_P(h^2)$ as in the proof of Theorem 4.1 the desired rate is obtained. The uniform convergence of (4.35) and (4.36) imply convergence in the squared $L_2(f(x^j))$-norm. This completes the proof of the assumption. □

Proof of Theorem 4.4

The Assumptions (A1'), (A2'), (A4') and (A8') are identical to the proof of theorem 4.2. (A5') follows by lemmata 4.5 and 4.6. (A6') and (A9') can be concluded by identical modifications as in the proof of Theorem 4.3. □

Chapter 5

Nonparametric Regression Tests Using Dimension Reduction Techniques

5.1 Introduction

Testing for parametric structure is an important issue in nonparametric regression analysis. A standard approach is to measure the distance between a parametric and a nonparametric fit with a squared deviation measure. Based on the L_2-distance, various test statistics have been proposed, e. g. Härdle and Mammen (1993), Li (1994), Hjellvik and Tjøstheim (1995), Zheng (1996), Li and Wang (1998), Dette (1999) or Fan et al. (2001). Due to the use of a kernel estimator these tests suffer from the curse of dimensionality, i. e. the estimators become worse as the dimension of the predictor increases. Formally, this results in a slower rate of convergence for local alternatives. Beside this asymptotic result, the procedures break down in small samples and have no power there.

A powerful technique to overcome the curse of dimensionality is to impose an additive structure on the unknown regression function. Because additive models maintain high flexibility compared to parametric specifications they are now widely used in nonparametric multivariate modelling. Furthermore, the additive components can be estimated with the same rate as a one-dimensional nonparametric regression and in that way the curse of dimensionality can be circumvented. To estimate additive models, different procedures have been proposed. The most prominent smoothing based techniques are the classical backfitting algorithm by Buja, Hastie and Tibshirani (1989), the marginal integration by Linton and Nielsen (1995) and Tjøstheim and Auestad (1994), smooth backfitting by Mammen, Linton and Nielsen (1999) and local partitioned regression by Christopeit and Hoderlein (2006). While marginal integration and local partitioned regression suffer from the curse of dimensionality, because they use a full-dimensional estimator in a first stage, backfitting procedures are completely free of that. The backfitting estimators require iterative procedures that make the asymptotic analysis more complex.

Compared to classical backfitting, smooth backfitting has different advantages. Opsomer and Rupert (1997) and Opsomer (2000) analyze the asymptotic properties of classical backfitting and show that the estimator is not fully oracle efficient. This concept requires that each additive component can be estimated as well as if all other components were known. Mammen, Linton and Nielsen (1999) have shown that smooth backfitting is fully oracle efficient. If the design is correlated, the implementation of classical backfitting estimators is problematic. Opsomer and Rupert (1997) illustrate the theoretical restriction of the correlation for covariates that are bivariate normally distributed. The declined performance of classical backfitting is also reported in simulation studies (see Sperlich, Linton and Härdle, 1999), while smooth backfitting performs much better (see Nielsen and Sperlich, 2005). Finally, the behavior of the smooth backfitting estimators is well understood, even if the true model is not additive. This is of particular importance for the analysis of a test statistic under the alternative hypothesis.

This article proposes to construct a test for parametric form by projecting the residuals under the null hypothesis onto the space of additive functions. The asymptotic results show that the test has the same rate of convergence for an arbitrary dimension of the covariates, which coincides with the rate for one-dimensional testing problems. Therefore this test circumvents the curse of dimensionality. The price for this is the incapacity to detect arbitrary alternatives. If the additive projection of model-implied residuals is zero, the test can not reject the null hypothesis. Since alternative test procedures fail to work at all in small sample size, this test still provides a powerful data analytic tool.

A similar test problem is considered in Fan and Jiang (2005). They use a generalized likelihood ratio test statistic to compare the parametric fit to the additive fit, where the additive model is estimated by classical backfitting. Because the fits are compared directly, this test can only be used to test for parametric specifications that are a subclass of the class of additive functions. The test proposed in this article uses the fact that the smooth backfitting estimator can be understood as an estimator of the additive projection. Therefore it is applicable to a larger class of hypotheses.

This chapter is organized as follows. In the second section the test statistic is motivated and the asymptotic results are obtained. For a small sample size, the asymptotic distribution does not approximate the null distribution very well. Therefore a wild bootstrap procedure is proposed and analyzed. Some extensions to the basic test statistic are discussed in Section 3. The finite sample performance is examined by a Monte-Carlo study and illustrated by a small data example. This is presented in Section 4. All proofs are deferred to the appendix.

5.2 The Test Statistic

5.2.1 Motivation of the Test Statistic

Let $Y \in \mathbb{R}$ and $X = (X^1, \dots, X^d)' \in \mathbb{R}^d$ for some $d \geq 1$ denote random variables and define the mean regression function

$$g(x) = \mathbf{E}(Y \mid X = x).$$

To specify a certain model it has to be judged whether this function falls into a parametric function class. So the null hypothesis to be tested is if

$$H_0\colon \mathbb{P}(g(X) = G(X, \theta)) = 1 \quad \text{for some } \theta \in \Theta, \tag{5.1}$$

where $\Theta \subseteq \mathbb{R}^{\widetilde{d}}$ is a finite dimensional parameter space and $G(x, \theta)$ is a known function. The common approach to test this hypothesis using kernel regression techniques is to use the equivalence of H_0 to

$$\mathbf{E}(g(X) - G(X, \theta))^2 w(X) = 0, \tag{5.2}$$

where $w(x)$ is some positive weighting function. To construct a test statistic, equation (5.2) is replaced by the sample counterpart, using a parametric estimator for θ and a kernel estimator of $g(x)$ (e. g. the Nadaraya-Watson estimator).[1] However, for high-dimensional regressors X the rate of convergence of the estimator of $g(x)$ becomes slower. Therefore the resulting test suffers from the curse

[1] Most authors do not use equation (5.2) directly, but transform it. For example, Härdle and Mammen (1993) use $\mathbf{E}(m(X) - \mathbf{E}(G(X, \theta) \mid X))^2 w(X)$. For other specifications see Li (1994), Zheng (1996) or Dette (1999).

of dimensionality as well. This is reflected in a rate of convergence of $nh^{d/2}$ for the test statistic, the incapacity to detect local alternatives that converge to the null hypothesis faster than $n^{-1/2}h^{-d/4}$ and the need for larger sample sizes.

A common approach to circumvent the curse of dimensionality in nonparametric regression is to impose an additive structure on the mean regression function, i. e.

$$g(x) = g^0 + g^1(x^1) + \cdots + g^d(x^d)$$

and for identifiability it is assumed that $\mathbf{E}\, g^j(X^j) = 0$ for all $j = 1, \dots, d$. Additive models provide an important class of structured multivariate nonparametric models, because they are more flexible than parametric families. Denote with $\mathcal{G} = \{g \colon \mathbb{R}^d \to \mathbb{R} \mid \mathbf{E}\, g(X)^2 < \infty\}$ the class of L_2-functions, with $\mathcal{G}^{ad} = \{g \in \mathcal{G} \mid g(x) = g^0 + g^1(x^1) + \cdots + g^d(x^d)\}$ its additive subclass and with $\mathcal{G}_{G,\Theta} = \{g \in \mathcal{G} \mid g(x) = G(x, \theta) \text{ for some } \theta \in \Theta\}$ a parametric subclass. Denote with $\mathcal{P}(Y \mid X = x)$ the L_2-projection of Y onto the space of additive functions $\mathcal{G}^{ad}$, defined as

$$\mathcal{P}(Y \mid X) = \operatorname*{argmin}_{\gamma \in \mathcal{G}^{ad}} \mathbf{E}(Y - \gamma(X))^2. \tag{5.3}$$

where the minimization is under the constraint $\mathbf{E}\,\gamma^j(x^j) = 0$.

To motivate the test statistic of this chapter consider for the beginning the testing problem

$$H_0 \colon g(x) = G(x, \theta) \in \mathcal{G}_{G,\Theta} \qquad \text{vs.} \qquad H_1 \colon g(x) \in \mathcal{G}^{ad} \backslash \mathcal{G}_{G,\Theta}, \tag{5.4}$$

with $\mathcal{G}_{G,\Theta} \subset \mathcal{G}^{ad}$. For example, $\mathcal{G}_{G,\Theta}$ is the class of linear functions. In this case, the conditional expectation and the additive projection are identical, i. e. $\mathcal{P}(Y \mid X = x) = g(x)$ and the null hypothesis is equivalent to

$$\mathbf{E}(\mathcal{P}(Y \mid X) - G(X, \theta))^2 w(X) = 0. \tag{5.5}$$

The additive projection is a sum over d one-dimensional functions. Using kernel regression techniques these components (and therefore the whole function) can be estimated with the one-dimensional rate of convergence. Having such estimators at hand a sample analogue to the left-hand side in (5.5) can be constructed. This can be used as a test statistic that converges faster than the test statistics based on (5.2). By this, the curse of dimensionality can be circumvented.

It is not advisable to compare the distance between the additive regression and the parametric estimator directly. While under H_0 they both converge to the same true function asymptotically, the nonparametric estimator has a bias that can dominate the test statistic. Therefore it is preferable to smooth the parametric estimator to imitate the bias of the kernel estimator. This is equivalent to basing the test on

$$\mathbf{E}(\mathcal{P}(Y - G(X,\theta) \mid X))^2 w(X) = 0. \tag{5.6}$$

But this equation holds under H_0 also for parametric families that are not additive, i. e. $\mathcal{G}_{G,\Theta} \not\subset \mathcal{G}^{ad}$. The object of interest is now the additive projection of residuals of the parametric estimation. Therefore a test based on (5.6) is applicable to more general hypotheses than (5.4).

The testing problem (5.4) has also been considered by Fan and Jiang (2005). They adapt the generalized likelihood ratio-test by Fan, Zhang and Zhang (2001) to this testing problem. The test statistic is obtained by constructing residual sums of squares under H_0 – fitting the model with a parametric estimator – and under H_1 – estimating the additive model by classical backfitting. Then, the logarithm of the ratio of these sums of squares serves as test statistic. Shortcomings of the classical backfitting estimator like restricted correlation structure, lack of oracle efficiency and unknown behavior under non-additive models have already been mentioned. Therefore the choice of this article is the smooth backfitting estimator by Mammen, Linton and Nielsen (1999) as estimator of the additive projection. Before the test statistic will be constructed, the next subsection reviews smooth backfitting estimation.

5.2.2 Smooth Backfitting

Based on a sample of independent and identically distributed random variables $(X_i, Y_i), i = 1, \dots, n$ it is desired to estimate

$$\mathcal{P}(Y - G(X,\theta) \mid X = x) = m^0 + m^1(x^1) + \cdots + m^d(x^d),$$

under the constraint

$$\int_{\mathcal{A}^j} m^j(x^j) f^j(x^j)\,\mathrm{d}x^j = 0, \qquad j = 1, \dots, d. \tag{5.7}$$

Here $f^j(x^j)$ denotes the marginal density of X^j and $\mathcal{A}^j$ is a compact subset of the support of X^j. The smooth backfitting procedures are based on usual kernel estimators. In this subsection the algorithm based on a Nadaraya-Watson estimator is presented. Alternatively, smooth backfitting could be based on local linear estimators. For more detailed expositions of the estimators and algorithms see Mammen, Linton and Nielsen (1999) or Nielsen and Sperlich (2005).

Having a parametric estimator $\widehat{\theta}$ at hand, the residuals of the parametric regression $\widehat{U}_i = Y_i - G(X_i, \widehat{\theta})$ can be constructed. The Nadaraya-Watson smooth backfitting estimators are motivated by the solution of the smoothed empirical version of the additive projection (5.3)

$$\min_{\substack{\bar{\mu}^0,\ldots,\bar{\mu}^1 \\ \bar{\mu}^0+\cdots+\bar{\mu}^1 \in \mathcal{G}^{ad}}} \int_{\mathcal{A}} \frac{1}{n} \sum_{i=0}^{n} \big(\widehat{U}_i - \bar{\mu}^0 - \bar{\mu}^1(x^1) - \cdots - \bar{\mu}^d(x^d)\big)^2 \prod_{j=1}^{d} K_h(x^j, X_i)\,\mathrm{d}x, \tag{5.8}$$

where the minimization is subject to the empirical version of (5.7), given by

$$\int_{\mathcal{A}^j} \bar{\mu}^j(x^j)\widehat{f}_h^j(x^j)\,\mathrm{d}x^j = 0, \quad j = 1, \ldots, d. \tag{5.9}$$

Here, $K_h(u^j, v^j)$ is a kernel weight, $\widehat{f}_h^j(x^j) = n^{-1}\sum_{i=1}^n K_h(x^j, X_i^j)$ is a kernel density estimator[2] and $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^d$. Usually, for smooth backfitting estimators, modified kernel weights are implemented. These are given by

$$K_h(u^j, v^j) = \frac{K(h^{-1}(v^j - u^j))}{\int_{\mathcal{A}^j} K(h^{-1}(v^j - w^j)\,\mathrm{d}w^j}$$

where $K(\cdot)$ integrates to one over its support. This ensures that $\int_{\mathcal{A}^j} K_h(u^j, v^j)\,\mathrm{d}u^j = 1$ for all v^j and is required to derive the asymptotic properties of the estimators. Simulation results suggest that unmodified kernels can be implemented as well, but by now no theoretical justification for doing so exists.

Solving the minimization problem (5.8) with respect to (5.9) the minimum $(\widetilde{m}_h^0, \widetilde{m}_h^1, \ldots, \widetilde{m}_h^d)$ is given as the implicit solution to the set of equations

$$\widetilde{m}_h^j(x^j) = \widehat{m}_h^j(x^j) - \sum_{k \neq j} \int_{\mathcal{A}^k} \widetilde{m}_h^k(x^k) \frac{\widehat{f}_h^{k,j}(x^k, x^j)}{\widehat{f}_h^j(x^j)}\,\mathrm{d}x^k - \widetilde{m}_j^0, \tag{5.10}$$

together with $\int \widetilde{m}_h^j(x^j)\widehat{f}_h^j(x^j)\,\mathrm{d}x^j = 0$ for $j = 1, \ldots, d$. Here, the two-dimensional kernel density estimator of the joint density of X^k and X^j is denoted with $\widehat{f}_h^{k,j,}(x^k, x^j) = n^{-1}\sum_{i=1}^n K_h(x^k, X_i^k)K_h(x^j, X_i^j)$ is and the marginal Nadaraya-Watson estimator is given by $\widehat{m}_h^j(x^j) = \widehat{f}_h^j(x^j)^{-1}n^{-1}\sum_{i=1}^n K_h(x^j, X_i^j)\widehat{U}_i$. The set

[2] To reduce notation it is assumed that the same bandwidth is used for all dimensions.

of equations (5.10) is then solved by an iterative procedure where the marginal estimators can be used as starting values. If modified kernels are used, the constant is given by $\widetilde{m}_j^0 = n^{-1}\sum_{i=1}^n \widehat{U}_i$ and demeaned data can be used.

