On the Choice of Regularization Parameters in Specification Testing: A critical Discussion

Stefan Sperlich

Département des sciences économiques and Research Center for Statistics

Université de Genève, Bd du Pont d'Arve 40

CH - 1211 Genève, Suisse, stefan.sperlich@unige.ch

October 15, 2012

Abstract

This article reviews and discusses the problem of choosing smoothing parameters and resampling schemes for specification tests in econometrics. While smoothing is used for the regularization of the non-specified parts of the null hypothesis and omnibus alternatives, the resampling serves for determining the critical value. Several of the existing selection methods are discussed, implemented, and compared. This has been done for cross sectional data along the example of additivity testing. Doubtless, all problems considered here carry over to specification testing with dependent data. Intensive simulations illustrate that this is still an open problem that easily corrupts these tests in practice. Possible ways out of the dilemma are proposed.[1]

Keywords: Nonparametric specification tests, adaptive testing, bandwidth choice, bootstrap, subsampling.

JEL code: C12, C14, C52

[1]The author acknowledges very helpful discussion of three anonymous referees and the editors which have improved the paper a lot.

1 Introduction

A decade after the review book of Hart (1997), non- and semiparametric specification testing is still quite a popular research field, especially in econometric theory. Any Internet search engine finds several hundred papers dealing with this topic even when limiting the search to the last five years. While the choice of appropriate smoothing parameters is fundamental for these tests, only a few articles address this issue. The Journal of Econometrics published in 2008 a special issue devoted exclusively to semi- and nonparametric testing. Only in Gao, Gijbels & Van Bellegem (2008) the smoothing parameter choice was explicitly considered – unavoidably, because they considered structural breaks. Moreover, for testing these choice problems are not equivalent to the ones in regression. For example, it is well known that the optimal smoothing parameters for testing have different rates from those which are optimal for estimation, Ibragimov & Khasminski (1981) and Ingster (1982,1993) being the classic references.

In the last couple of years there has been a growing amount of literature on adaptive testing. This "adaptiveness" refers to the unknown smoothness of the alternative and therefore deals with the choice of smoothing parameters for the test statistic, see e.g. Ledwina (1994), Kulasekera & Wang (1997), Spokoiny (1998), Kallenberg & Ledwina (1995), Horowitz & Spokoiny (2001), Guerre & Lavergne (2005). Even though these methods have had little impact in the sense that we could not find published papers using them, they have been useful for a better understanding of the problem. Most of these papers concentrate on testing problems where the null hypothesis is fully parametric. Among others, Gao & Gijbels (2008) extensively discussed the role of bandwidth selection in kernel testing when the null is parametric. In that case it is clear that the testing bandwidth is just a nuisance parameter which may not even be identified when size and power of the test are the main focus. It is not clear to what extend the proposed methods will help if the null hypothesis is semi- or nonparametric. However, this is not that rare; even additivity (or more general "separability") tests belong to this family. Other popular examples are tests for particular covariance structures including independence, tests for symmetry of distributions or substitution matrices, tests for profit maximization, etc.

Where a semi- or nonparametric null hypothesis has to be estimated explicitly, an additional smoothing parameter is needed. In many cases this is chosen by cross validation or simply ad-hoc without further justification. When bootstrap is used to determine the critical value, these tests entail another parameter choice problem: for pre-estimating the model under the null hypothesis to later generate the bootstrap samples. In most cases the bandwidths for estimation and the bootstrap must have different rates, see Härdle & Marron (1991), González-Manteiga, Martínez-Miranda & Pérez-González (2004) or Cao-Abad & González-Manteiga (1993). Although these authors have already mentioned the problem of choosing an appropriate bandwidth, in practical applications this problem has hardly been addressed. As a consequence, in most published procedures for testing or constructing confidence bands with a semi- or nonparametric null hypothesis, there is no guarantee that the test would hold the level, or the bands the nominal coverage probability, see for example Dette, von Lieres und Wilkau & Sperlich (2005) or Rodriguez-Poó, Sperlich & Vieu (2012). However, in the former it is not referred to as a bandwidth problem but rather as a problem of design and dimensionality because the size distortion is much smaller for covariates being independent among each other. In the latter paper the problem is avoided by using subsampling instead of bootstrap. It should be mentioned that there, the authors can use a parametric bootstrap drawing the bootstrap errors from a distribution known up to one parameter. Although that parameter depends on further nonparametric nuisance parameters, the knowledge of distribution greatly mitigates the impact of bandwidths on the estimates of critical values.

To illustrate the outlined problem we concentrate on the testing for additivity in cross sectional data. We limit the presentation to statistics considered in Dette, von Lieres und Wilkau & Sperlich (2005), Rodriguez-Poó, Sperlich & Vieu (2012), and the methods of Horowitz & Spokoiny (2001), Guerre & Lavergne (2005), and Gao & Gijbels (2008). The aim is not to find the most efficient additivity test or to propose new ones. A discussion of what are reasonable test statistics can be found in Roca-Pardiñas & Sperlich (2007). Here, our focus is only directed at a reasonable choice of regularization parameters that hold the level and guarantee non trivial power.

In the next section we review the estimation and testing procedures used for illustration. In Section 3 we discuss the different scenarios from which the practitioner has to make his choice, including modifications of test statistics, and resampling methods. Section 4 summarizes the main findings from our simulations, and Section 5 concludes.

2 Model, Estimators and Test Statistics

Before we introduce estimators and test statistics for a d-dimensional regression problem let us briefly discuss the consequences of implicit conventions or explicit conditions you find in almost all related articles. Often the tests are analyzed in detail only for the univariate case, indicating that the extension to the multivariate one is straight. There, optimal smoothing requires the use of a bandwidth matrix, say $\mathbf{H} \in I\!R^{d \times d}$, which transforms the whole data matrix unless it is restricted to being diagonal. Even when abstracting from practical problems of determining a complete $\mathbf{H}$, such cross-covariates transformations are often undesirable for reasons of interpretability, especially for separable models. Presumably for that reason it is mainly just in the context of density estimation where suggestions for the choice of $\mathbf{H}$ can be found. A most natural transformation is to turn all variables toward orthogonality and to normalize them afterward, see Duong and Hazelton (2003) for details. Intuitively one might conclude that this is also recommendable for multivariate regression, but this can hardly be seen from the asymptotics. Actually, we are not aware of a study that explicitly considers a practical selector for $\mathbf{H}$ in multivariate regression containing off-diagonals. Furthermore, for convenience it is typically assumed that the density of covariates (and the regression function, respectively) has the same smoothness in all directions, and that all covariates have either the same variance or the same support. Having done so, $\mathbf{H}$ gets reduced to $h \cdot \mathbf{I}_d$. In practice one therefore divides each covariate by its standard deviation and then takes the same bandwidth h for all. In order to concentrate on the various selection problems emerging in the testing context we follow this custom, being aware that this is just a makeshift. For certain additive model estimators, some authors proposed methods which allow for different diagonal elements of $\mathbf{H}$.

2.1 Model and Estimation

Consider the following general regression model:

$$Y_i = m\left(X_i\right) + u_i \qquad i = 1, 2, ..., n, \tag{1}$$

with $\{(X_i,Y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. (for simplicity), $m : \mathbb{R}^d \to \mathbb{R}$ the unknown function of interest, $m(x) = E(Y|X = x)$, and u_i random errors with $E[u_i] = E[u_i|x_i] = 0$ and finite variance $\sigma^2(x_i)$. The internalized Nadaraya-Watson estimator is defined as

$$\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\hat{f}_k(X_i)\right)^{-1} \mathbf{K}_k(x - X_i), \tag{2}$$

where $\hat{f}_k(X_i) = \frac{1}{n}\sum_{i=1}^n \mathbf{K}_k(X_j - X_i)$ is a kernel density estimator, $\mathbf{K}_k\{(v_1, \ldots, v_d)\} = \prod_{\delta=1}^d K_k\left(v_\delta\right)$ a product kernel with $K_k(v_\delta) = k^{-1}K\left(v_\delta k^{-1}\right)$ for $\delta = 1, \ldots, d$. For the ease of presentation we follow the common line of applying the same bandwidth k to all covariates neglecting the possibility of using bandwidth matrices. As said, all covariates should then be brought to the same scale or range. Commonly, the kernel is assumed to be Lipschitz continuous with compact support and $\int |K(v)|\, dv < \infty, \int K(v)dv = 1$. Furthermore, k is a bandwidth assumed to tend to zero for sample size n going to infinity, but nk^d going to infinity. Let us call $\hat{m}$ the estimator for the alternative; k can then be considered as the "testing bandwidth".

Assume we are interested in additive modeling. We write

$$E\left(Y|X = x\right) = m_S\left(x\right) = \psi + \sum_{\delta=1}^{d} m_\delta\left(x_\delta\right) \ , \tag{3}$$

setting $E_{X_\delta}\left\{m_\delta(X_\delta)\right\} = \int m_\delta(x) f_\delta(x)dx = 0$ for all covariates X_δ for identification. Here, m_δ, $\delta = 1, \ldots, d$ are the marginal impact functions for each regressor. Therefore, ψ is a constant equal to the unconditional expectation of Y. Writing $m(X) = \psi + m_\delta(X_\delta) + m_{-\delta}(X_{-\delta})$ where $X_{-\delta}$ is the vector X of all explanatory variables without X_δ, i.e. $X_{-\delta} = \left(X_{i1}, \ldots, X_{i(\delta-1)}, X_{i(\delta+1)}, \ldots, X_{id}\right)$ the identification condition directly induces an estimator for m_δ. There exist different motivations to obtain

$$\hat{m}_\delta(x_\delta) = \frac{1}{n}\sum_{i=1}^{n} Y_i K_h\left(x_\delta - X_{i\delta}\right) \frac{\hat{f}_{-\delta}\left(X_{i,-\delta}\right)}{\hat{f}\left(X_{i\delta}, X_{i,-\delta}\right)} \tag{4}$$

like the argument of Kim, Linton and Hengartner (1999), saying that there exist kernel weights $\omega(X_\delta, X_{-\delta})$ so that $E[\omega(X_\delta, X_{-\delta})Y|X_\delta = x_\delta] = \psi + m_\delta(x_\delta)$. Finally, with $\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ one sets $\hat{m}_S(X_j) = \hat{\psi} + \sum_{\delta=1}^{d} \hat{m}_\delta(X_{j\delta})$ for each $j = 1, 2, ..., n$. The densities in (4) are kernel densities with bandwidths h for variable X_δ and $h_{-\delta}$ else, see Dette el al (2005) or Hengartner & Sperlich (2005) for details. In the following, (4) will be the estimator for the null hypothesis with bandwidth h (and $h_{-\delta}$, respectively). As for k, we neglect bandwidth matrices and use the same bandwidth for all covariates supposing they have comparable distributions, recall our discussion from above.

For completeness we conclude this subsection with a brief, non-complete review of available alternative estimators for nonparametric additive models. The probably first key reference is the classical backfitting estimator of Hastie& Tibshirani (1990). The lack of asymptotic theory for this estimator led to alternative proposals like the marginal integration estimator of Linton and Nielsen (1995), and its improved version of Kim et al (1999) which we are using here. Comparison studies were provided by Sperlich, Linton & Härdle (1999) and Dette et al (2005). About the same time appeared the smooth backfitting estimator of Mammen, Linton & Nielsen (1999). Today, penalized spline versions of the classical backfitting are very popular in practice due to its easy implementation, see Fahrmeir, Kneib & Lang (2004) and Wood (2008).

2.2 Test Statistics

The null hypothesis of interest is $H_0 : m(\cdot) = m_S(\cdot)$ versus $H_1 : m(\cdot) \neq m_S(\cdot)$. We do not aim to introduce new testing procedures but rather take statistics which have already been studied in Dette et al. (2005) together with other additivity tests, motivated by Rodríguez-Poó et al. (2004), and which performed excellently in the study of Roca-Pardiñas & Sperlich (2007). Finally, we add a new test for additivity which is based on the statistic introduced in Guerre & Lavergne (2005). For more details about

the tests the reader is referred to these papers. Consider the test statistics

$$\begin{aligned}\tau_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-\hat{m}_S\left(X_i\right)\right)^2 w(X_i)\ ,\\ \tau_2 &= \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i\left(\hat{m}\left(X_i\right)-\hat{m}_S\left(X_i\right)\right) w(X_i)\ ,\end{aligned}$$

where $\hat{e}_i = Y_i - \hat{m}_S\left(X_i\right)$, i.e. the residuals under the null hypothesis, and $w(\cdot)$ an optional weight function that might be used for trimming. While τ_1 calculates directly the integrated squared difference between the null and alternative models, τ_2 seeks to mitigate the bias problem inherited from the estimate $\hat{m}$ which suffers from the curse of dimensionality. While in τ_1 this large bias enters squared, in τ_2 it only appears in simple terms multiplied by residuals which under H_0 should even be independent of the smoothing bias. Therefore, using leave-(Y_i, X_i)-out estimators for $\hat{m}_S(X_i)$ and $\hat{m}(X_i)$ in τ_2 would give $E_{H_0}\{\tau_2\} = 0$ asymptotically, see Gozalo & Linton (2001). Else, in Dette et al. (2005) it has been proved that each $nk^{\frac{d}{2}}\left(\tau_j - \mu_j\right)$ converges under the null to a normal with mean zero and variance v_j^2, where

$$\begin{aligned}v_1^2 &= Var_{H_0}\left\{\tau_1\right\} = 2\int\sigma^4(x)w^2(x)dx\int\left(\mathbf{K}*\mathbf{K}\right)^2(x)dx\ ,\\ v_2^2 &= Var_{H_0}\left\{\tau_2\right\} = 2\int\sigma^4(x)w^2(x)dx\int\mathbf{K}^2(x)dx\ ,\\ \mu_1 &= E_{H_0}\left\{\tau_1\right\} = \frac{1}{nk^d}\int\sigma^2(x)w(x)dx\int\mathbf{K}^2(x)dx + o\left(\frac{1}{nk^d}\right)\ ,\\ \mu_2 &= E_{H_0}\left\{\tau_2\right\} = \frac{1}{nk^d}\int\sigma^2(x)w(x)dx\ \mathbf{K}(0) + o\left(\frac{1}{nk^d}\right)\ .\end{aligned}$$