The smooth backfitting estimators $(\widetilde{m}_h^0, \widetilde{m}_h^1, \dots, \widetilde{m}_h^d)$ are defined as the solution to (5.10). Mammen, Linton and Nielsen (1999) give general conditions under which the algorithm converges and investigate asymptotic properties of the solutions. If the additive model holds, then the estimators enjoy an oracle property for the variance. This means that the backfitting estimator of one additive component converges with rate $\sqrt{nh}$ and has the same variance as the infeasible oracle estimator which is based on knowledge of all other components of the additive function. Smooth backfitting based on local linear estimators is fully oracle efficient, which means that these estimators have the same asymptotic variance and bias as the oracle estimator. But under H_0 there is no bias at all. Therefore it is no disadvantage to base the test statistic on Nadaraya-Watson smooth backfitting.

In contrast to classical backfitting, the behavior of the smooth backfitting estimators can be investigated even if the additive model does not hold. This alleviates the analysis of the test statistic under the alternative hypothesis.

5.2.3 Asymptotic Results

Estimating the additive projection by smooth backfitting, the test statistic can be constructed. Based on equation (5.6), the null hypothesis will now be formulated more generally as

$$(5.11) \qquad H_0\colon \mathbb{P}(\mathcal{P}(Y - G(X,\theta) \mid X) = 0) = 1 \quad \text{for some } \theta \in \Theta,$$

where $\Theta \subseteq \mathbb{R}^{d'}$ is a finite dimensional parameter space and $G(X,\theta)$ is a known function.

Using the smooth backfitting estimators $\widetilde{m}_h^1(x^1), \dots, \widetilde{m}_h^d(x^d)$ the empirical version of equation (5.6) can be constructed. For the parametric estimator $\widehat{\theta}$ of θ some assumptions beyond consistency will be specified below. The test statistic is defined as

$$(5.12) \qquad \widehat{T} = \int_{\mathcal{A}} \Big(\sum_{j=0}^{d} \widetilde{m}_h^j(x^j)\Big)^2 \widehat{f}_h(x) w(x)\,\mathrm{d}x,$$

where $\widehat{f}_h(x) = n^{-1}\sum_{i=1}^n K_h(x, X_i)$ is a full dimensional kernel density estimator. Note that by solving the square inside the integral of $\widehat{T}$ and choosing the weight function such that[3] $\int \widehat{f}_h(x)w(x)\,\mathrm{d}x^{-j} = \int \widehat{f}_h^{j,k}(x^j, x^k)w(x^j, x^k)\,\mathrm{d}x^k = \widehat{f}_h^j(x^j)w(x^j)$ the test statistic can be written as sum over one- and two-dimensional integrals only. Therefore only one- and two-dimensional kernel density estimators are required.

As an alternative the expectation in equation (5.6) could be replaced by a sum over the data points. This would result in the test statistic

$$\widetilde{T} = \sum_{i=1}^{n} \Big(\sum_{j=0}^{d} \widetilde{m}_h^j(X_i^j) \Big)^2 w(X_i),$$

which shares the asymptotic properties of $\widehat{T}$ but is computationally more cumbersome, because the backfitting estimators have to be evaluated at the data points. From the algorithm the estimators are obtained at (ideally) equispaced grid points, that allow to evaluate integrals quickly. A third implementation would be given by omitting the kernel density estimator in the integration. This corresponds to the choice of $w(x) = \widetilde{w}(x)/f(x)$ asymptotically.

The hypothesis (5.11) is more general than (5.4), but it is weaker than (5.1). Of course it could be the case that the conditional expectation is not in the parametric class, but the additive projection $\mathcal{P}(Y - G(X,\theta) \mid X) = \mathcal{P}(m(X) - G(X,\theta) \mid X)$ is still zero. As an example consider $Y = \theta_1 X^1 + \theta_2 X^2 + X^1 X^2 + \varepsilon$ as true data generating process where X^1, X^2 and ε are independent (truncated) normal random variables (with expectation zero). As parametric class choose $G(x,\theta) = \theta_1 x^1 + \theta_2 x^2$. Then obviously the conditional expectation of Y given X is not in that class. However, if X^1 and X^2 are independent, it holds that $\mathcal{P}(Y \mid X = x) = G(x,\theta)$. However, if X^1 and X^2 are correlated, the null hypothesis is violated and the test will reject.

For the more general testing problem (5.1) this approach can still be useful as a data analytic tool. If a full-dimensional test for (5.1) is not available because the sample size is too small to estimate a full-dimensional nonparametric regression, the class of alternatives is still larger than for usual parametric goodness-of-fit tests.

To derive asymptotic results, the following assumptions have to be imposed

Assumption 5.1. *For the nonparametric estimation*

1. *The data $(Y_i, X_i), i = 1, \ldots, n$ are independent and identically distributed with density $f(y,x)$.*

[3] Denote $\mathrm{d}x^{-j} = \mathrm{d}x^1 \ldots \mathrm{d}x^{j-1}\,\mathrm{d}x^{j+1} \ldots \mathrm{d}x^d$.

2. *$f(x) = \int f(y,x)\,\mathrm{d}y$ is twice continuously differentiable on $\mathcal{A} \subset \mathbb{R}^d$ with bounded derivatives.*

3. *The two-dimensional densities $f(x^j, x^k) = \int f(x)\,dx^{-(j,k)}$ are twice continuously differentiable on $\mathcal{A}^j \times \mathcal{A}^k$ with bounded derivatives.*

4. *The marginal densities $f(x^j) = \int f(x)\,dx^{-j}$ are twice continuously differentiable on $\mathcal{A}^j$ with bounded derivatives. $f(x^j)$ is bounded from below on $\mathcal{A}^j$.*

5. *The continuously differentiable weighting function $w(x)$ is positive and bounded on $\mathcal{A}$.*

6. *The conditional variances*

$$v^j(x^j) = \mathbf{E}\big((Y - G(X,\theta))^2 \mid X^j = x^j\big)$$

are square-integrable on $\mathcal{A}^j$.

7. $\mathbf{E}((Y - G(X,\theta))^4) < \infty$.

8. *The kernel function $K\colon [-1,1] \to \mathbb{R}$ is Lipschitz continuous, bounded and symmetric around 0. The kernel and its convolution are square-integrable*

$$\|K\|_2^2 = \int_{-1}^{1} K(u)^2\,\mathrm{d}u \qquad \|K * K\|_2^2 = \int_{-1}^{1}\Big(\int_{-1}^{1} K(u)K(u+v)\,\mathrm{d}u\Big)^2\,\mathrm{d}v.$$

9. *The bandwidth sequence satisfies $h = O(n^{-1/5})$.*

The assumption of independent and identically distributed data could be relaxed to allow for β-mixing random variables with mixing coefficients decaying sufficiently fast. The moment conditions are minimal to obtain asymptotic normality of the test statistic and the required smoothness of the unknown functions is standard in nonparametric regression.

If the implementation of a data-driven bandwidth is desired, there are two proposals. Nielsen and Sperlich (2005) investigate the implementation of cross-validation in a simulation study, but give no theoretical result. Mammen and Park (2006) use penalizing functions and prove the validity of their method. For an iterative estimator cross-validation is supposed to be more time-consuming and therefore the penalizing function approach is favorable. In practice the bandwidth for the test could be chosen by using the optimal bandwidth of the additive projection of Y on X. However, this proposal will not yield an optimal bandwidth for the test. For optimal testing, a data adaptive method as in Horowitz and Spokoiny (2001) could be implemented, but this is beyond the scope of the present work.

Assumption 5.2. *For the parametric estimation*

1. *Under the null hypothesis it holds that $\widehat{\theta} - \theta = O_P(n^{-1/2})$ for all $\theta \in \Theta$.*

2. *Under the alternative hypothesis there exists a $\widetilde{\theta} \in \Theta$ such that $\widehat{\theta} - \widetilde{\theta} = O_P(n^{-1/2})$.*

3. *For the link function it holds that $\nabla_\theta G(x, \theta)$ and $\nabla^2_\theta G(x, \theta)$ are continuous in x and θ. $\nabla_\theta G(x, \cdot)$ and $\nabla^2_\theta G(x, \cdot)$ are dominated by square integrable functions on $\mathcal{A}$.*

For usual parametric estimators this assumption is no restriction. It is formulated in a rather general way to cover many possible cases of different null hypotheses. Considering again a linear model, $\widehat{\theta}$ would be the usual (general) least squares estimator, for which Assumption 5.2 is clearly fulfilled. The first theorem states the asymptotic behavior of $\widehat{T}$ under H_0

Theorem 5.1. *Let Assumptions 5.1 and 5.2 be fulfilled. Then it holds under H_0 that*

$$n\sqrt{h}\widehat{T} - h^{-1/2} B_T \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T),$$

where

$$\Sigma_T = 2\|K * K\|_2^2 \sum_{j=1}^{d} \int v^j(x^j)^2 w(x^j)^2 \, \mathrm{d}x^j$$

$$B_T = \|K\|_2^2 \sum_{j=1}^{d} \int v^j(x^j) w(x^j) \, \mathrm{d}x^j.$$

As expected, the test statistic is asymptotically normal with a one-dimensional rate of convergence of $n\sqrt{h}$. The variance Σ_T is given as the sum over integrals of the d marginal conditional variances $v^j(x^j)$. Thus, the variance (and the bias B_T as well) increases with the dimension. As usual the power decreases with an increasing variance of the test statistic under H_0. In finite samples this effect is present and will be found in the simulation.

To check the consistency of $\widehat{T}$, the behavior under the alternative has to be examined. Formally, the alternative hypothesis is stated as

$$H_1\colon \mathbb{P}(\mathcal{P}(Y - G(X,\theta) \mid X) = 0) < 1 \quad \text{for any } \theta \in \Theta.$$

This treats the case of fixed alternatives, i. e. $\mathbf{E}(\mathcal{P}(Y - G(X,\bar{\theta}) \mid X))^2 = c > 0$, where $\bar{\theta}$ is given in Assumption 5.2 and c is a fixed positive constant.

Theorem 5.2. *Let Assumptions 5.1 and 5.2 be fulfilled. Then, under H_1 it holds that*

$$\widehat{T} \xrightarrow{P} \mathbf{E}(\mathcal{P}(Y - G(X,\widetilde{\theta}) \mid X))^2 w(X).$$

Hence, the standardized test statistic of $\widehat{T}$ diverges to infinity in probability. Therefore the test is consistent against any fixed alternative, where the additive projection of the model-implied residuals is nonzero. Return to the restricted testing problem (5.4) where the parametric class is a subclass of the additive model. Then, $\widehat{T}$ is consistent against all fixed alternatives in $\mathcal{G}^{ad}$.

Of additional interest is the behavior of the test against local alternatives, i. e. alternatives that converge to H_0 for $n \to \infty$. Consider the sequence of local alternatives

$$H_{1n}\colon\ \mathcal{P}(Y - G(X,\widetilde{\theta}) \mid X) = g_n(X),$$

where $g_n(x) \in \mathcal{G}^{add}$ is a nonzero function.

Theorem 5.3. *Let Assumptions 5.1 and 5.2 hold. If there exists a constant B_L such that*

$$\frac{\lambda_n}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \frac{g_n(X_i)^2}{f^j(X_i^j)} f(X_i) w(X_i) \xrightarrow{P} B_L$$

with $\lambda_n = O(nh^{1/2})$ then it holds under H_{1n} that

$$n\sqrt{h}\widehat{T} - h^{-1/2}(B_T + \|K\|_2^2 B_L) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T)$$

with B_T and Σ_T as in Theorem 5.1.

Usually a kernel-based test for parametric structure can only detect local alternatives that converge to zero at rate $n^{-1/2}h^{-d/4}$ where d is the dimension of the covariates. The implementation of the additive estimator circumvents this curse of dimensionality. However, the price that has to be paid for the circumvention is reflected in the smaller class of alternatives against which the test has power. While a kernel based test using a full-dimensional estimator has power against functions in $\mathcal{G}$, the test proposed here has only power against alternatives with $\mathcal{P}(Y - G(X, \theta)) \neq 0$. However, for the considered situation, where d is large and n is relatively small, the asymptotic results of the full-dimensional test are not available.

Taking again the linear model as an example, the parametric structure could be checked by testing against a higher dimensional model (or against a quadratic term). Such tests have power against contiguous alternatives, i. e. alternatives that converge to the null hypothesis at rate $n^{-1/2}$. But the class of alternatives that can be detected is further restricted by the construction of these tests. For example it may only be the class of quadratic deviations from the null model.

The price for enlarging this class by the additive test is the slower rate of alternatives that can be considered. If $h = O(n^{-1/5})$ is used, it is given by $n^{-9/20}$ against $n^{-1/2}$. Therefore, analyzing the parametric specification with $\widehat{T}$ provides a data analyzing tool, that is clearly more flexible than parametric test procedures.

5.2.4 Bootstrap Implementation

The asymptotic distribution of $\widehat{T}$ is driven by a U-statistic and a large number of lower order terms that are omitted. However, it is well known in nonparametric hypothesis testing, that the convergence of the test statistic to the underlying U-statistic is rather slow (see Hjellvik and Tjøstheim, 1995, or Li and Wang, 1998). Therefore, it is not advisable to rely on the asymptotic normality approximation in small or moderate samples. Beside that, the quantities arising in the expected value and variance of the test statistic have to be estimated, since they involve the unknown conditional variances $v^j(x^j)$. This could in principle be done by regressing the squared residuals $\widehat{U}_i^2$ nonparametrically on all dimensions of the predictors X^j. Nadaraya-Watson-type estimators would be given by $\widehat{v}_h^j(x^j) = \widehat{f}_h^j(x^j)^{-1} n^{-1} \sum_{i=1}^n K_h(x^j, X_i^j) \widehat{U}_i^2$. Assuming smoothness of the conditional variances, it is not difficult to show that the use of these estimators leads to a consistent test statistic. Alternatively, other pre-estimators could be considered.

However, for small samples the asymptotic approximation is usually not valid. Therefore this subsection proposes the implementation of a wild bootstrap procedure. The bootstrap algorithm is performed in the following way

1. Construct parametric residuals $\widehat{U}_i = Y_i - G(X_i, \widehat{\theta})$.

2. Construct the residuals of the additive projection $\widehat{\varepsilon}_i = \widehat{U}_i - \sum_{j=0}^{d} \widetilde{m}^j_{\widetilde{h}}(X^j_i)$.
3. Generate independent and identically distributed random variables $\eta^*_1, \ldots, \eta^*_n$ from a distribution $\widehat{F}$ for all $i = 1, \ldots, n$.
4. Construct the bootstrap sample $(Y^*_i, X^*_i), i = 1, \ldots, n$ by $Y^*_i = G(X_i, \widehat{\theta}) + \widehat{\varepsilon}_i \eta^*_i$ and $X^*_i = X_i$.
5. Calculate $\widehat{T}^*$ from the bootstrap sample.
6. Repeat steps 3 to 5 B times to obtain critical values for $\widehat{T}$.

The predictors $X_1, \ldots, X_n$ remain unchanged in every bootstrap iteration. This is computationally convenient, since all density estimators and kernel weights in the smooth backfitting algorithm remain unchanged. Even though the iterative backfitting procedure has to be calculated B times, the computation time of the algorithm is not too high if a fast implementation is used.

Denote with $\mathbf{E}^*(\cdot) = \mathbf{E}(\cdot \mid (X_1, Y_1), \ldots, (X_n, Y_n))$ the conditional expectation of a random variable given the whole data sample. To derive the validity of the bootstrap method the following assumption is required formally.

Assumption 5.3. *For the bootstrap*

1. *For the bootstrap distribution $\widehat{F}$ it holds that $\mathbf{E}^* \eta^*_i = 0$, $\mathbf{E}^*(\eta^*_i)^2 = 1$ and $\mathbf{E}^*(\eta^*_i)^4 < \infty$ for all $i = 1, \ldots, n$.*
2. *Denote with $\widehat{\theta}^*$ the parametric estimator calculated from the bootstrap sample. Then it holds that $\widehat{\theta}^* - \widehat{\theta} = O_P(n^{-1/2})$.*
3. *The bandwidth sequence satisfies $n\widetilde{h} \to \infty$.*

It is not formally required that $\mathbf{E}^* \eta^3_i = 1$, since the proof of the bootstrap result will not be based on a formal Edgeworth expansion. But simulations provide evidence that mimicking three moments leads to higher order approximations of the distribution of the test statistic, which improves the finite sample behavior (see Li and Wang, 1998, for formal evidence of this finding in kernel based tests). The second part of Assumption 5.3 is not restrictive. It is not difficult to establish for usual parametric estimators $\widehat{\theta}$.

Theorem 5.4. *Let Assumptions 5.1–5.3 hold. Then under H_0 it holds that*

$$n\sqrt{h}\widehat{T}^* - h^{-1/2} B_T \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T),$$

conditional on the data for $n \to \infty$ with probability tending to one.

Alternatively, the bootstrap observations could be constructed under the null hypothesis, i. e. $Y_i^* = G(X_i, \widehat{\theta}) + \widehat{U}_i \eta_i^*$. In that case, the same result as in Theorem 5.4 could be proved. But, under the alternative, the residuals $\widehat{U}_i$ do not consistently estimate the true error $Y - G(X, \theta)$. This means that the bootstrap generates the distribution of the test under the closest parametric approximation to the true model with different error terms than in the true model. Because of the erratic behavior of the parametric residuals $\widehat{U}_i$ under the alternative, the simulated marginal conditional variances $v^j(x^j)$ can be too large. This can result in a loss in power of the bootstrap.

In contrast, the corrected residuals $\widehat{\varepsilon}_i$ are consistent under H_1 as well. Here the problem is that they underestimate the true error term under the null hypothesis if the bandwidth is too small. The result is a distortion of the level of the test. This can be reduced by using a different (larger) bandwidth $\widetilde{h}$ to construct the residuals $\widehat{\varepsilon}_i$ than to calculate the test statistic.

5.3 Extensions

5.3.1 Post-hoc-type Tests

If the F-test-type statistic $\widehat{T}$ leads to a rejection, the researcher will be interested in finding out by which regressor X^j this is caused. This can be done by testing which of the additive components $m^j(x^j)$ are significantly different from zero. The corresponding null hypothesis is given by

$$H_0^j \colon \mathcal{P}(m^j(X^j) = 0) = 1 \quad \text{for some } \theta \in \Theta$$

and as test statistic serves

$$\widehat{T}^j = \int_{\mathcal{A}|} \big(\widetilde{m}^j(x^j)\big)^2 \widehat{f}_h(x^j) w(x^j) \, \mathrm{d}x^j .$$

The asymptotic behavior of $\widehat{T}^j$ under H_0^j is given by

Theorem 5.5. *Let Assumptions 5.1 and 5.2 be fulfilled. Then it holds under H_0^j that*

$$n\sqrt{h}\widehat{T}^j - h^{-1/2} B_T^j \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T^j),$$

where

$$\Sigma_T^j = 2\|K * K\|_2^2 \int v^j(x^j)^2 w(x^j)^2 \, dx^j \quad \textit{and} \quad B_T^j = \|K\|_2^2 \int v^j(x^j) w(x^j) \, dx^j .$$

Under H_0 it holds that

$$n\sqrt{h}\begin{pmatrix}\widehat{T}^1\\ \vdots\\ \widehat{T}^d\end{pmatrix} - h^{-1/2}\begin{pmatrix}B_T^1\\ \vdots\\ B_T^d\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \begin{pmatrix}\Sigma_T^1 & & 0\\ & \ddots & \\ 0 & & \Sigma_T^d\end{pmatrix}).$$

Consider again the situation to test (5.4). If all other components of $g(x)$ except for $g^j(x)$ were known, an oracle test could be implemented using unobservable data $Y_i - \sum_{k\neq j} g^k(x^k)$. Denote with $\widehat{g}^j(x^j, \widehat{\theta})$ the parametric estimate of the unknown component. Then, a test statistic is given by

$$\widetilde{T}^j = \int_{\mathcal{A}^j} \Big(\widehat{f}_h^j(x^j)^{-1}\sum_{i=1}^n K_h(x^j, X_i^j)\big(Y_i - \sum_{k\neq j} g^k(x^k) - \widehat{g}^j(x^j, \widehat{\theta})\big)\Big)^2 \widehat{f}_h^j(x^j)\,\mathrm{d}x^j.$$

The asymptotic distribution of this test is derived by an application of Proposition 1 of Härdle and Mammen (1993) and is given by

$$\check{T}^j = \int_{\mathcal{A}^|} \big(\widetilde{m}^j(x^j)\big)^2 \widehat{f}_h(x^j) w(x^j)\,\mathrm{d}x^j.$$

Obviously, the first part of Theorem 5.5 shows that $\widehat{T}^j$ has an oracle property in the sense that this test statistic has the same asymptotic distribution as the oracle test.

The second part of the theorem states that the d different test statistics $\widehat{T}^1, \dots,$ $\widehat{T}^d$ are asymptotically independent. This can be used to test various additive components simultaneously. For studentized versions of these statistics, theory for multiple testing can be applied to obtain correct critical values. In finite samples correlation might be present and can be approximated by wild bootstrap. The joint distribution can be simulated in the same way as described in the last section. The only difference is that in step 5 of the algorithm all test statistics $\widehat{T}^{1,*}, \dots, \widehat{T}^{d,*}$ have to be calculated.

5.3.2 Omission of Additive Components

Apart from the parametric specification it is also of interest to test whether one component of the predictors has an influence on the conditional expectation at all. For this, assume that $g(x) \in \mathcal{G}^{ad}$ and consider the testing problem

(5.13) $$H_0^{j'} : g^j(x^j) = 0 \qquad \text{vs.} \qquad H_1^{j'} : g^j(x^j) \neq 0.$$

To test this, the smooth backfitting algorithm is applied directly to Y_i instead of $\widehat{U}_i$. Denote the corresponding estimators of the additive components with $\widetilde{g}^1, \dots, \widetilde{g}^d$. Then, a test statistic is given by

$$\widetilde{T}^j = \int_{\mathcal{A}^|} \left(\widetilde{g}^j(x^j)\right)^2 \widehat{f}_h(x^j) w(x^j) \,\mathrm{d}x^j.$$

From all theorems it is obvious that the parametric estimator does not influence the asymptotic distribution. Therefore this test statistic is under $H_0^{j'}$ asymptotically equivalent to $\widehat{T}^j$ under H_0^j and the first part of Theorem 5.5 applies. Only the marginal conditional variance has to be adjusted to $\widetilde{v}^j = \mathbf{E}((Y - g(X))^2 \mid X^j)$. But it is important to note that testing problem (5.13) is restricted to the case that $g(x)$ is fully additive. If $g(x) \notin \mathcal{G}^{ad}$ it can be the case that the conditional expectation is independent from X^j, but the additive projection is not. Therefore an application of $\widetilde{T}^j$ can produce misleading results.

The test statistic for this problem is again $\widehat{T}^j$ but for the calculation of the additive estimators Y_i is used instead of $\widehat{U}_i$. From the proofs it is obvious that $\widehat{U}_i = U_i + O_P(n^{-1/2})$ which means that the parametric estimation does not influence the asymptotic distribution. Then, the first part of Theorem 5.5 holds for this test statistic under $H_0^{j'}$ with variance given by $\widetilde{v}^j = \mathbf{var}(Y \mid X^j = x^j)$.

5.4 Simulation and Application

5.4.1 Monte Carlo Study

The simulation study will examine the performance of the test in finite (rather small) samples. Two data generating processes are used. First, a linear model will be simulated. As second specification, the nonlinear model of Fan and Jiang (2005) is simulated to compare the performance of the test derived in this chapter to the results based on the classical backfitting estimator.

The first model is given by

$$Y_i = \sum_{j=1}^{d} 2(4X_i^j - 2) + \lambda(4X_i^1 - 2)^2 + U_i, \tag{5.14}$$

with $X_j^d \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$ and $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. This specification was also used in Zhang and Dette (2004) to compare univariate test statistics. The model under the null hypothesis is given by $\lambda = 0$ and samples for different values of $\lambda \in [0, 0.75]$ are generated to estimate the power of $\widehat{T}$. The sample size is $n = 100$ and for 16 different values of λ 1 000 simulation runs are used. Under H_0 the model is linear and the parameters are estimated using ordinary least squares.

A lower dimensional model with $d = 3$ is considered first. To construct the additive residuals $\widehat{\varepsilon}_i$, the bandwidth $\widetilde{h} = 1.5n^{-1/5}/\sqrt{12}$ is implemented. This is motivated by the rule of thumb, $\mathbf{var}\, X_i^j = 1/12$ and the need to use larger bandwidths $\widetilde{h}$ to obtain consistent estimators of the residuals. The test statistic $\widehat{T}$ is calculated with a bandwidth $h = h_0 n^{-1/5}/\sqrt{12}$ and different values for h_0 are used to check the sensibility of the power to the bandwidth. All additive projections are calculated by approximating the integrals in (5.10) with 51 grid points. The test statistic is calculated as

$$\widehat{T} = \int_{\mathcal{A}} \Big(\sum_{j=0}^{d} \widetilde{m}_h^j(x^j) \Big)^2 \, \mathrm{d}x.$$

Recall that this corresponds to the choice of $w(x) = 1/\widehat{f}_h(x^j)$ in equation (5.12). The bootstrap samples are generated using $\eta_i^* = V_i/\sqrt{2} - (V_i^2 - 1)/2$ with $V_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ (see Mammen, 1993).

In Figure 5.1 a quantile plot of the distribution of the test statistic for $\lambda = 0$ and $h_0 = 1$ against a normal random variable is presented. Obviously the test statistic is not normally distributed. This provides gives evidence that the asymptotic results for $\widehat{T}$ do not hold in small samples and relying on the normal distribution would lead to wrong critical values. Therefore it is advisable to approximate the distribution by the bootstrap procedure described in Section 5.2.4.

To calculate the empirical power, three different bandwidth constants $h_0 = 0.5, 1, 1.5$ are used. The results are displayed in Figure 5.2. For all values of the bandwidth the test has good power against the alternatives. For small bandwidths the test tends to be too conservative. It is a typical result in nonparametric goodness-of-fit testing that small changes in the power are observed for different bandwidths. In general, large bandwidths have more power against low frequency alternatives, while small bandwidths allow to detect high frequency alternatives. Since in this specification the deviation from the null hypothesis is of low frequency (the function is only quadratic), the power should increase more rapidly for larger bandwidths. However, the effect is not strong.

To examine the influence of the error distribution, model (5.14) is now considered with different distributions of ε_i. Additionally to the standard normal distribution, it is now also simulated from a standardized t-distribution with 5

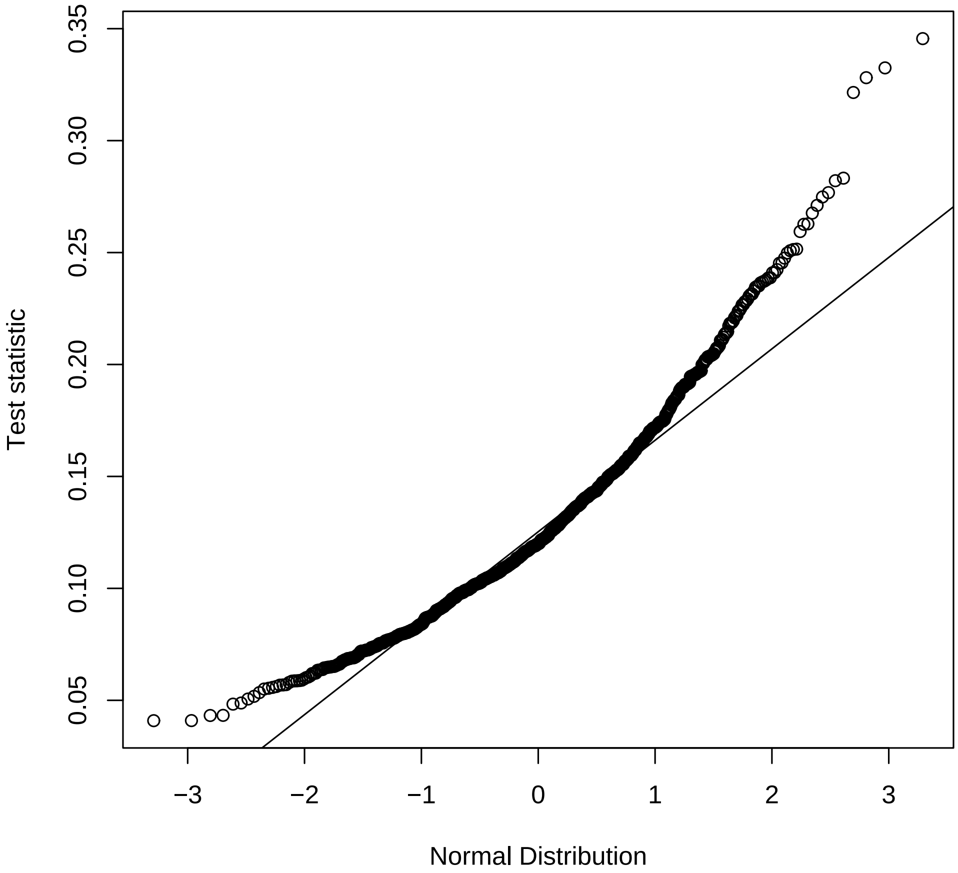

Figure 5.1: Quantile plot of the distribution of $\widehat{T}$ under H_0 for model (5.14). For the simulation $d = 3, n = 100$ and $h_0 = 1$ are used.

degrees of freedom and a standardized χ^2-distribution with 5 degrees of freedom. The asymptotic results of the second section have been established under the assumption that the errors have finite forth moment, which is minimal for the finiteness of the variance of $\widehat{T}$. Therefore the $t(5)$-distribution seems to be close to the boundary of the domain of attraction. Beside the leptokurtic errors, the χ^2-distribution is skewed and asymmetric. All other settings are unchanged (in particular $d = 3, n = 100$). Only four different values of λ are considered and the results are given in Table 5.1. No severe differences are found between the three different distributions for all bandwidths. The numerical results give evidence that the test is robust against different error distributions.