The next statistic was defined to avoid the calculation of high dimensional $\hat{m}$:

$$\tau_3 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{K}_k\left(X_i - X_j\right)\hat{e}_j\right]^2 w(X_i)\ , \tag{5}$$

where for ease of presentation and implementation $\mathbf{K}\left(\cdot\right)$ is the same kernel function as in the last subsection, and k again its bandwidth. It is straightforward to derive from the above mentioned paper that $nk^{\frac{d}{2}}\left(\tau_3 - \mu_3\right)$ converges under the null to a normal variable with mean zero and variance v_3^2, where now

$$\begin{aligned}\mu_3 &= E_{H_0}\left\{\tau_3\right\} = \frac{1}{nk^d}\int\left(\mathbf{K}*\mathbf{K}\right)^2(x)\,dx\int\sigma^2(x)f^2(x)w(x)dx\ ,\\ v_3^2 &= Var_{H_0}\left\{\tau_3\right\} = \int\sigma^4(x)f^4(x)w^2(x)\ .\end{aligned}$$

As the skedasticity functions $\sigma^2(x)$ are typically large where the density of x is small, one obtains an efficient though natural weighting in the bias and variance term.

Finally, Guerre & Lavergne (2005) constructed a statistic which is particularly useful when applying their bandwidth choice method. It also avoids the estimation of $\hat{m}$ but requires the nonparametric estimation of the high dimensional densities of x. It is useful to introduce here the notation of τ_4 as a function of bandwidth k :

$$\begin{aligned} \tau_4(k) &= \frac{1}{n-1} \sum_{i \neq j} \{Y_i - \hat{m}_S(X_i)\} \frac{\mathbf{K}_k(X_i - X_j)}{\sqrt{\hat{f}_k(X_i)\hat{f}_k(X_j)}} \{Y_j - \hat{m}_S(X_j)\} \\ &= \frac{1}{n-1} \sum_{i \neq j} \hat{e}_i w_{ij}(k) \hat{e}_j \ , \end{aligned} \tag{6}$$

where $\hat{e}_i$ are the residuals under H_0, and $w_{ij}(k)$ the smoothing weights of the statistic. Later on, for the bandwidth choice procedure of Guerre & Lavergne (2005) we need the variances of τ_4 and $\tau_4(k) - \tau_4(k_P)$, i.e. the difference of two statistics calculated for different bandwidths k and k_P respectively. To emphasize its dependency on k and k_P, we write them here explicitly as functions of the smoothing weights $w_{ij}(k)$, i.e.

$$\begin{aligned} v_4^2(k) &= Var_{H_0}\{\tau_4(k)\} = 2 \sum_{i,j} \sigma^2(x_i) w_{ij}^2(k) \sigma^2(x_j) \ , \\ v_{4b}^2(k, k_P) &= Var_{H_0}\{\tau_4(k) - \tau_4(k_P)\} = 2 \sum_{i,j} \sigma^2(x_i) \{w_{ij}(k) - w_{ij}(k_P)\}^2 \sigma^2(x_j) dx \ . \end{aligned}$$

This will help us later to understand better the idea of the procedure for selecting k.

All tests have been proved to be consistent. We also studied other test statistics, for example all those given in Dette et al. (2005) but have not presented them here because they showed even less satisfactory performance.

2.3 Calculating the critical value by Resampling

Asymptotics are her of little help to calculate the critical value as bias and variance contain unknown expressions which have to be estimated nonparametrically, and as the convergence rate to them and the normal distribution is slow. Therefore it is common to use resampling (bootstrap or subsampling) methods to approximate the critical value for the particular sample statistic. The commonly used bootstrap procedure is:

1. With bandwidth h, calculate the estimate $\hat{m}_S$ under the null hypothesis of additivity and its resulting residuals $\hat{e}_i$, $i = 1, \ldots, n$.

2. With bandwidth k, calculate the estimator $\hat{m}$ (or the respective kernel expressions of the test statistic involving k) for the conditional expectation without the additivity restriction, and the corresponding residuals $\hat{u}_i$, $i = 1, \ldots, n$.

3. With the results from step 1 and 2 we can calculate our test statistics τ_j.

4. Repeat step 1 but now with a bandwidth h_b which depends on h from step 1. We call the outcome $\hat{m}_{S,h_b}$, set $\epsilon_i = Y_i - \hat{m}_{S,h_b}(X_i)$, $i = 1, \ldots, n$. Draw random variables e_i^* with $E[(e_i^*)^l] = (y_i - \hat{m}(x_i))^l =: u_i^l$ (or $\hat{e}_i^l$, or even ϵ_i^l, see discussion below) for $l = 2, 3$ (respectively $l = 2$, see below again). Set $Y_i^* = \hat{m}_{S,h_b}(X_i) + e_i^*$, $i = 1, \ldots, n$. Repeat this B times. This defines B different bootstrap samples $\{(X_i, Y_{i,b}^*)\}_{i=1}^n$, $b = 1, \ldots, B$.

5. For each bootstrap sample from step 4 calculate test statistic $\tau_{j,b}^*$, $j = 1, 2, 3, 4$, $b = 1, \ldots, B$. Then, for each test statistic τ_j the critical value is approximated by the corresponding quantiles of the distribution of the B bootstrap analogues: $F^*(v) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\tau_{j,b}^* \leq v\}$. Note that they are generated under H_0.

This procedure is well known, has been proved to be consistent for many tests and has therefore been applied, with slight modifications, to many non- or semiparametric testing problems. Unfortunately, it is not known how to choose the smoothing parameter h_b in practice for the pre-estimation of the model that is used to generate the bootstrap samples. Note that in finite samples the correct choice will depend not only on h but also on the choice of k, and the used residuals (see step 4.).

A more and more popular alternative to bootstrapping is the subsampling procedure, see Politis, Romano & Wolf (1999). To date, as subsampling is commonly believed to converge slower than bootstrapping, it has been used almost exclusively when bootstrap fails, see Neumeyer & Sperlich (2006) for a purely nonparametric testing context. They also study the automatic choice of subsample size N which turned out to work in their simulations. As this method can be remodeled to serve as a procedure for finding h_b, we introduce subsampling together with the automatic choice of the subsample size.

Let $\mathcal{Y} = \{(X_i, Y_i) \,|i = 1, ..., n\}$ be the original sample and denoted by $\tau_{\mathcal{Y}}(k)$ the original statistic calculated from this sample, leaving aside the index $j = 1, 2, 3, 4$. To determine the critical values we need to approximate

$$Q\left(z\right) \;=\; P\left(n\sqrt{k^d}\tau_{\mathcal{Y}}(k) \leq z\right) \;. \tag{7}$$

Recall that under H_0 this distribution converges to an $N(\mu, v^2)$, for μ and v given above. For finite sample size n, drawing B subsamples $\mathcal{Y}_b$ - each of size N - we can approximate Q under H_0 by

$$\hat{Q}\left(z\right) \;:=\; \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(N\sqrt{k_N^d}\tau_{\mathcal{Y}_b}(k_N) \leq z\right) \;. \tag{8}$$

Note that the awkward notation comes from the fact that we have to adjust all bandwidths to the new sample size N. For example, imagine $k = c_0 \cdot n^{-\rho}$ for c_0 being a constant. Then, the subsample test $\tau(k_N)$ is calculated like τ but with bandwidth $k_N = c_0 \cdot N^{-\rho}$. Certainly, under the alternative H_1, not only $n\sqrt{k^d}\tau_{\mathcal{Y}}(k)$ but also $N\sqrt{k_N^d}\tau_{\mathcal{Y}_b}(k_N)$ converges to infinity. Then, demanding $N/n \to 0$ guarantees that $n\sqrt{k^d}\tau_{\mathcal{Y}}(k)$ converges faster to infinity than the subsample analogues. Finally, $\hat{Q}$ underestimates the quantiles of Q which yields the rejection of H_0.

Actually, the optimal N is a function of the nominal level α. In order to find that, we now test a null hypothesis of which we know that it is true but suffers from the same smoothing bias as our original testing problem does. Such a null hypothesis is for example $H_0^* :\; m(x) - m_S(x) = \hat{m}(x) - \hat{m}_S(x)$ which is to be tested by an analogous statistic one uses to test the original problem $H_0 :\; m(x) = m_S(x)$. For the desired level α, apply that statistic to H_0^* and use subsampling to determine its p-value. Now draw some pseudo sequences $\mathcal{Y}_l^*$, $l = 1, \ldots, L$ from $\mathcal{Y}$ of size n with the same distribution as $\mathcal{Y}$, and repeat this test L times. From these repetitions you can determine the empirical rejection level (estimated size) for the given α. As you know that H_0^* is true and therefore should be rejected exactly in 100α percent of all cases, you look for a subsample size N producing this rejection level. In practice you choose from a grid of potential N the one whose rejection level comes closest to α from below. For further details see Politis et al (1999), Delgado, Rodriguez-Poó & Wolf (2001) or Neumeyer & Sperlich (2006). A drawback of this procedure are the computational costs.

3 The Choice of Parameters

The above introduced statistics and procedures raise many questions of practical importance which typically remain open: bandwidth choice h and $h_{-\delta}$ in step 1, bandwidth choice k in step 2, how to generate the bootstrap residuals e_i^* in step 4, and how to choose h_b. Finally, how many bootstrap samples are necessary to get a reasonable approximation of the distribution in step 5. We will discuss all these questions except the last one, giving mainly space to the discussions that are testing specific.

Recall that the bandwidth choice problem in estimation is different from that in testing. Testing additivity is just an illustrative example here for a nonparametric testing problem with a semi- or nonparametric null hypothesis. We are generally interested in nonparametric specification tests where both the regression model and the testing problem could equally well have nothing to do with additivity but for example with distribution assumptions, the link function, variable selection, heteroscedasticity, autocorrelation, endogeneity, jump detection, etc. The article is aimed to discuss the resulting bandwidth choice problems and its impact on the test performance.

3.1 The Choice of Bandwidths h

The problem of finding an optimal h is somewhat different from that of finding the optimal smoothing parameter k which is directly linked to the optimal rate of the test statistic. In the latter case it is clear that a theoretical optimal choice depends on the optimal rate at which the test can detect a deviation from the null hypothesis, see the next subsection. In most cases, the estimator of the null model can have faster convergence rates than that of the alternative, so the asymptotics of the test statistics provide no theoretical guideline for an optimal choice of h (or $h_{-\delta}$). To date we have to rely on practical issues; but an optimal bandwidth choice for the null model in nonparametric testing is a potential topic for further investigation, cf. Section (4).

As there exist data adaptive methods for finding the optimal bandwidth k for the alternative (next subsection) one could argue that h should be chosen accordingly

to k. This way one would guarantee that the same smoothness is imposed on the regression function regardless of whether it is estimated under the null hypothesis or not. However, it is not clear whether this is always wanted. Moreover, we will see that the adaptive choice of k is computationally intensive and that also h_b (cf. Section 2.3) depends on h. So each bandwidth selection would depend on each other, and it is unclear where to start. Apart from the fact that we could not find any suggestion in the literature pointing in this direction, this would lead to a computationally quite complex procedure. Intuitively, it seems to be desirable to look for a reasonable estimation of the null model. This is only guaranteed with a reasonable bandwidth choice of h beforehand. All in all, given the state of the art we recommend to follow this intuitive argument and use cross validation (cv henceforth) or plug-in methods, see Vieu (1993) for a classic survey, and Köhler, Schindler & Sperlich (2012) for a recent review on bandwidth selection methods in kernel regression.

For internalized marginal integration estimators no explicit bandwidth selection procedure is available. Here, we follow the cross validation idea. Recall that in practice it is often preferred to simply divide all explanatory variables by its standard deviation and then use the same bandwidth for all directions. On the one hand, this is certainly suboptimal from a strictly statistical point of view, but on the other hand this follows the idea of comparable smoothness in all directions. In such a case we have only two parameters to choose: h and $h_{-\delta}$. This can be done by minimizing the cv criterion

$$CV(h, h_{-\delta}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \hat{m}_S^-(X_i) \right\}^2 \ , \tag{9}$$

where $\hat{m}_S^-(X_i) = \hat{\psi} + \sum_{\delta=1}^{d} \hat{m}_\delta^-(X_{j\delta})$ with

$$\hat{m}_\delta^-(X_{j\delta}) = \frac{1}{n-1} \sum_{i \neq j}^{n} Y_i K_h \left(X_{j\delta} - X_{i\delta} \right) \frac{\hat{f}_{-\delta}\left(X_{i,-\delta} \right)}{\hat{f}\left(X_{i\delta}, X_{i,-\delta} \right)} \ . \tag{10}$$

As explained in Kim et al (1999) the fraction of estimated densities is a weighting necessary to avoid an omitted-variable bias. The leave-out or not leave-out of X_j has a minor impact in that fraction, so that the definition of the cv criterion via (9) and (10) constitutes an acceptable simplification of the classical cross validation, like it has already been proposed by Nielsen & Sperlich (2005) in the context of smoothed

backfitting. So we minimize here the average mean squared error for the null model. This does not necessarily give the bandwidths to minimize the estimation errors of the individual additive components; but recall that all test statistics look for the prediction (or fitting) power of the total regression.