The advantage of the test statistics is that the asymptotic convergence is independent of the dimension of the regressors, circumventing the curse of dimen-

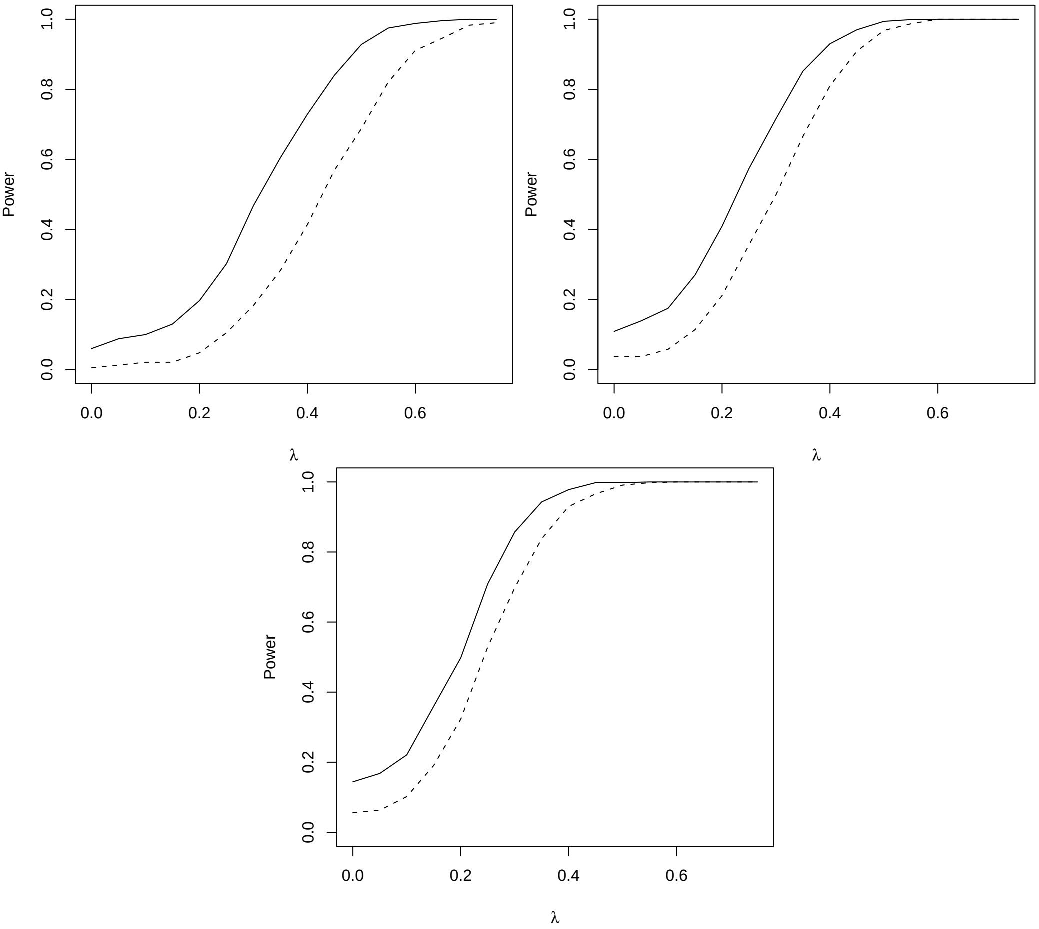

Figure 5.2: Empirical power of the test statistic for model (5.14) with $d = 3, n = 100$ for different bandwidths $h_0 = 0.5$ (upper left picture), $h_0 = 1$ (upper right picture) and $h_0 = 1.5$ (lower picture). The rejection rates are given for different significance levels $\alpha = 0.10$ (solid line) and $\alpha = 0.05$ (dashed).

sionality. To illustrate this, it will be simulated from model (5.14) with $d = 10$ and $n = 100$. The variance of $\sum_{j=0}^{d} \widehat{m}_{\widetilde{h}}^{j}(X_i^j)$ increases with d and therefore $\widetilde{h} = 2.5n^{-1/5}/\sqrt{12}$ has to be enlarged to obtain consistent estimates of the residuals. All other specifications are unchanged. The power of this high-dimensional model is displayed in Figure 5.3. The test still has good power but compared to the three-dimensional model, the increase in power with increasing λ is slower. This can be explained by the asymptotic results, because the variance Σ_T of the test is larger, if the number of dimensions increases (see Theorem 5.1).

Table 5.1: Power of $\widehat{T}$ under different error distributions

	$\alpha = 0.10$				$\alpha = 0.05$			
λ	0	0.25	0.5	0.75	0	0.25	0.5	0.75
				$h_0 = 0.5$				
$\mathcal{N}(0,1)$	0.060	0.302	0.928	0.999	0.005	0.105	0.688	0.990
$t(5)$	0.049	0.283	0.895	0.998	0.003	0.088	0.613	0.961
$\chi^2(5)$	0.056	0.298	0.911	0.998	0.013	0.079	0.655	0.955
				$h_0 = 0.5$				
$\mathcal{N}(0,1)$	0.109	0.574	0.994	1.000	0.037	0.356	0.968	1.000
$t(5)$	0.103	0.558	0.987	1.000	0.026	0.313	0.938	1.000
$\chi^2(5)$	0.115	0.582	0.981	1.000	0.028	0.327	0.930	0.999
				$h_0 = 0.5$				
$\mathcal{N}(0,1)$	0.144	0.709	0.998	1.000	0.056	0.529	0.991	1.000
$t(5)$	0.123	0.711	0.996	1.000	0.038	0.508	0.985	1.000
$\chi^2(5)$	0.130	0.720	0.995	1.000	0.047	0.493	0.981	1.000

A second data generating process is given by

$$Y_i = (X_i^1)^3 + \sin(3\pi X_i^2) + \sin(3\pi X_i^3)(1 + \lambda X_i^3) + \varepsilon_i. \tag{5.15}$$

The covariates $(X_i^1, X_i^2, X_i^3)'$ are independently drawn from a multivariate normal distribution with covariance matrix

$$\Sigma_1 = \frac{1}{9}\begin{pmatrix} 1 & \frac{1}{4} & 0 \\ \frac{1}{4} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The variables are truncated onto $[-0.5, 0.5]^3$, such that the actual correlation between X^1 and X^2 is smaller than $1/4$. The sample size is $n = 200$. Under

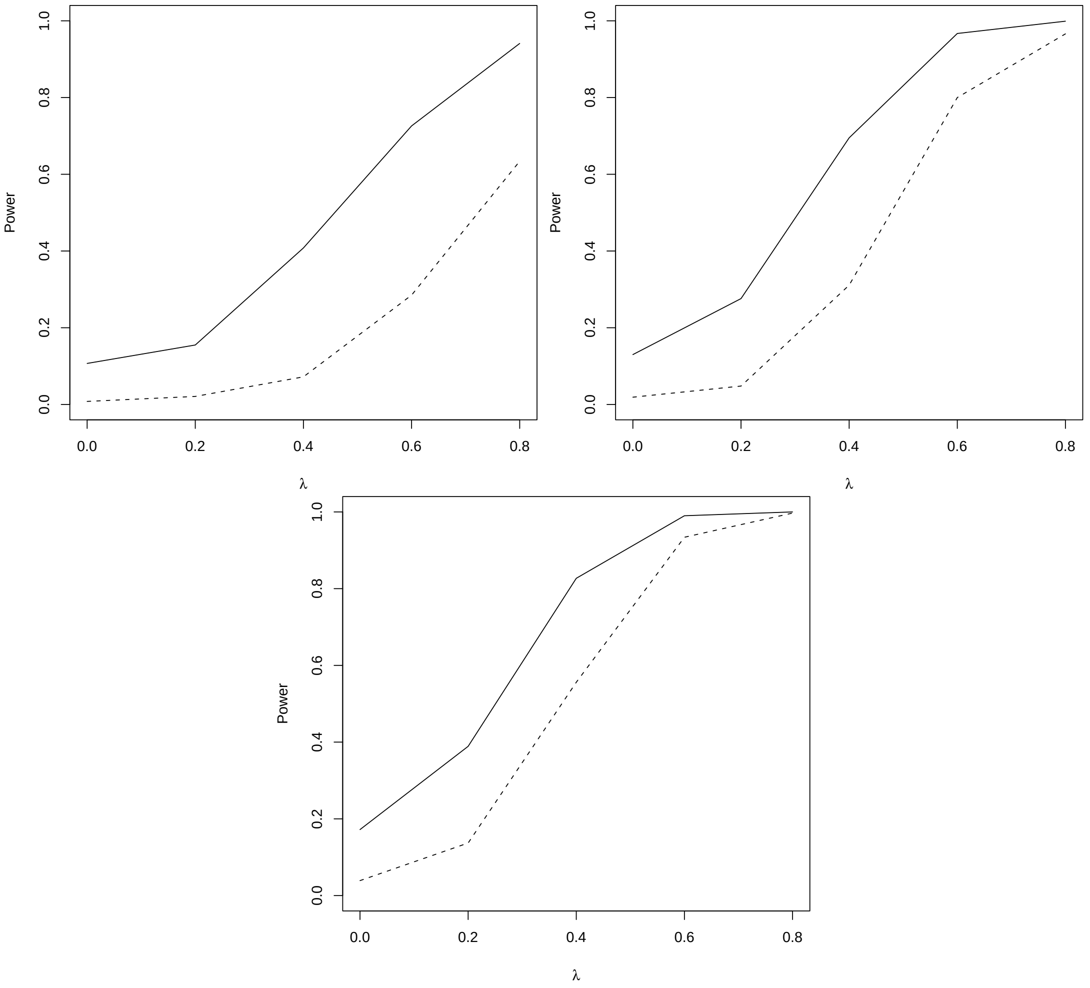

Figure 5.3: Empirical power of the test statistic for model (5.14) with $d = 10, n = 100$ for different bandwidths $h_0 = 0.5$ (upper left picture), $h_0 = 1$ (upper right picture) and $h_0 = 1.5$ (lower picture). The rejection rates are given for different significance levels $\alpha = 0.10$ (solid line) and $\alpha = 0.05$ (dashed).

the null hypothesis, the model is estimated using nonlinear least squares. The bandwidths are given by $h = h_0 n^{-1/5}/3$ and $\widetilde{h} = n^{-1/5}/2$ for all directions. The power is estimated over a grid of $\lambda \in [0, 1]$ by 500 simulation runs for each specification. This model is also used in Fan and Jiang (2005) and the power functions in Figure 5.4 can directly be compared with Figure 3 in that article. Both tests have very similar power functions and differences may vanish if the number of simulation runs increases. Again, the test has very good power across all bandwidths. For low values of λ which correspond to small deviations from

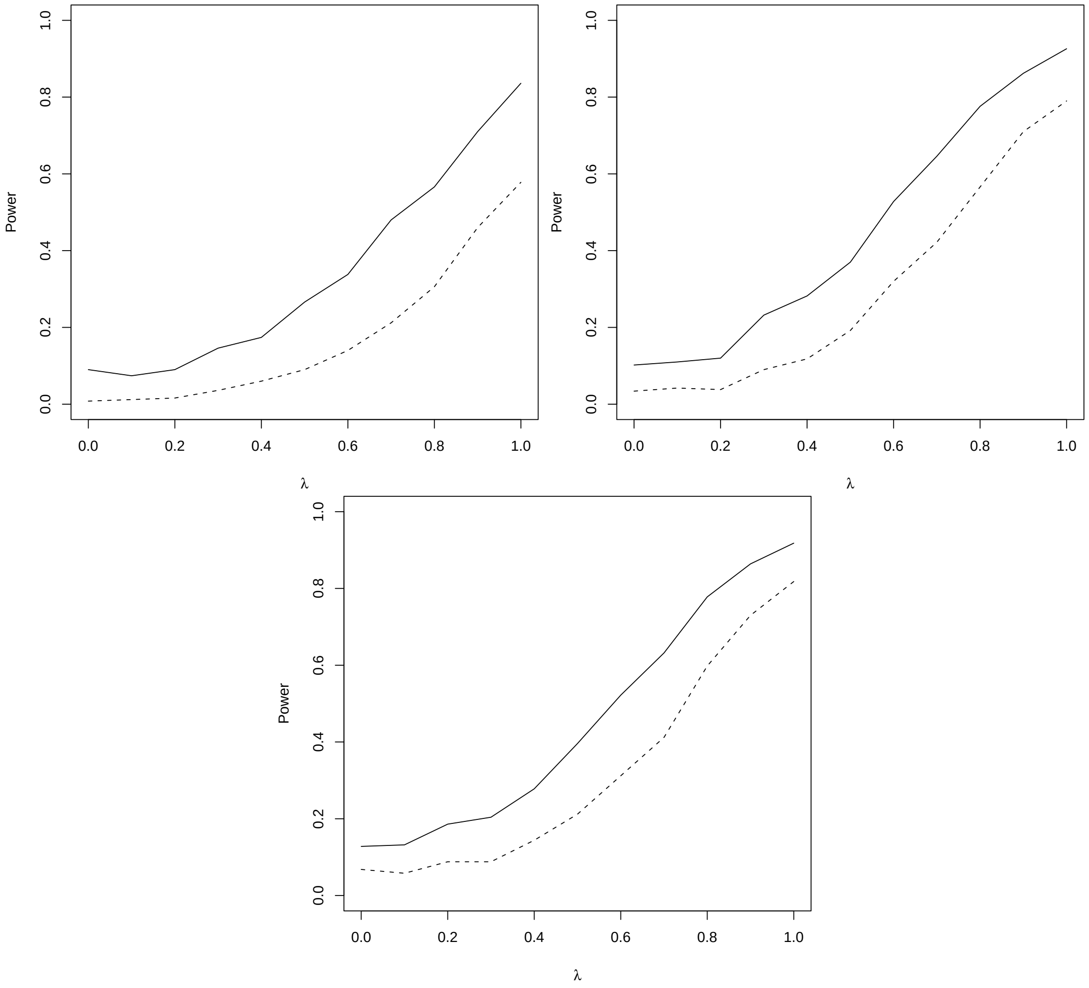

Figure 5.4: Empirical power of the test statistic for model (5.15) with $n = 200$ and covariance Σ_1 for different bandwidths $h_0 = 0.5$ (upper left picture), $h_0 = 1$ (upper right picture) and $h_0 = 1.5$ (lower picture). The rejection rates are given for different significance levels $\alpha = 0.10$ (solid line) and $\alpha = 0.05$ (dashed).

the null hypothesis the rejection rate is very low. But under the alternative, the limit of the parametric estimator $\bar{\theta}$ can be different from θ_0 and therefore the functional relation between the power and λ can be almost constant in that region.

In the simulation above, X^3 is independent from $(X^1, X^2)'$ and the correlation is limited to a rather small level of $1/4$. Because the smooth backfitting estimators are superior to classical backfitting in the case of correlated covariates it will now

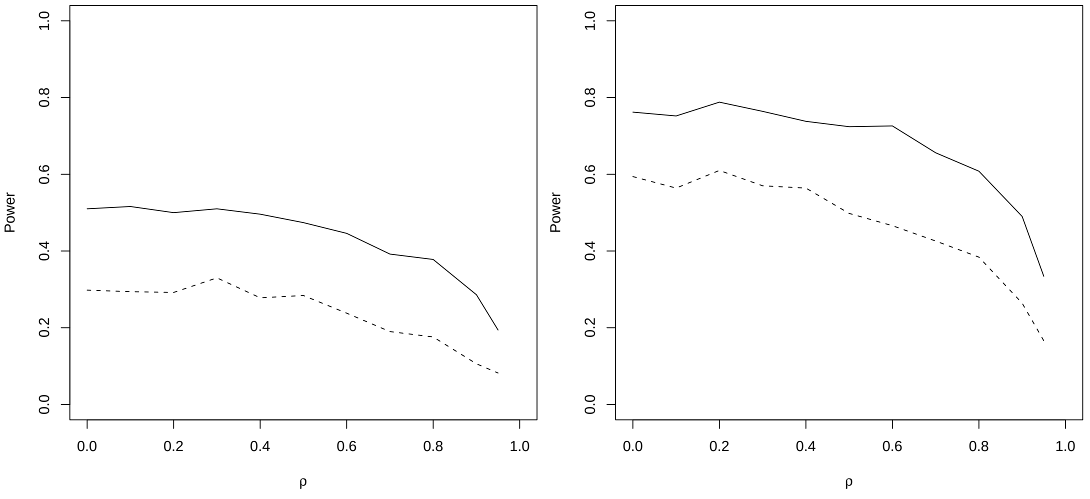

Figure 5.5: Empirical power of the test statistic for model (5.15) with $n = 200$ and covariance matrix Σ_ϱ with increasing correlation ϱ for different alternatives $\lambda = 0.6$ (left picture) and $\lambda = 0.8$ (right picture). The rejection rates are given for different significance levels $\alpha = 0.10$ (solid line) and $\alpha = 0.05$ (dashed). The bandwidth is given by $h_0 = 1.0$.

be simulated from model (5.15) using the covariance-matrix

$$\Sigma_\varrho = \frac{1}{9}\begin{pmatrix} 1 & \varrho & \varrho \\ \varrho & 1 & \varrho \\ \varrho & \varrho & 1 \end{pmatrix}$$

for increasing values of $\varrho \in [0, 0.95]$. The power is examined for two different alternatives with $\lambda = 0.6$ and $\lambda = 0.8$. The results are presented in Figure 5.5. Clearly the increasing correlation is associated with a loss of power, but this is not dramatic. The rejection rates are decreasing very slowly up to correlations of 0.8. Only for very extreme correlations the test looses its power. The stability of the power results from the smooth backfitting estimator. The theoretical results of Opsomer and Ruppert (1997) for classical backfitting restrict the correlation in the present setting to values $|\varrho| \leq 0.4$.

The Monte Carlo study provides evidence that some of the asymptotic properties of $\widehat{T}$ still hold in finite samples. In particular, the results for a high-dimensional model are very convincing. The results are also very stable against changes in the error distribution, different bandwidths and correlation structure of the covariates.

5.4.2 Application to Consumer Demand Data

To demonstrate the test statistic in practice, it will be applied to test a parametric specification of a demand system. A demand system with G goods is given by budget shares $W = (W^1, \ldots, W^G)'$, corresponding prices $P = (P^1, \ldots, P^G)'$ and total expenditure X. The parametric model under investigation is the Almost Ideal Demand System (AIDS), specified as

$$W^j = \alpha_j + \sum_{k=1}^{G} \gamma_{jk} P^k + \beta_j(\log X - a(P, \theta)) + \varepsilon,$$

for $j = 1, \ldots, G$ with

$$a(P, \theta) = \sum_{j=1}^{G} \alpha_j P^j + \frac{1}{2} \sum_{j,k} \gamma_{j,k} P^j P^k.$$

The test procedure is applied to household budget data from the Italian Central Statistical Office (ISTAT). This dataset was used by Bollino, Perali and Rossi (2000) and is distributed with the R-extension package micEcdat[4]. The sample consists of a demand system with three goods, namely food, housing and fuel and a miscellaneous good, where all other shares are aggregated. The sample size is 1 729. The parametric model is estimated by the iterative linear least squares estimator with Stone price index (see Blundell and Robin, 1999, for details). The parametric specification is tested for each good separately. The bandwidth is given by $h_j = h_0 \widehat{s}_j n^{-1/5}$ where $\widehat{s}_j$ denotes the empirical standard deviation of predictor j. To construct the additive residuals the bandwidth $\widetilde{h_j} = 1.5 h_j$ is implemented. The bootstrap distribution was given by η_i^* as in the simulation study and the bootstrap shares are normalized to add up to one.

Using the full sample the estimated p-values based on 999 bootstrap iterations was zero for all three goods. This result is not surprising, because the p-value depends on the sample size. Therefore a subsample of size 500 was selected randomly and critical values were calculated based on 399 bootstrap iterations. The results for different bandwidth constants are presented in Table 5.2. The model is rejected for the miscellaneous good group for all bandwidth choices. This is not surprising since the basis of aggregation is very large. For the other two goods the model is not rejected for larger bandwidths. This provides evidence that the AIDS model is an appropriate approximation if the goods are not aggregated in

[4] Downloadable from `www.cran.r-project.org`

Table 5.2: p-values for testing the AIDS

h_0	Food	Housing/Fuel	Miscellaneous
1.0	0.00	0.00	0.00
1.5	0.01	0.00	0.00
2.0	0.01	0.01	0.01
2.5	0.09	0.11	0.01
3.0	0.11	0.09	0.01
3.5	0.17	0.18	0.01
4.0	0.21	0.24	0.02

too large classes. A more sophisticated model including household characteristics (as in Bollino, Perali and Rossi, 2000) should be able to improve the fit.

Appendix

For abbreviation the random variables $W_i = (Y_i, X_i)$ and $U_i = Y_i - G(X_i, \theta)$ are introduced.

Proof of Theorem 5.1

The proof will use an expansion of the smooth backfitting estimator

$$\widetilde{m}_h^j(x^j) = \widehat{m}_h^j(x^j) + \frac{1}{n}\sum_{i=1}^{n} r_{ij}(x^j)\widehat{U}_i + o_P(n^{-1/2}), \tag{5.16}$$

uniformly in x^j with $r_{ij}(\cdot)$ absolutely uniformly bounded functions. This expansion is stated in Theorem 6.1 Mammen und Park (2005) under the assumption that the residuals ($\widehat{U}_i$ in this case) are independent and identically distributed and have conditional mean zero given X_i. Going through the proof of that theorem, this assumption is used to show that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\int \frac{\widehat{f}_h(x^j,x^k)}{\widehat{f}_h(x^j)\widehat{f}_h(x^k)}K_h(x^k,X_i^k)\,\mathrm{d}x^k\widehat{U}_i&\\ -\frac{1}{n}\sum_{i=1}^{n}\int \frac{f(x^j,X_i^k)}{f(x^j)f(X_i^k)}K_h(x^k,X_i^k)\,\mathrm{d}x^k\widehat{U}_i&\\ =\frac{1}{n}\sum_{i=1}^{n}\Delta_{k,j}(x^j,h)\widehat{U}_i = o_P(n^{-1/2})&\end{aligned}$$

holds uniformly in x^j (see equation (6.22) in Mammen and Park, 2005). To extend this, consider the decomposition

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Delta_{k,j}(x^j,h)\widehat{U}_i = \frac{1}{n}\sum_{i=1}^{n}\Delta_{k,j}(x^j,h)U_i&\\ +\frac{1}{n}\sum_{i=1}^{n}\Delta_{k,j}(x^j,h)\big(G(X_i,\theta)-G(X_i,\widehat{\theta})\big).&\end{aligned}$$

Because $\mathbf{E}(U_i \mid X_i) = 0$ for the first term on the right equation (6.22) in Mammen and Park (2005) applies. For the second part the mean value theorem is applied for the parametric function $G(X_i,\theta) - G(X_i,\widehat{\theta}) = (\theta-\widehat{\theta})^T\nabla_\theta G(X_i,\bar{\theta})$, where $\bar{\theta}$, depending on X_i, lies between $\widehat{\theta}$ and θ. Note that $\Delta_{k,j}(x^j,h) = O_P(h)$ uniformly in x^j and h. Using the rate of convergence of the parametric estimator it can be deduced that

$$\begin{aligned}\Big|\frac{1}{n}\sum_{i=1}^{n}\Delta_{k,j}(x^j,h)(G(X_i,\theta)-G(X_i,\widehat{\theta}))\Big| &= O_P(hn^{-1/2})\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^T\nabla_\theta G(X_i,\bar{\theta})\\ &= o_P(n^{-1/2}),\end{aligned}$$

where $\mathbf{1} = (1,\dots,1)^T$. This completes the proof of (5.16).

Now, turn to the test statistic. First, the full-dimensional density estimator will be replaced with the true density, since $\widehat{f}_h(x) = f(x) + o_P(1)$ uniformly in x. Without loss of generality it can be assumed that $\widetilde{m}_0 = n^{-1}\sum_{i=1}^n \widehat{U}_i = 0$. Then, expansion (5.16) is used to decompose the test statistic as follows

$$\begin{aligned}\widehat{T} &= \int\Big(\sum_{j=1}^{d}\widetilde{m}_h(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x\\ &= \int\Big(\sum_{j=1}^{d}\widehat{m}_h(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x + \int\Big(\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}r_{ij}(x^j)\widehat{U}_i\Big)^2 f(x)w(x)\,\mathrm{d}x\end{aligned}$$

$$+ o_P(n^{-1/2})\Big(\int \sum_{j=1}^d \widehat{m}_h(x^j) f(x) w(x)\,\mathrm{d}x + \int \sum_{j=1}^d \frac{1}{n}\sum_{i=1}^n r_{ij}(x^j)\widehat{U}_i f(x) w(x)\,\mathrm{d}x\Big)$$

$$+ 2\int \Big(\sum_{j=1}^d \widehat{m}_h(x^j)\Big)\Big(\sum_{j=1}^d \frac{1}{n}\sum_{i=1}^n r_{ij}(x^j)\widehat{U}_i\Big) f(x) w(x)\,\mathrm{d}x + o_P(n^{-1})$$

(5.17)

$$= \widehat{T}_1 + \widehat{T}_2 + o_P(n^{-1/2})(\widehat{T}_3 + \widehat{T}_4) + \widehat{T}_5 + o_P(n^{-1}).$$

The theorem follows from showing the following convergence results for the components

(5.18)
$$n\sqrt{h}\widehat{T}_1 - h^{-1/2}B_T \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T) \qquad \widehat{T}_2 = o_P(n^{-1}h^{-1/2})$$
$$\widehat{T}_3 = o_P(n^{-1/2}h^{-1/4}) \quad \widehat{T}_4 = o_P(n^{-1/2}h^{-1/4})\widehat{T}_5 = o_P(n^{-1}h^{-1/2})$$

Convergence in probability of $\widehat{T}_2, \dots, \widehat{T}_5$ The terms of lower order are considered first. Replacing the numerator with its limit and expanding $\widehat{U}_i$ it holds that

$$\widehat{T}_3 = O_P(1)\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^d \int K_h(x^j, X_i^j) f(x^j)^{-1} w(x) f(x)\,\mathrm{d}x$$
$$\times \big(U_i + (\theta - \widehat{\theta})^T \nabla_\theta G(X_i, \bar{\theta})\big)$$
$$= O_P(1)\frac{1}{n}\sum_{i=1}^n \widetilde{K}(X_i)U_i + O_P(n^{-1/2})\frac{1}{n}\sum_{i=1}^n \widetilde{K}(X_i)\mathbf{1}^T\nabla_\theta G(X_i, \bar{\theta})$$
$$= o_P(n^{-1/2}h^{-1/4}),$$

with $\widetilde{K}(X_i) = \sum_{j=1}^d \int K_h(x^j, X_i^j) f(x^j)^{-1} w(x) f(x)\,\mathrm{d}x$ which is bounded by a constant. The last line follows from direct calculations, using $\mathbf{E}\,\widetilde{K}(X_i)U_i = 0, \mathbf{E}(\widetilde{K}(X_i)U_i)^2 = O(1)$ and similarly for the second term. $\widehat{T}_4 = o_P(n^{-1/2}h^{-1/4})$ is shown analogously.

Next, consider

$$\widehat{T}_5 = O_P(1)\frac{1}{n^2}\sum_{i,i'} \int \Big(\sum_{j=1}^d K_h(x^j, X_i^j) f(x^j)^{-1}\Big)\Big(\sum_{j=1}^d r_{i'j}(x^j)\Big) w(x) f(x)\,\mathrm{d}x \widehat{U}_i\widehat{U}_{i'}$$
$$= O_P(1)\frac{1}{n^2}\sum_{i=1}^n \widetilde{K}_i(X_i)\widehat{U}_i^2 + O_P(1)\frac{1}{n^2}\sum_{i<i'}(\widetilde{K}_i(X_{i'}) + \widetilde{K}_{i'}(X_i))\widehat{U}_i\widehat{U}_{i'}$$
$$= O_P(1)\big(\widehat{T}_{5,1} + \widehat{T}_{5,2}\big).$$

Obviously $\widetilde{K}_{i'}(X_i) = \sum_{j,j'} \int K_h(x^j, X_i^j) r_{i'j'}(x^{j'}) f(x^j)^{-1} w(x) f(x)\,\mathrm{d}x$ is bounded. It follows from similar arguments as for $\widehat{T}_3$ that $\widehat{T}_{5,1} = o_P(n^{-1/2}h^{-1/4})$. Concerning, $\widehat{T}_{5,2}$ the expansion

$$\begin{aligned}\widehat{U}_i\widehat{U}_{i'} = U_iU_{i'} + (\theta - \widehat{\theta})^T\big(\nabla_\theta G(X_i,\bar{\theta})U_{i'} + \nabla_\theta G(X_{i'},\bar{\theta})U_i\big)\\ + (\theta - \widehat{\theta})^T\nabla_\theta G(X_i,\bar{\theta})(\theta - \widehat{\theta})^T\nabla_\theta G(X_{i'},\bar{\theta})\end{aligned}$$

is used. Calculating mean and variance it is obtained that $\widehat{T}_{5,2} = o_P(n^{-1/2}h^{-1/4})$. Therefore it holds that

$$\widehat{T}_5 = o_p(n^{-1/2}h^{-1/4}).$$

The convergence of $\widehat{T}_2$ is shown in the same way.

Asymptotic distribution of $\widehat{T}_1$ Replacing the numerator of the Nadraya-Watson estimator with the density and solving the square, $\widehat{T}_1$ can be written as

$$\widehat{T}_1 = (\widehat{T}_{1,1} + \widehat{T}_{1,2} + \widehat{T}_{1,3})(1 + o_P(1)) + \widehat{T}_{1,4},$$

where

$$\widehat{T}_{1,1} = \sum_{i=1}^n \sum_{k=i+1}^n h_n(W_i;W_k) \qquad \widehat{T}_{1,2} = \frac{1}{2}\sum_{i=1}^n h_n(W_i;W_i),$$

with a kernel given by

$$h_n(W_i;W_k) = \frac{2}{n^2}U_iU_k\widetilde{K}(X_i,X_k),$$

where

$$\widetilde{K}(X_i,X_k) = \sum_{j,j'} \int K_h(x^j,X_i^j)K_h(x^{j'},X_k^{j'})\frac{f(x^j,x^{j'})}{f(x^{j'})f(x^j)}w(x^j,x^{j'})\,\mathrm{d}x^j\,\mathrm{d}x^{j'} \tag{5.19}$$

and

$$\widehat{T}_{1,3} = \frac{1}{n^2}\sum_{i,k} U_i\big(G(X_k,\theta) - G(X_k,\widehat{\theta})\big)\widetilde{K}(X_i,X_k)$$

$$\widehat{T}_{1,4} = \int\Big(\sum_{j=1}^d\sum_{i=1}^n K_h(x^j,X_i^j)\big(G(X_i,\theta) - G(X_i,\widehat{\theta})\big)\widehat{f}_h(x^j)^{-1}\Big)^2 f(x)w(x)\,\mathrm{d}x.$$

Now it has to be shown that

$$n\sqrt{h}\widehat{T}_{1,1} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma_T) \tag{5.20}$$

$$n\sqrt{h}\widehat{T}_{1,2} - h^{-1/2}B_T \xrightarrow{P} 0 \tag{5.21}$$

$$n\sqrt{h}\widehat{T}_{1,3} \xrightarrow{P} 0 \tag{5.22}$$

$$n\sqrt{h}\widehat{T}_{1,4} \xrightarrow{P} 0 \tag{5.23}$$

to proof the asymptotic distribution of $\widehat{T}_1$.