Further discussion of potential bandwidth selectors for this estimator will not help for a deeper understanding of the considered testing problem. However, since we already reviewed alternative estimators for nonparametric additive models, we should conclude here giving a brief review for choosing the corresponding smoothing parameters. Hastie and Tibshirani (1990) proposed cross validation and generalized cv criteria. Linton and Nielsen (1995) and Severance-Lossin and Sperlich (1999) proposed different plug-in estimators for the original marginal integration. As said, for its internalized version we are not aware of any specific proposal. Mammen and Park (2005) as well as Nielsen & Sperlich (2005) introduced several plug-in and cv methods for the smooth backfitting estimator. For the classical backfitting with splines often the so-called generalized cv is recommended. A more detailed discussion but for the Bayesian approach can be found in Fahrmeir, Kneib & Konrath (2010). We stop here also because for each specific null hypothesis one has a specific bandwidth choice problem; additivity is only an example.

3.2 The Choice of Bandwidths k

It is known that a bandwidth k which is optimal for estimation is usually suboptimal for testing. More specifically, the optimal smoothing parameter for testing has faster convergence rates; we actually should undersmooth. As cv bandwidths have a tendency to undersmooth in practice, they are quite popular in nonparametric testing.

As an alternative, let us consider the adaptive testing approach introduced in Spokoiny (1998). The (where necessary modified and adapted) method is the same for each of our four test statistics, so we skip the index j of τ_j, $j = 1, 2, 3, 4$ for a moment but use the notation $\tau(k)$ indicating the testing bandwidth applied. Adapted to our problem it works as follows. Consider simultaneously a family of tests $\{\tau(k),\, k \in \mathfrak{K}\}$, where $\mathfrak{K} = \{k_1, k_{2,\ldots,}, k_P\}$ is a finite set of reasonable bandwidths. The theoretical maximal

number P depends on n; for the sequences $\{k_j\}_{j=1}^{P}$ see the particular paper.

Horowitz & Spokoiny (2001) proposed to look at

$$\tau^{max} = \max_{k \in \mathfrak{K}} \frac{\tau(k) - E_{H_0}[\tau(k)]}{Var^{1/2}[\tau(k)]} \ , \text{ where} \tag{11}$$

$E_{H_0}[\cdot]$ indicates the expectation under H_0, and the variance has to be estimated with an estimator of σ^2 that is consistent under H_1. This studentizing under the null is to correct for the deviations in distribution caused by the different bandwidths k. Instead of $Var^{1/2}[\tau(k)]$ one could take something proportional to it without loosing consistency (though asymptotic efficiency) as long as it corrects for the standard deviation caused by the different $k \in \{k_1, \ldots, k_P\}$. We will call this method HS when explicitly estimating the required moments, and k_o denotes the bandwidth giving τ^{max}.

Guerre & Lavergne (2005) suggested a procedure to select k for τ_4 (in a simpler testing context) of which they stated it would select a bandwidth even more tailored for testing. They presupposed to dispose of centered (under H_0) test statistics - therefore we will take again $\{\tau_4(k) - E_{H_0}[\tau_4(k)]\}$. Instead of dividing by its standard deviation, they selected k along the criterion

$$k_o = arg \ \max_{k \in \mathfrak{K}} \left\{ \tau_4(k) - \widehat{E}_{H_0}[\tau_4(k)] - \kappa_n \hat{v}_{4b}(k, k_P) \right\}, \ \kappa_n = 2\sqrt{2 \ln P}. \tag{12}$$

In our context k_P is the largest bandwidth in $\mathfrak{K}$. Their final test statistic was $\tau_4(k_o)/v_4(k_P)$. Let us denote this method by GL in the following.

A quite different approach was proposed by Gao & Gijbels (2008). For testing a fully parametric null hypothesis they considered a simplified version of τ_4 and calculated the Edgeworth expansion of its size and power function. Then, for the obtained expressions they chose bandwidth k such that it maximized the power while holding the size. These expansions and its bootstrap approximates are subject to serious changes when the null hypothesis becomes non- or semiparametric. Although this has not yet been analyzed in detail, it is clear that the approximation will not only change importantly but also be much less reliable, e.g. concerning the distribution. Especially the size function - as it depends on the (estimated) mull model - will be affected by the smoothing and bootstrap bias. Summarizing, it is unclear how the semiparametric analogue to this

solution would look like, and it is unlikely that this approximation would work well until the samples size is huge.

In contrast, it is obvious how to extend the methods of Horowitz & Spokoiny (2001) or the one of Guerre & Lavergne (2005) to problems with semi- or nonparametric null hypotheses, see Rodríguez-Poó et al (2004). A particularity of the bootstrap analogues of $\tau^{max} = \tau(k_o)$ is, that one first needs to calculate the bootstrap statistics $\tau_b^*(k)$ for all $k \in \mathfrak{K}$ to get $(\tau^{max})_b^*$ as this is not necessarily equal to $\tau_b^*(k_o)$ when k_o refers to the bandwidth that maximizes only the original statistic. In other words, for each bootstrap sample this bandwidth can be different. In fact, the bootstrap tries to simulate an extreme value distribution. The empirical moments of the bootstrap statistics $\tau_b^*(k)$ can be used to estimate $E_{H_0}[\tau(k)]$, respectively $Var^{1/2}[\tau(k)]$, in practice. This is what we will try and study in our simulations for τ^{max} in (11), called HSB as it refers to the purely bootstrap based HS version. The direct estimation of these expectations and variances is often quite hard if not infeasible.

3.3 The Choice of Bootstrap Residuals

From a theoretical point of view, wild bootstrap errors should be drawn from the residuals of the alternative model, i.e. u_i should be used in Section 2.3, step 4. Clearly this would maximize the power, as the variance of ϵ_i (or $\hat{e}_i$) can increase a lot with increasing distance between H_0 and the true model. For consistency it is sufficient to have this variance bounded, but it is evident that this phenomena deteriorates the power of the test. In the study of Dette et al (2005) the power loss was negligible when using residuals from the null instead of taking residuals from the alternative.