Asymptotic distribution of $\widehat{T}_{1,1}$ A change of variables is applied to obtain

$$
\begin{aligned}
h_n(W_i, W_k) = {} & \frac{2h}{n^2} U_i U_k \sum_{j=1}^{d} \int K_h(vh + X_i^j, X_i^j) K_h(vh + X_i^j, X_k^j) \frac{w(X_i^j + vh)}{f(X_i^j + vh)} \, \mathrm{d}v \\
& + \frac{2h^2}{n^2} U_i U_k \sum_{j' \neq j} \int K_h(vh + X_i^j, X_i^j) K_h(v'h + X_k^{j'}, X_k^{j'}) \\
& \qquad \frac{f(X_i^j + vh, X_k^{j'} + v'h)}{f(X_i^j + vh) f(X_k^{j'} + v'h)} w(X_i^j + vh, X_k^{j'} + v'h) \, \mathrm{d}v \, \mathrm{d}v'.
\end{aligned}
$$

This shows that the terms with $j \neq j'$ are of lower order. Note that the unmodified kernel differs from the modified kernel only at the boundary and the distance is of order $O(h)$. Therefore, asymptotically the modification has no influence on integrated statistics.

To derive the asymptotic distribution a central limit theorem for degenerated U-Statistics is used (see Lemma 3.1 by de Jong, 1987). According to this theorem it has to be shown that

$$
(5.24) \qquad \frac{\max_{1 \leq i \leq n} \sum_{k=1, k \neq i}^{n} \mathbf{E}\, h_n(W_i; W_k)^2}{\mathbf{var}\, \widehat{T}_{1,1}} \to 0 \qquad \text{and} \qquad \frac{\mathbf{E}\, \widehat{T}_{1,1}^4}{(\mathbf{var}\, \widehat{T}_{1,1})^2} \to 3
$$

and $n^2 h \, \mathbf{var}\, \widehat{T}_{1,1} \to 2\Sigma_T$ to deduce (5.20).

To show these three statements, first $\mathbf{var}\, h_n(\cdot, \cdot)$ is considered. Omitting the lower order terms in $h_n(\cdot, \cdot)$, it is derived that

$$
\begin{aligned}
\mathbf{var} h_n(W_i; W_k) & = \mathbf{E}\, h_n(W_i; W_k)^2 \\
& = \frac{4}{n^4 h^2} \sum_{j}^{d} \int (y_i - G(x_i, \theta))^2 (y_k - G(x_k, \theta))^2 \int K(u) K(u + (x_i^j - x_k^j)/h) \, \mathrm{d}u \\
& \quad \times \frac{w(x_i^j)}{f(x_i^j)} f(y_i, x_i) f(y_k, x_k) \, \mathrm{d}y_i \, \mathrm{d}x_i \, \mathrm{d}y_k \, \mathrm{d}x_k + o(n^{-4})
\end{aligned}
$$

$$= \frac{4}{n^4h^1}\|K*K\|_2^2\sum_{j=1}^d\int(y_i-G(x_i,\theta))^2(y_k-G(x_k^{-j},x_i^j,\theta))^2$$

$$\times\frac{w(x_i^j)^2}{f(x_i^j)^2}f(y_i,x_i)f(y_k,x_k^{-j},x_i^j)\,\mathrm{d}y_i\,\mathrm{d}x_i\,\mathrm{d}y_k\,\mathrm{d}x_k^{-j}(1+O(h))+o(n^{-4})$$

$$= \frac{4}{n^4h^1}\|K*K\|_2^2\sum_{j=1}^d\int(y_i-G(x_i,\theta))^2\frac{f(y_i,x_i)}{f(x_i^j)}\,\mathrm{d}y_i\,\mathrm{d}x_i^{-j}$$

$$\times(y_k-G(x_k^{-j},x_i^j,\theta))^2\frac{f(y_k,x_k^{-j},x_i^j)}{f(x_i^j)}\,\mathrm{d}y_k\,\mathrm{d}x_k^{-j}w(x_i^j)^2\,\mathrm{d}x^j+o(n^{-4}h^{-1})$$

(5.25)

$$= \frac{4}{n^4h^1}\|K*K\|_2^2\sum_{j=1}^d\int v(x^j)^2w(x^j)^2\,\mathrm{d}x^j+o(n^{-4}h^{-1}).$$

First, variables are changed to $v=(x_i^j-x_k^j)/h$. For abbreviation the notation $x_k^{-j}=(x_k^1,\dots,x_k^{j-1},x_i^j,x_k^{j+1},\dots,x_k^d)$ is introduced. Then, the final result is obtained from rearrangements of the terms.

Using (5.25) and the independence of the data it is easy to obtain

$$\max_{1\le i\le n}\sum_{\substack{k=1\\k\ne i}}^n\mathbf{E}\,h_n(W_i;W_k)^2=O(n^{-3}h^{-1}),$$

as well as

(5.26)

$$\mathbf{var}\,\widehat{T}_{1,1}=\sum_{i<k}\mathbf{var}\,h_n(W_i;W_k)=\frac{n(n-1)}{2}\,\mathbf{var}\,h_n(W_1;W_2)=\frac{2}{n^2h}\Sigma_T(1+o(1)).$$

Because $h_n(\cdot;\cdot)$ is centered the covariances cancel out. From this, he first condition in (5.24) follows.

Finally, the forth moment of $\widehat{T}_{1,1}$ has to be considered

(5.27)

$$\mathbf{E}\,\widehat{T}_{1,1}^4=\sum_{i_1<i_2}\mathbf{E}\,h_n(W_{i_1};W_{i_2})^4+3\sum_{i_1<i_2}\sum_{\substack{i_3<i_4\\(i_3,i_4)\ne(i_1,i_2)}}\mathbf{E}\,h_n(W_{i_1};W_{i_2})^2h_n(W_{i_3};W_{i_4})^2$$

$$+24\sum_{i_1<i_2}\sum_{i_3\ne i_1,i_2}\mathbf{E}\,h_n(W_{i_1};W_{i_2})^2h_n(W_{i_1};W_{i_3})h_n(W_{i_2};W_{i_3})$$

$$+3\sum_{i_1}\sum_{i_2\ne i_1}\sum_{i_3\ne i_1,i_2}\sum_{i_4\ne i_1,i_2,i_3}\mathbf{E}\,h_n(W_{i_1};W_{i_2})h_n(W_{i_2};W_{i_3})h_n(W_{i_3};W_{i_4})h_n(W_{i_4};W_{i_1}).$$

Here, all vanishing terms are already omitted. Similar calculations as done for $\mathbf{E}\, h_n(W_1; W_2)^2$ show that

$$\begin{aligned}\mathbf{E}\, h_n(W_1; W_2)^4 &= O(n^{-8}h^{-3})\\ \mathbf{E}\, h_n(W_1; W_2)^2 h_n(W_1; W_3)^2 &= O(n^{-8}h^{-2})\\ \mathbf{E}\, h_n(W_1; W_2)^2 h_n(W_1; W_3) h_n(W_2; W_3) &= O(n^{-8}h^{-2})\\ \mathbf{E}\, h_n(W_1; W_2) h_n(W_2; W_3) h_n(W_3; W_4) h_n(W_1; W_4) &= O(n^{-8}h^{-1}).\end{aligned}$$

By combinatorial arguments it follows that the forth moment of $\widehat{T}_{1,1}$ is asymptotically dominated by terms with $\mathbf{E}\, h_n(W_1; W_2)^2 h_n(W_3; W_4)^2 = (\mathbf{E}\, h_n(W_1; W_2)^2)^2$. In total it holds that

$$\frac{\mathbf{E}\, \widehat{T}_{1,1}^4}{(\mathbf{var}\, \widehat{T}_{1,1})^2} = \frac{12n^{-4}h^{-2}\Sigma_T^4(1 + o(1))}{(2n^{-2}h^{-1}\Sigma_T^2(1 + o(1)))^2} \longrightarrow 3.$$

which is the second condition in (5.24) and asymptotic normality of $\widehat{T}_{1,1}$ is established.

Convergence in probability of $\widehat{T}_{1,2}$ Starting with the expected value it holds that

$$\begin{aligned}\mathbf{E}\, \widehat{T}_{1,2} &= 2^{-1} n\, \mathbf{E}\, h_n(W_i, W_i)\\ &= \frac{1}{nh}\|K\|_2^2 \sum_{j=1}^{d} \int (y - G(x, \theta))^2 \frac{w(x)}{f(x^j)} f(y, x)\, \mathrm{d}y\, \mathrm{d}x + o(n^{-1}h^{-1})\\ &= \frac{1}{nh}\|K\|_2^2 \sum_{j=1}^{d} \int v(x^j) w(x^j)\, \mathrm{d}x^j + o(n^{-1}h^{-1}).\end{aligned}$$

First, the lower order parts of $h_n(\cdot; \cdot)$ are omitted and then a Taylor expansion is applied.

Convergence in probability is shown using Chebychev's inequality and calculating

$$\mathbf{var}\, \widehat{T}_{1,2}^2 = n\, \mathbf{var}(h_n(W_1, W_1)) = O(n^{-3}h^{-1}) = o(n^{-2}h^{-1}).$$

Convergence in probability of $\widehat{T}_{1,3}$ Expand $G(X_i,\theta)-G(X_i,\widehat{\theta})$ to obtain

$$\begin{aligned}\widehat{T}_{1,3} &= (\theta-\widehat{\theta})^T\frac{1}{n^2}\sum_{i=1}^n \widetilde{K}(X_i,X_i)U_i\mathbf{1}^T\nabla_\theta G(X_i,\bar{\theta})\\ &\quad+(\theta-\widehat{\theta})^T\frac{1}{n^2}\sum_{i<k}\widetilde{K}(X_i,X_k)\big(U_i\mathbf{1}^T\nabla_\theta G(X_k,\theta)-U_k\mathbf{1}^T\nabla_\theta G(X_i,\theta)\big)\\ &\quad+(\theta-\widehat{\theta})^T\frac{1}{n^2}\sum_{i\neq k}\widetilde{K}(X_i,X_k)U_i\mathbf{1}^T\nabla^2_\theta G(X_k,\bar{\theta})(\theta-\widehat{\theta})\\ &= O_P(n^{-1/2})\widehat{S}_1+O_P(n^{-1/2})\widehat{S}_2+O_P(n^{-1})\widehat{S}_3.\end{aligned}$$

Note that the intermediate point $\bar{\theta}$ depends on $\widehat{\theta}$. Direct calculations yield $\mathbf{E}\,\widehat{S}_1=O(n^{-1}),\mathbf{E}\,|\widehat{S}_1|=O(n^{-1})$ and $\mathbf{E}\,\widehat{S}_3=O(1),\mathbf{E}\,|\widehat{S}_3|=O(1)$. $\widehat{S}_2$ is a U-statistic with non-degenerated kernel

$$\widetilde{h}_n(W_i,W_k)=\widetilde{K}(X_i,X_k)\big(U_i\mathbf{1}^T\nabla_\theta G(X_k,\theta)-U_k\mathbf{1}^T\nabla_\theta G(X_i,\theta)\big).$$

By similar calculations as in the analysis of $h_n(\cdot;\cdot)$ it is shown that $\mathbf{E}\,\widetilde{h}_n(W_i,W_k)=0$ and $\mathbf{E}\,\widetilde{h}_n(W_i,W_k)^2=O(h^{-1})=o(n)$. This allows to apply Lemma 3.1 of Powell, Stock and Stoker (1989) to obtain

$$\widehat{S}_2-\widetilde{S}_2=o_P(n^{-1/2}),$$

where $\widetilde{S}_2$ is the projection of the U-statistic, given by

$$\widetilde{S}_2=\mathbf{E}\,\widetilde{h}_n(W_1,W_2)+\frac{2}{n}\sum_{i=1}^n\mathbf{E}\big(\widetilde{h}_n(W,W_i)\mid W_i\big)-\mathbf{E}\,\widetilde{h}_n(W_1,W_2).$$

Here, W is distributed as W_i independently of W_i. Since $\mathbf{E}\big(\widetilde{h}_n(W,W_i)\mid W_i\big)$ is a sequence of iid random variables with mean zero and finite second moment (note that two change of variables can be applied for $\widetilde{K}_n(X,X_i)$) it is obvious that $\widetilde{S}_2=O_P(n^{-1/2})$. This completes the proof of (5.22).

Convergence in probability of $\widehat{T}_{1,4}$ This follows directly from

$$|\widehat{T}_{1,4}|\leq\big(\sup_x|G(X_i,\theta)-G(X_i,\widehat{\theta})|\big)^2\int w(x)f(x)\,\mathrm{d}x=O_P(n^{-1}),$$

because the kernel is assumed to be positive and then the kernel density estimator cancels.

This completes the proof of the theorem. □

Proof of Theorem 5.2

Under the alternative, the residuals can be decomposed into a bias and a variance part $\widehat{U}_i = \widehat{U}_i^V + \widehat{U}_i^B$ where

$$\widehat{U}_i^V = U_i - \mathcal{P}(U \mid X_i) + G(X_i, \widetilde{\theta}) - G(X_i, \widehat{\theta}) \quad \text{and} \quad \widehat{U}_i^B = \mathcal{P}(U \mid X_i).$$

This defines a decomposition of the marginal Nadraya-Watson estimator $\widehat{m}_h^j(x^j) = \widehat{m}_h^{j,V}(x^j) + \widehat{m}_h^{j,B}(x^j)$, where $\widehat{m}_h^{j,S}(x^j) = \widehat{f}_h(x^j)^{-1} n^{-1} \sum_{i=1}^n K_h(x^j, X_i^j)\widehat{U}_i^S$ for $S = B, V$. Recall that the smooth backfitting estimator is defined via the marginal Nadraya-Watson estimator in equation (5.10). Replacing $\widehat{m}_h^j$ with the two components $\widehat{m}_h^{j,B}$ and $\widehat{m}_h^{j,V}$ respectively, a bias part $\widetilde{m}_h^{j,B}$ and a variance part $\widetilde{m}_h^{j,V}$ of the smooth backfitting estimator is defined as the solution to the respective version of equation (5.10) and it holds that $\widetilde{m}_h^j(x^j) = \widetilde{m}_h^{j,V}(x^j) + \widetilde{m}_h^{j,B}(x^j)$. Since $\mathbf{E}\big(U - \mathcal{P}(U \mid X) \mid X^j\big) = 0$ for all $j = 1, \ldots, d$,[5] representation (5.16) applies for $\widetilde{m}_h^{j,V}(x^j)$. For the bias part of the Nadaraya-Watson estimator it holds that

$$\widehat{m}_h^{j,B}(x^j) = \mathcal{P}(U \mid x^j) + \sum_{k \neq j} \int \mathcal{P}(U \mid x^k) \frac{\widehat{f}_h(x^j, x^k)}{\widehat{f}_h(x^j)} \, \mathrm{d}x^k \\ + h^2 \int \beta(x) \frac{f(x)}{f(x^j)} \, \mathrm{d}x^{-j} + o_P(h^2),$$

uniformly in x^j with

$$\beta(x) = \int u^2 K(u) \, \mathrm{d}u \sum_{j=1}^d \frac{\partial}{\partial x^j} \mathcal{P}(U \mid x^j) \frac{\partial}{\partial x^j} \log f(x) + \frac{1}{2} \frac{\partial^2}{\partial (x^j)^2} \mathcal{P}(U \mid x^j).$$

Because $\mathcal{P}(U \mid x)$ is an additive function, this is proofed in the same way as equation (112) in Mammen, Linton and Nielsen (1999). This representation of

[5] Consider the definition of the marginal conditional expectation, given as minimizer over $\mu(x^k)$ of

$$\int (u - \mathcal{P}(U \mid x) - \mu(x^k))^2 f(u, x) \, \mathrm{d}u \, \mathrm{d}x = \int (u - \mathcal{P}(U \mid x))^2 f(u, x) \, \mathrm{d}u \, \mathrm{d}x \\ + 2 \int (u - \mathcal{P}(U \mid x)) \mu(x^k) f(u, x) \, \mathrm{d}u \, \mathrm{d}x + \int \mu(x^k)^2 f(x^k) \, \mathrm{d}x^k.$$

The first term on the right cannot be minimized over $\mu(x^k)$. Because the additive projection is defined as minimization of (5.3) $u - \mathcal{P}(U \mid x)$ is orthogonal to the space of additive functions in x. As $\mu(x^k)$ is an additive function, the second term is zero and the third term is minimized by $\mu(x^k) = 0$.

the bias part allows to apply Theorem 3 in Mammen, Linton and Nielsen (1999) and it is obtained that

$$\widetilde{m}_h^{j,B}(x^j) = \mathcal{P}(U \mid x^j) + h^2\mathcal{P}(\beta(X) \mid x^j) - h^2\gamma_{n,j} + o_P(h^2) \tag{5.28}$$

uniformly in x^j with

$$\gamma_{n,j} = \int u^2 K(u)\,\mathrm{d}u \int \frac{\partial}{\partial x^j}\mathcal{P}(U \mid x^j)\frac{\partial}{\partial x^j} f(x) + \frac{1}{2} f(x^j)\frac{\partial^2}{\partial (x^j)^2}\mathcal{P}(U \mid x^j)\,\mathrm{d}x^j.$$

For this see equation 6.6 in Mammen and Park (2005) and note the wrong proof of equation (114) in Mammen, Linton and Nielsen (1999).

Decomposing into bias and variance part, the test statistic can be written as

$$\begin{aligned}
\widehat{T} &= \int \Big(\sum_{j=1}^d \widetilde{m}_h^j(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x \\
&= \int \Big(\sum_{j=1}^d \widetilde{m}_h^{j,V}(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x + \int \Big(\sum_{j=1}^d \widetilde{m}_h^{j,B}(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x \\
&\qquad + 2\int \Big(\sum_{j=1}^d \widetilde{m}_h^{j,V}(x^j)\Big)\Big(\sum_{j=1}^d \widetilde{m}_h^{j,B}(x^j)\Big) f(x)w(x)\,\mathrm{d}x \\
&= \widehat{T}_1 + \widehat{T}_2 + \widehat{T}_3.
\end{aligned}$$

Since representation (5.16) applies, $\widehat{T}_1$ can be treated as in Theorem 5.1 and it holds that

$$\widehat{T}_1 - h^{-1/2}B_T = O_P(n^{-1}h^{-1/2}).$$

Using representation (5.28), we have that

$$\widehat{T}_2 = \int \Big(\sum_{j=1}^d \mathcal{P}(U \mid x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x + O_P(h^2).$$

For the cross term it holds that

$$\begin{aligned}
\widehat{T}_3 = 2&\int \Big(\sum_{j=1}^d \widehat{m}_h^{j,V}(x^j)\Big) P(U \mid x) f(x)w(x)\,\mathrm{d}x \\
&+ O_P(n^{-1/2}) \int \mathcal{P}(U \mid x) f(x)w(x)\,\mathrm{d}x \\
&+ O_P(h^2) \int \sum_{j=1}^d \widehat{m}_h^{j,V}(x^j) f(x)w(x)\,\mathrm{d}x + O_P(n^{-1/2}h^2)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{n}\sum_{i=1}^{n}\widehat{U}_i\sum_{j=1}^{d}\int K_h(X_i^j,x^j)f(x^j)^{-1}P(U\mid x)f(x)w(x)\,\mathrm{d}xO_P(1)+O_P(n^{-1/2})\\
&= O_P(n^{-1/2}).
\end{aligned}
$$

In total $\widehat{T}$ is dominated by $\widehat{T}_2$ which converges to a constant under H_1. □

Proof of Theorem 5.3

Under the local alternative, the residuals are decomposed according to $\widehat{U}_i = \widetilde{U} + g_n(X_i)$. Analogously to the proof of (5.16) it can be shown that under H_{1n} the following extension of the estimator holds

$$
\widetilde{m}_h^j(x^j) = \widehat{m}_h^j(x^j) + \frac{1}{n}\sum_{i=1}^{n} r_{ij}(x^j)\widehat{U}_i + o_P(\lambda_n^{-1/2}).
$$

With this extension the test statistic is decomposed as in (5.17) and the lower order terms are bounded as in the proof of Theorem 5.1. In total it is obtained that

$$
\begin{aligned}
\widehat{T} &= \int\Big(\sum_{j=1}^{d}\widehat{m}_h^j(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x + o_P(n^{-1}h^{-1/2})\\
&= \int\Big(\sum_{j=1}^{d}\widehat{f}_h(x^j)^{-1}\frac{1}{n}\sum_{i=1}^{n}K_h(x^j,X_i^j)\widetilde{U}_i\Big)^2 f(x)w(x)\,\mathrm{d}x\\
&\quad+\int\Big(\sum_{j=1}^{d}\widehat{f}_h(x^j)^{-1}\frac{1}{n}\sum_{i=1}^{n}K_h(x^j,X_i^j)g_n(X_i)\Big)^2 f(x)w(x)\,\mathrm{d}x\\
&\quad+2\int\Big(\sum_{j=1}^{d}\widehat{f}_h(x^j)^{-1}\frac{1}{n}\sum_{i=1}^{n}K_h(x^j,X_i^j)\widetilde{U}_i\Big)\\
&\qquad\times\Big(\sum_{j=1}^{d}\widehat{f}_h(x^j)^{-1}\frac{1}{n}\sum_{i=1}^{n}K_h(x^j,X_i^j)g_n(X_i)\Big)f(x)w(x)\,\mathrm{d}x + o_P(n^{-1}h^{-1/2})\\
&= \widehat{T}_1+\widehat{T}_2+\widehat{T}_3+o_P(n^{-1}h^{-1/2}).
\end{aligned}
$$

Under H_{1n} it holds that $\widetilde{U}_i = \breve{U}_i + G(X_i,\widetilde{\theta}) - G(X_i,\widehat{\theta})$ and $\mathbf{E}(\breve{U}_i \mid X^j)=0$ for all $j=1,\dots,d$ where $\breve{U}_i = Y_i - G(X_i,\theta) - g_n(X_i)$. Therefore the first term $\widehat{T}_1$ can be treated in the same way as the test statistic under the null hypothesis and it follows from (5.18) that

$$
n\sqrt{h}\widehat{T}_1 - h^{-1/2}B_T \xrightarrow{\mathcal{D}} \mathcal{N}(0,\Sigma_T). \tag{5.29}
$$

Next, turn to

$$\begin{aligned}\widehat{T}_3 = O_P(1)\frac{1}{n^2}\sum_{i,k}^{n}\widetilde{K}(X_i,X_k)g_n(X_i)\breve{U}_k \\ + O_P(1)\frac{1}{n^2}\sum_{i,k}^{n}\widetilde{K}(X_i,X_k)g_n(X_k)\big(G(X_i,\widetilde{\theta})-G(X_i,\widehat{\theta})\big) \\ = O_P(1)(\widehat{T}_{3,1}+\widehat{T}_{3,2})\end{aligned}$$

Decomposing both parts into the terms with $i = k$ and $i \neq k$ it is direct to show that $\widehat{T}_{3,1} = O_P(n^{-1/2}\lambda_n^{-1/2})$ and $\widehat{T}_{3,2} = O_P(n^{-1/2}\lambda_n^{-1/2})$. This follows by calculating the mean and the variance for the parts with $i = k$ and applying Lemma 3.1 in Powell, Stock and Stoker (1989) for the parts with $i \neq k$. Then it follows that

$$\widehat{T}_3 = o_P(n^{-1}h^{-1/2}).$$

Finally, consider

$$\begin{aligned}\widehat{T}_2 &= O_P(1)\frac{1}{n^2}\sum_{i=1}^{n}g_n(X_i)^2\widetilde{K}(X_i,X_i) + O_P(1)\frac{2}{n^2}\sum_{i<k}g_n(X_i)g_n(X_k)\widetilde{K}(X_i,X_k) \\ &= O_P(1)(\widehat{T}_{2,1}+\widehat{T}_{2,2}).\end{aligned}$$

As in the proof of (5.21) it follows that

$$n\sqrt{h}\widehat{T}_{2,1} - h^{-1/2}\|K\|_2^2 B_L \xrightarrow{P} 0. \tag{5.30}$$

For $\widehat{T}_{2,2}$ again Lemma 3.1 in Powell, Stock and Stoker (1989) is applied to obtain

$$\widehat{T}_{2,2} = o(\lambda_n^2) + o_P(\lambda_n^2) = o_P(n^{-1}h^{-1/2}). \tag{5.31}$$

Putting together (5.29)–(5.31), the statement of the theorem follows. □

Proof of Theorem 5.4

Introduce $\widehat{U}_i^* = Y_i^* - G(X_i,\widehat{\theta}^*)$ and decompose

$$\widehat{U}_i^* = \widehat{\varepsilon}_i\eta_i^* + G(X_i,\widehat{\theta}) - G(X_i,\widehat{\theta}^*).$$

By Assumption 5.3 the bootstrap version of the parametric estimator $G(x,\widehat{\theta}^*)$ can be expanded as $G(x,\widehat{\theta})$. Analogously to equation (5.16), the bootstrap versions of the backfitting estimators can be expanded to

$$\widetilde{m}_h^{j,*}(x^j) = \widehat{m}_h^{j,*}(x^j) + \frac{1}{n}\sum_{i=1}^{n}r_{ij}(x^j)\widehat{U}_i^* + o_P(n^{-1/2}).$$

Here $\widehat{m}_h^{j,*}(x^j)$ are the marginal Nadraya-Watson estimators based on the bootstrap data. Using this extension and similar arguments as to show (5.18) yield

$$\begin{aligned}\widehat{T}^* &= \int \Big(\sum_{j=1}^d \widehat{m}_h^{j,*}(x^j)\Big)^2 f(x)w(x)\,\mathrm{d}x + o_P(n^{-1}h^{-1/2})\\ &= (\widehat{T}^*_{1,1} + \widehat{T}^*_{1,2} + \widehat{T}^*_{1,3})(1 + o_P(1)) + \widehat{T}^*_{1,4} + o_P(n^{-1}h^{-1/2}).\end{aligned}$$

Here,

$$\begin{aligned}\widehat{T}^*_{1,1} &= \sum_{i=1}^n \sum_{k=i+1}^n h_n(W_i^*; W_k^*) \qquad\qquad \widehat{T}^*_{1,2} = \frac{1}{2}\sum_{i=1}^n h_n(W_i^*; W_i^*)\\ \widehat{T}^*_{1,3} &= \frac{1}{n^2}\sum_{i,k} \widehat{\varepsilon}_i \eta_i^* \big(G(X_k, \widehat{\theta}) - G(X_k, \widehat{\theta}^*)\big)\widetilde{K}(X_i, X_k)\\ \widehat{T}^*_{1,4} &= \int \Big(\sum_{j=1}^d \sum_{i=1}^n K_h(x^j, X_i^j)\big(G(X_i, \widehat{\theta}) - G(X_i, \widehat{\theta}^*)\big)\widehat{f}_h(x^j)^{-1}\Big)^2 f(x)w(x)\,\mathrm{d}x.\end{aligned}$$

with $\widetilde{K}(X_i, X_k)$ as in (5.19) and the kernel is given by

$$h_n(W_i^*; W_k^*) = \frac{2}{n^2}\widehat{\varepsilon}_i \eta_i^* \widehat{\varepsilon}_k \eta_k^* \widetilde{K}(X_i, X_k).$$

By expanding the parametric bootstrap estimator and using that $\mathbf{E}\,\widehat{\varepsilon}_i\eta_i^* = 0$ and $\mathbf{E}(\widehat{\varepsilon}_i\eta_i^*)^2 = \mathbf{E}(\widehat{\varepsilon}_i)^2 = O(n^{-1})$ it is shown as in the proof of Theorem 5.1 that

$$\widehat{T}^*_{1,3} = o_P(n^{-1}h^{-1/2}) \quad \text{and} \quad \widehat{T}^*_{1,4} = o_P(n^{-1}h^{-1/2}).$$

Then, the statement of the theorem follows from

$$n\sqrt{h}\widehat{T}^*_{1,1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma_T) \tag{5.32}$$

$$n\sqrt{h}\widehat{T}^*_{1,2} - h^{-1/2}B_T \xrightarrow{P} 0, \tag{5.33}$$

where the convergence in distribution is conditional on the data with probability tending to one.

Asymptotic distribution of $\widehat{T}^*_{1,1}$ By construction, $\mathbf{E}^*\, h_n(W_i^*; W_k^*) = 0$ and $\mathbf{E}^*\, h_n(W_i^*; W_k^*)^2 = \frac{4}{n^4}\widehat{\varepsilon}_i^2\widehat{\varepsilon}_k^2\widetilde{K}(X_i, X_k)^2$. To further analyze the second term, recall that

$$\begin{aligned}\widehat{\varepsilon}_i &= U_i + G(X_i, \theta) - G(X_i, \widehat{\theta}) - \sum_{j=0}^d \widetilde{m}_{\widetilde{h}}^j(x^j)\\ &= U_i + O_P(n^{-1/2}) + O_P(n^{-1/2}\widetilde{h}^{-1/2}).\end{aligned}$$

This follows from an application of expansion (5.16) to $\widetilde{m}^j_{\widetilde{h}}(x^j)$ and the assumptions on the parametric estimator. As seen above, it holds that $\widetilde{K}_h(X_i, X_k)^2 = O_P(h^{-1})$ if $i \neq k$. This yields in total that

$$\mathbf{E}^*\, h_n(W_i^*; W_k^*)^2 = h_n(W_i; W_k)^2 + O_P(n^{-9/2}h^{-1}\widetilde{h}^{-1/2}).$$

The asymptotic normality follows by showing that the conditions of Lemma 3.1 hold with probability tending to one, i. e.
(5.34)

$$\frac{\max_{1\le i\le n}\sum_{k=1,k\neq i}^{n}\mathbf{E}^*\, h_n(W_i^*; W_k^*)^2}{\mathbf{var}^*\,\widehat{T}^*_{1,1}} \xrightarrow{P} 0 \qquad \text{and} \qquad \frac{\mathbf{E}^*(\widehat{T}^*_{1,1})^4}{(\mathbf{var}^*\,\widehat{T}^*_{1,1})^2} \xrightarrow{P} 3$$

and $n^2h\,\mathbf{var}^*\,\widehat{T}^*_{1,1} \xrightarrow{P} 2\Sigma_T$.
Consider the variance first

$$n^2h\,\mathbf{var}^*\,\widehat{T}^*_{1,1} = n^2h\sum_{i<k} h_n(W_i, W_k)^2 + O_P(n^{-1/2}\widetilde{h}^{-1/2}) = n^2h\widehat{T}_{1,1} + o_P(1) \xrightarrow{P} 2\Sigma_T.$$

The limit follows from (5.25) and convergence in probability from (5.26).
Recall from the calculations in (5.27) that $\widehat{T}^4_{1,1}$ is dominated by terms with $h_n(W_i, W_k)^2 h_n(W_{i'}, W_{k'})^2$ (remember that all cross terms converge to zero). Then it is obtained that

$$\begin{aligned} n^4h^2\,\mathbf{E}^*(\widehat{T}^*_{1,1})^4 &= n^4h^2\widehat{T}^4_{1,1} + O_P(n^{-1}\widetilde{h}^{-1}) \\ &= 3n^4h^2\sum_{i_1<i_2}\ \sum_{\substack{i_3<i_4\\(i_3,i_4)\neq(i_1,i_2)}} h_n(W_{i_1}; W_{i_2})^2 h_n(W_{i_3}; W_{i_4})^2 + o_P(1) \\ &\xrightarrow{P} 12\Sigma_T^2. \end{aligned}$$

Convergence in probability follows from Chebychev's inequality and the fact that $\mathbf{var}\, h_n(W_1, W_2)^2 = O(n^{-8}h^{-3})$. From this, the second condition in (5.34) is obtained.
Finally, an application of Markov's inequality with the first moment shows that

$$n^2h\sum_{k=1}^{n} h_n(W_i, W_k)^2 = O_P(n^{-1})$$

for all i. This shows the first condition in (5.34) and therefore statement (5.32) follows.