Arguments in favor of using ϵ_i exist only under other, mainly practical aspects: often the size distortion in bootstrap tests is worse when using u_i or $\hat{e}_i$. Moreover, when using adaptive procedures (i.e. the statistic is evaluated over a range of testing bandwidths k) as described in Subsection 3.2, then it is not that clear which of the $u_i(k)$ to use or whether the u_i should be estimated independently of the k-choice for the test. So the size argument seems to be decisive though we admit that, if no adaptive choice of k is

made, it would be desirable to use u_i as long as one can control for the size distortion.

The second question is what kind of distribution should be used for generating the random errors. In step 4 of the bootstrap procedure described in Subsection 2.3 a distribution is commonly taken that gives e_i^* with $E[(e_i^*)^l] = \epsilon_i^l$ for $l = 2$ up to 3 (or even more). The so-called golden-cut wild bootstrap is quite popular, see e.g. Härdle & Mammen (1993). More recently, in the context of size distortion of bootstrap tests, Davidson & Flachaire (2008) argued that for problems with moderate sample size the disadvantages of the higher-order-moment adapting bootstraps outweigh their asymptotic advantages. To check this, we compare different methods in our simulations.

3.4 The Choice of Bootstrap Bandwidth h_b

In general, for many test statistics one knows that the mean of $\hat{m}_h(x) - m(x)$ under the conditional distribution of $Y_1, ..., Y_n | X_1, ..., X_n$, respectively of $\hat{m}_h^*(x) - \hat{m}_{h_b}(x)$ under the conditional distribution of $Y_1^*, ..., Y_n^* | X_1, ..., X_n$, is

$$E_{Y|X}(\hat{m}_h(x) - m(x)) \approx h^2 \frac{\mu(K)}{2} m''(x) \ , \tag{13}$$

$$E^*(\hat{m}_h^*(x) - \hat{m}_{h_b}(x)) \approx h^2 \frac{\mu(K)}{2} \hat{m}_{h_b}''(x) \ , \tag{14}$$

where $\mu(K) = \int u^2 K(u) du$, see Rosenblatt (1969). Then, to guarantee that (13) is well approximated by the bootstrap version (14) we need that $\{\hat{m}_{h_b}''(x) - m''(x)\} \to 0$. The optimal bandwidth h_b for estimating the second derivative is known to be much larger (in rates) than the optimal h for estimating the function itself. We can even give the optimal rate. For example, the optimal rate to estimate m_S'' is of order $n^{-1/9}$ (instead of $n^{-1/5}$). This gives the optimal bootstrap bandwidth to construct uniform confidence bands, see Härdle & Marron (1991) and Cao-Abad & González-Manteiga (1993); compare also with González-Manteiga, Martínez-Miranda & Pérez-González (2004) for the $d > 1$ dimensional case. They give an explicit plug-in formula for the optimal pre-smoothing bandwidth but in a different context and containing many unknown expressions. In our simulation study we will come back to this optimal rate for h_b. However, it will also be seen that the typical comment h_b *has to be oversmoothing*

compared to h is not very helpful; neither taking any $h_b > h$ nor setting $h_b = h \cdot n^{\frac{1}{5}-\frac{1}{9}}$ will guarantee a size$\approx$ α with non-trivial power. In fact, the proper choice of h_b has even not been solved appropriately if we just use it for constructing confidence bands in estimation, see Neumeyer & Polzehl (1998) or Claeskens & Van Keilegom (2003). Typically it is recommended an undersmoothing h or interpreting the band as a band around the estimator, not the function. Then one could set $h = h_b$, at least in the univariate case $d = 1$. In testing, this choice problem is unfortunately more involved. Our simulations show very well that one has a serious size problem due to the smoothing bias in the bootstrap samples. There exist econometric papers that do simulations under H_0 with large bandwidths, and under H_1 with small bandwidths – replicating their simulations one can see that the large bandwidths cause trivial power under H_1 and the small bandwidths reject under H_0 much too often.

The practitioner needs a clear guideline how to choose h_b which we unfortunately cannot find in the literature. A proper and detailed development of such a selection strategy is beyond the scope of this paper; we rather intend to review and highlight the existing but commonly ignored problems in nonparametric specification testing in econometrics. Only in the note of Barrientos & Sperlich (2007) it was suggested to apply the same idea used for the automatic choice of a proper subsample size N in subsampling (cf. Section 2.3) but has so far not really been studied. This will be tried and analyzed in our simulation study as it is the only hint we found in the literature.

More specifically: imagine h_b as a function of α, and for simplification fix also k. The procedure can be extended to GL-, HS- and HSB- data adaptive choices of k.

1. For a given nominal level α consider the testing problem $H_0^* : \ m(x) - m_S(x) = \hat{m}(x) - \hat{m}_S(x)$ which will be tested with an analogous statistic one wants to use for the original problem $H_0 : \ m(x) = m_S(x)$.

2. Draw some pseudo sequences $\mathcal{Y}_l^*$, $l = 1, \ldots, L \geq 100$ from $\mathcal{Y}$ of size n with the same distribution as $\mathcal{Y}$.

3. Take h_b from a bandwidth grid and do a bootstrap test to check H_0^* for all the L samples $\mathcal{Y}_l^*$ generated above.

4. Let α^* be the percentage of rejecting H_0^* in the L samples. If $\alpha^* = \alpha$, then you've found the appropriate h_b, else go back to step 3 to try with a different h_b.

5. In practice, none of the h_b of the grid will produce $\alpha^* = \alpha$ exactly. Then take for the original test the bandwidth h_b which minimizes $(\alpha - \alpha^*)$ for $\alpha^* \leq \alpha$.

4 Simulation Results

To study all the points listed in the last section, we performed a comprehensive simulation study. We give here only a summary; for example, limiting the presentation to $w(x) \equiv 1$ for τ_j, $j = 1, 2, 3, 4$, one particular model, one specific (random) design, and sample size n to 150. We draw i.i.d. three dimensional explanatory variables

$$X_i \sim N(0, \Sigma_X) \quad \text{with} \quad \Sigma_X = \begin{pmatrix} 1 & 0.2 & 0.4 \\ 0.2 & 1 & 0.6 \\ 0.4 & 0.6 & 1 \end{pmatrix},$$

and i.i.d. error terms $e_i \sim N(0, \sigma_e^2)$ to generate

$$Y_i = X_{1,i} + X_{2,i}^2 + 2\ \sin(\pi X_{3,i}) + aX_{2,i}X_{3,i} + e_i, \quad i = 1, \ldots, n$$

with $a = 0$ to generate an additive separable model, and $a = 2$ for the alternative. The null hypothesis is additivity. Unless otherwise indicated, $\sigma_e = 1$. For the unrealistic situation where Σ_X is the identity matrix (i.e. with covariates being independent from each other) the problem is greatly simplified. A much stronger correlated design than ours leads already to identification problems for moderate sample sizes.