Convergence in probability of $\widehat{T}^*_{1,2}$ Expand the residual to obtain

$$\begin{aligned}
\widehat{T}^*_{1,2} &= \frac{1}{n^2}\sum_{i=1}^{n} U_i\eta_i^*\widetilde{K}(X_i,X_i) + \frac{1}{n^2}\sum_{i=1}^{n}(G(X_i,\theta) - G(X_i,\widehat{\theta}) \\
&\quad - \sum_{j=0}^{d}\widetilde{m}^j_{\widetilde{h}}(x^j))\eta_i^*\widetilde{K}(X_i,X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} U_i\eta_i^*\widetilde{K}(X_i,X_i) + O_P(n^{-3/2}h^{-1/2}\widetilde{h}^{-1/2}),
\end{aligned}$$

because $\sum_{j=0}^{d}\widetilde{m}^j_{\widetilde{h}}(x^j) = O_P(n^{-1/2}\widetilde{h}^{-1/2})$ and $\widetilde{K}(X_i,X_i) = O_P(h^{-1})$. Using iterated expectations, the first term is analyzed as in the proof of (5.21).
This completes the proof of Theorem 5.4. □

Proof of Theorem 5.5

The asymptotic normality follows by the same calculations as in the proof of Theorem 5.1. □

References

[1] Aït-Sahalia, Y. (1996a). Nonparametric pricing of interest rate derivative securities, *Econometrica* **64**, 527–560.

[2] Aït-Sahalia, Y. (1996b). Testing continuous time models of the spot interest rate, *The Review of Financial Studies* **9**, 385–426.

[3] Aït-Sahalia, Y., Bickel, P. and T. Stoker (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities, *Journal of Econometrics* **105**, 363–412.

[4] Bandi, F. and G. Moloche (2001). On the functional estimation of multivariate diffusion processes, *Working paper*, University of Chicago.

[5] Bandi, F. and T. Nguyen (2003). On the functional estimation of jump-diffusion models, *Journal of Econometrics* **116**, 293–328.

[6] Bandi F. and P. Phillips (2003). Fully nonparametric estimation of scalar diffusion models, *Econometrica* **71**, 241–283.

[7] Bierens, H. (1982). Consistent model specification test, *Journal of Econometrics* **20**, 105–134.

[8] Bierens, H. (1990). A consistent conditional moment test of functional form, *Econometrica* **58**, 1443–1458.

[9] Bierens, H. and W. Ploberger (1997). Asymptotic theory of integrated conditional moment tests, *Econometrica* **65**, 1129–1151.

[10] Blundell, R. and J. M. Robin (1999). Estimation in large and disaggregated demand systems: an estimator for conditionally linear systems, *Journal of Applied Econometrics* **14**, 209–232.

[11] Bollino, C. A., F. Perali and N. Rossi (2000). Linear household technologies, *Journal of Applied Econometrics* **15**, 275–287.

[12] Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction (2nd edition)*, Lecture Notes in Statistics, Vol. 110, Springer-Verlag, New York.

[13] Brugiere (1993). Théorème de limite central pour un estimateur non paramétrique de la variance d'un processus de diffusion multidimensionnelle, *Ann Inst. Henri Poincaré* **29**, 357–289.

[14] Buja, A., Hastie, T. J. and R. J. Tibshirani (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**, 453–555.

[15] Christopeit, N. and S. G. N. Hoderlein (2006). Local partitioned regression, *Econometrica* **74**, 787–818.

[16] Cai, Z. and Y. Hong (2003). Nonparametric methods in continuous-time finance: a selective overview, In: *Recent Advances and trends in Nonparametric Statistics*, Eds: M.G. Akritis and D.M. Politis, 282–302

[17] Chapman, D. and N. Pearson (2000). Is the short rate drift actually nonlinear?, *Journal of Finance* **55**, 355–388.

[18] Das, M., Newey, W. and F. Vella (1999). Nonparametric estimation of sample selection models, *Working Paper*, Columbia University.

[19] de Jong, P. (1987). A central limit theorem for generalized quadratic forms, *Probability Theory and Related Fields* **75**, 261–275.

[20] Delgado, M. and W. Gonzáles-Manteiga (2001). Significance Testing in nonparametric regression based on the bootstrap, *Annals of Statistics* **29**, 1469–1507.

[21] Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators, *Annals of Statistics* **27**, 1012–1050.

[22] Dette, H. and I. Spreckelsen (2004). Some comments on specification tests in nonparametric absolutely regular processes, *Journal of Time Series Ananlysis* **25**, 159–172.

[23] Fan, J. (2005). A selective overview of nonparametric methods in financial econometrics (with discussion), *Statistical Science* **20**, 317–357.

[24] Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

[25] Fan, J. and J. Jiang (2005). Nonparametric Inference for Additive Models, *Journal of the American Statistical Association* **100**, 890–907.

[26] Fan, J. and Q. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression, *Biometrika* **85**, 645–660.

[27] Fan, J. and C. Zhang (2003). A reexamination of diffusion estimators with applications to financial model validation, *Journal of the American Statistical Association* **98**, 118–134.

[28] Fan, J., Zhang, C. M. and J. Zhang (2001). Generalized likelihood ratio statistics and Wilks phenomenon, *Annals of Statistics* **29**, 153–193.

[29] Fan, Y. and Q. Li (1996). Consistent model specification tests: omitted variables and semiparametric functional forms, *Econometrica* **64**, 865–890.

[30] Fan, Y. and Q. Li (1999). Central limit theorem for degenerate U-statistics of absolutely regular processes with applications to model specification testing, *Journal of Nonparametric Statistics* **10**, 245–271.

[31] Fan, Y. and Q. Li (2000). Consistent model specification tests: Kernel based tests versus Bierens' ICM tests, *Econometric Theory* **16**, 1016–1041.

[32] Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations, *Journal of Applied Probability* **30**, 790–804.

[33] Gobet, E., Hoffmann, M. and M. Reiß (2004). Nonparametric estimation of scalar diffusions based on low-frequency data, *Annals of Statistics* **32**, 2223–2253.

[34] Gozalo, P. (1997). Nonparametric bootstrap analysis with applications to demographic effects in demand functions, *Journal of Econometrics* **81**, 357–393.

[35] Haag, B. R. and S. G. N. Hoderlein (2005). Bootstrap specification testing in systems of equations, *Working Paper*, University of Mannheim.

[36] Haag, B. R., Hoderlein, S. G. N. and K. Pendakur (2005). Testing and imposing Slutsky symmetry in nonparametric demand systems, *Working Paper*, University of Mannheim.

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

[38] Härdle, W. and E. Mammen (1993). Comparing nonparametric vs. parametric regression fits, *Annals of Statistics* **21**, 1926–1947.

[39] Hall, P. (1984). Central limit theorems for integrated squared errors of multivariate nonparametric density estimators, *Annals of Statistics* **11**, 1156–1174.

[40] Hjellvik, V. and D. Tjøstheim (1995). Nonparametric tests of linearity for time series *Biometrika* **82**, 351–368.

[41] Hidalgo, J. (1992). Adaptive estimation in time series regression models with heteroscedasticity of unknown form, *Econometric Theory*, **8**, 161–187.

[42] Hoderlein, S. G. N. (2005) Nonparametric demand systems, instrumental variables and a heterogeneous population, *Working paper*, University of Mannheim.

[43] Hong, Y. (1993). Consistent specification testing using optimal nonparametric kernel estimation, *Working Paper,* Cornell University.

[44] Horowitz, J. and W. Härdle (1994). Testing a parametric model against a semiparametric alternative, *Econometric Theory* **10**, 821–848.

[45] Horowitz, J. and V. Spokoiny (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69**, 599–631.

[46] Kim, W. and G. Tripathi (2003). Nonparametric estimation of homogeneous functions. *Econometric Theory* **19**, 640–663.

[47] Kreiß, J. P., M. H. Neumann and Q. Yao (2002). Bootstrap tests for simple structures in nonparametric time series regression, *Working paper*, University of Brausnchweig.

[48] Lavergne, P. and Q. Vuong (1996). Nonparametric selection of regressors: the nonnested case, *Econometrica* **64**, 207–219.

[49] Lavergne, P. and Q. Vuong (2000). Nonparametric Significance Testing, *Econometric Theory* **16**, 576–601.

[50] Lewbel, A. (1995). Consistent nonparametric hypothesis tests with an application to Slutsky symmetry, *Journal of Econometrics* **67**, 379–401.

[51] Lewbel, A. (2001). Demand systems with and without errors, *American Economic Review* **91**, 611–618.

[52] Lewbel, A. and O. Linton (2005). Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions, *Working Paper*, Boston College.

[53] Li, Q. and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function, *Journal of Econometrics* **87**, 145-165.

[54] Linton, O. B. and J. P. Nielsen (1995). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika* **82**, 93–100.

[55] Mas-Colell, A., M. D. Whinston and J. R. Green (1995). *Microeconomic Theory.* Oxford University Press. Oxford.

[56] Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes.* Springer, Heidelberg.

[57] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency, *Journal of Time Series Analysis* **17**, 571–599.

[58] Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models, *Annals of Statistics* **21**, 255–285.

[59] Mammen, E., Linton, O. B. and J. P. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics* **27**, 1443–1490.

[60] Mammen, E., Marron, J. S., Turlach B. A. and M. P. Wand (2001). A general framework for constrained smoothing, *Statistical Science* **16** 232-248.

[61] Mammen, E. and B. U. Park (2005). Bandwidth selection for smooth backfitting in additive models, *Annals of Statistics* **33**, 1260–1294.

[62] Mammen, E. and S. Sperlich (2006). Testing interaction in additive models with smooth backfitting, *Working paper*, University of Mannheim.

[63] Moloche, G. (2001). Local nonparametric estimation of scalar diffusions, *Working paper*, MIT.

[64] Nadaraya, E. A. (1964). On estimating regression, *Theory of Probability and Its Applications* **10**, 186–190.

[65] Newey, W. 1985. Maximum Likelihood specification testing and conditional moment tests, *Econometrica* **53**, 1047–1070.

[66] Nielsen, J. P. and S. Sperlich (2005). Smooth backfitting in practice, *Journal of the Royal Statistical Society B* **67**, 43–61.

[67] Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators, *Journal of Multivariate Analysis* **73**, 166–179.

[68] Opsomer, J. D. and D. Ruppert (1997). Fitting a bivariate additive model by local polynomial regression, *Annals of Statistics* **25**, 186–211.

[69] Parzen, E. (1962). On estimation of a probability density and mode, *Annals of Mathematical Statistics* **35**, 1065–1076.

[70] Powell, J. L., Stock, J. H. and T. M. Stoker (1997). Semiparametric estimation of index coefficients, *Econometrica* **57**, 1403–1430.

[71] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics* **27**, 642–669.

[72] Ruppert, D. and M.P. Wand (1994). Multivariate locally weighted least squares regression, *Annals of Statistics* **22**, 1346–1370.

[73] Schienle, M. (2006). An additive nonparametric approach to a nonlinear multivariate regression of recurrent markov processes, *Working paper*, University of Mannheim.

[74] Sperlich, S., Linton, O. B. and W. Härdle (1999). Integration and backfitting methods in additive models: Finite sample properties and comparison, *Test* **8**, 419–458.

[75] Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate rsik, *Journal of Finance* **52**, 1973–2002.

[76] Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models, *Journal of Econometrics* **30**, 415-443.

[77] Tjøstheim, D. and B. H. Auestad (1994). Nonparametric identification of nonlinear time series: projections, *Journal of the American Statistical Association* **89**, 1398–1409.

[78] Veretennikov, A. (1997). On polynomial mixing bounds for stochastic differential equations, *Stochastic Processes and Their Applications* **70**, 115–127.

[79] Watson, G. S. (1964). Smooth regression analysis, *Sankhya Series A* **26**, 359–372.

[80] Whang, Y. and D. Andrews. 1993. Tests of specifications for parametric and semiparametric models, *Journal of Econometrics* **57**, 277–318.

[81] Wooldridge, J. (1992). A test for functional form against nonparametric alternatives, *Econometric Theory* **4**, 210–230.

[82] Yatchew, A. (1992). Nonparametric regression test based on least squares, *Econometric Theory* **8**, 435–451.

[83] Zheng, C. and H. Dette (2004). A power comparison between nonparametric regression tests, *Statistics and Probability Letters* **66**, 289–301.

[84] Zheng, J. (1996). A Consistent test of functional form via nonparametric estimation techniques, *Journal of Econometrics* **75**, 263–290.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig angefertigt und mich keiner anderen als der in ihr angegebenen Hilfsmittel bedient zu haben. Insbesondere sind sämtliche Entlehnungen aus anderen Schriften als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, 7. Juni 2006

Berthold Haag

Lebenslauf

Persönliche Daten

Berthold Robert Oliver Haag

geboren am 6. Oktober 1975 in Tübingen

Ausbildung

04/2003-07/2006	Promotion in Wirtschaftswissenschaften an der Universität Mannheim.
11/2003	Abschluß als Diplom-Volkswirt.
02/2003	Abschluß als Diplom-Mathematiker.
03/2000–08/2000	Auslandssemester an der Universitá degli Studi la Sapienza in Rom.
04/1997–11/2003	Studium der Mathematik und der Volkswirtschaftslehre an der Ruprecht-Karls-Universität Heidelberg.
10/1996–03/1997	Studium der Mathematik an der Eberhard-Karls-Universität Tübingen.
08/1986–06/1995	Ludwig-Uhland-Gymnasium in Tübingen.