All results in the tables are calculated from 1000 replications using 250 bootstrap samples or subsamples, respectively. For real data applications 250 bootstrap samples are certainly very few; but in our simulations the results differed little when we increased the number to 500. We used the multiplicative quartic kernel throughout.

When we used the weighting function $w(\cdot)$ for trimming to get rid of the boundary effects, we certainly got somewhat different numerical results. The relative findings -

may it be size problems or the ranking of the tests by power - did not change. Therefore we present here only results without trimming or other additional weighting.

We start with the bandwidth choice for estimating the null model. Let the parameter responsible for the size, h_b, depend on both α (the level), k, and h but not vice verse. Then it is no problem when h is chosen by cv in each simulation run as proposed in Section 3.1. For the nuisance directions $X_{-\delta}$, see equation (4), it is known that a much larger bandwidth $h_{-\delta}$ can (cf. Hengartner & Sperlich, 2005) and has to be used (cf. Dette et al., 2005), always supposing that these covariates have smooth densities. We tried therefore with setting $h_{-\delta} = c \cdot h$ where $c \in \{3, 4, 5, 6, 7, 8\}$, based on the recommendations of the above quoted papers. The optimum seems to be between 5 and 6 when choosing $h_{-\delta}$. But now recall our discussion about the fact that minimizing the average mean squared error for estimation might not be the optimal choice for optimizing the test performance. That already the smoothing parameter choice for the null model can have an important impact on the test performance will easily be seen when in the following we always contrast the choices $h_{-\delta} = 5 \cdot h$ with $h_{-\delta} = 6 \cdot h$.

We also tried different bootstrap residuals (cf. Section 2.3). Our simulations mainly seem to confirm the arguments discussed above: The power loss caused by not using residuals from the alternative was negligible; so the size problem was decisive for our final decision. Below we report only results referring to $e_i^* = \varepsilon_i \cdot \epsilon_i$, where the ε_i are i.i.d. $N(0, 1)$. When they were drawn from the golden-cut distribution

$$\varepsilon_i = \begin{cases} -(\sqrt{5}+1)/2 & \text{with probability} \quad p = (\sqrt{5}+1)/(2\sqrt{5}) \\ (\sqrt{5}+1)/2 & \text{with probability} \quad 1-p \end{cases},$$

all our results became much worse (less precise sizes and less power). However, we admit that it might be interesting to study in new the effect of which residuals to take (i.e. u_i, $\hat{e}_i$ or ϵ_i) when trying different choice procedures for h_b.

For the bandwidth of the test, recall Section 3.2, we let k run over a grid of $P = 10$ bandwidths from $k_{max} = k_P = 6$, the range of the support of each covariate, to $k_{min} = k_1 = 3/n^{1/d}$, i.e. such that the approximate X support $[-3, 3]^d$ can be covered by n cubes of volume $(2k_{min})^d$, which is the support of each multiplicative quartic kernel. For τ_j, $j = 1, \ldots, 4$ we first study the results for all k and compare them with τ_j^{max}(HSB),

$j = 1, \ldots, 4$. We verified that the τ_j^{max} did not take values at the boundaries k_{min} or k_{max}. We speak of "adaptive" tests when referring to the τ_j^{max} HS or HSB method, or the automatic choice of k along GL. In these cases we call the chosen bandwidth k_o. Non-adaptive procedures certainly vary over the range from k_1 to k_P.

Another challenging point is the choice of h_b. We first give results obtained for particular h_b and let k run. In a second step we do it vice versa. To choose h_b as a function of h and validate the "oversmoothing" argument including the optimal rate $n^{-1/9}$ (recall discussion and justification in Section 3.4) we set

$$h_b = h \cdot n^{1/5-1/\gamma} \ , \text{ for } \gamma \in \{4, 5, 6, 7, 8, 9, 10\}. \tag{15}$$

Supposing that our cv bandwidth follows $h \propto n^{-1/5}$ one starts with an $h_b < h$ for $\gamma = 4$ and goes up to $h_b \propto n^{-1/10}$ including the optimal rate $n^{-1/9}$. As h_b is chosen data adaptively and thus changes for each sample, we report always the pre-fixed γ.

Clearly, since we calculated simulation results for all combinations of bandwidths, tests statistics, selection and resampling methods, its presentation would generate a serious data and information overflow. So we have decided to condense the presentation of numerical results in the following way. First, we concentrate on one bootstrap method as described above. Second, we only present the rejection levels for the nominal size of $\alpha = 5\%$ (error of the first kind) and some p-values. Third, recall the further settings mentioned at the beginning of this section.

Figures 1 (for $h_{-\delta} = 5h$) and 2 (for $h_{-\delta} = 6h$) show the real size and rejection levels for nominal size $\alpha = 5\%$ over different test bandwidths k for given $\gamma = 5$ and 9 respectively, i.e. fixing the rate of (over-)smoothing for the pre-estimation to later generate the bootstrap samples. The results corresponding to k_o refer to the adaptive tests $\tau_j^{max}(HSB)$, cf. Section 3.2. It can be seen how power and size problems vary over the range of test bandwidth k. At least for the presented $\gamma = 5$ and 9 test τ_4 and τ_1 have serious size problems, τ_2 is much too conservative, and even for τ_3 the proper choice of k seems to be crucial. Finally, τ_1 and τ_2 exhibit quite poor power. Interesting is also the impact of the bandwidth for estimating the null hypothesis, which in our case is only identified via the variation of $h_{-\delta}$. All graphs in Figure 2 have basically the

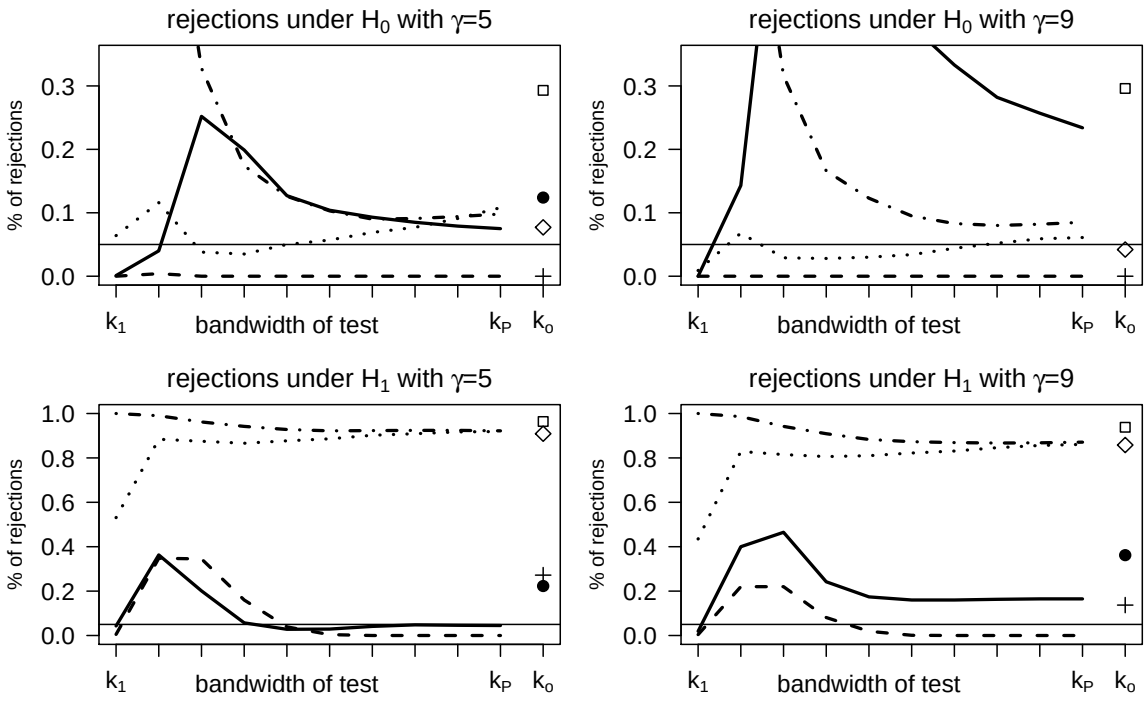

Figure 1: When $h_{-\delta} = 5h$: Real sizes (upper line) and rejection levels (lower line) over k, where k_o stands for adaptive choice referring to the bootstrap studentized (HSB) τ_j^{max}. Nominal size was $\alpha = 5\%$ (thin line). τ_1: solid line; τ_1^{max}: bullet (out of range in the upper right graph); τ_2: dashed line; τ_2^{max}: plus; τ_3: dotted line; τ_3^{max}: diamond; τ_4: dots and dashes; τ_4^{max}: square.

same shape like in Figure 1 but on a different scale. Similar statements hold for the versions with HSB adaptive choices of k. Note that we tried some more test statistics but found all others to perform worse.

Next, in Figures 3 (for $h_{-\delta} = 5h$) and 4 (for $h_{-\delta} = 6h$) are given the proportions of rejections for a nominal level $\alpha = 5\%$ under H_0 and H_1 over the range of $\gamma = 4$ to $\gamma = 9$. We first set k equal to $k_6 := k_1 + 5(k_P - k_1)/(P - 1)$, then repeated the comparison with the HSB adaptive choice k_o, i.e. for $\tau_j^{max}(HSB)$, $j = 1, \ldots, 4$. These results basically show that the problem is not simply solved by different smoothing in the pre-estimation. Oversmoothing, as generally recommended from a theoretical point of view, seems even to go into the wrong direction for some statistics. In particular, the hope that the intuition (see equations (13) and (14)) might give us a hint or even provide a rule of thumb for the choice of h_b, is not confirmed. Again, τ_3 outperforms

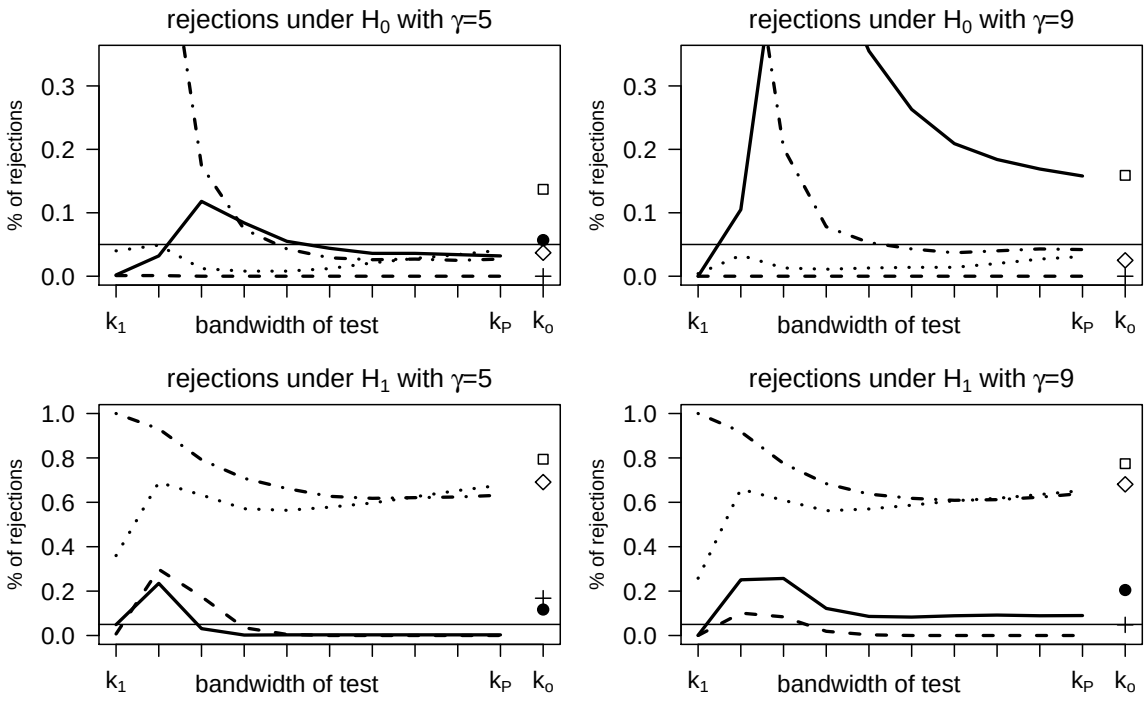

Figure 2: When $h_{-\delta} = 6h$: Real sizes (upper line) and rejection levels (lower line) over k, where k_o stands for adaptive choice referring to the bootstrap studentized (HSB) τ_j^{max}. Nominal size was $\alpha = 5\%$ (thin line). τ_1: solid line; τ_1^{max}: bullet (out of range in the upper right graph); τ_2: dashed line; τ_2^{max}: plus; τ_3: dotted line; τ_3^{max}: diamond; τ_4: dots and dashes; τ_4^{max}: square.

the other tests having almost as much power as τ_4 but with mitigated size problems. Test τ_2 is the only one that holds throughout the level but shows only some power in its k-adaptive version (right hand side, i.e. with k_o).

In Section 3.2 we discussed several strategies to find a data driven bandwidth k_o tailored toward the test. However, recall that both methods HS and GL are not always easy to implement; the GL requires for example good estimates of μ_4 and of the variances $v_4^2(k)$ and $v_{4b}^2(k, k_P)$. In fact, estimates of the conditional variance of Y are necessary which are also consistent under the alternative, and in the respective papers estimators where proposed which are only reasonable for one or two dimensional problems. Moreover, it is well known that the asymptotic expressions are little helpful for moderate samples. This is why we suggested to use the bootstrap to approximate also the required moments and called that approach HSB when applied to τ^{max} in (11). For

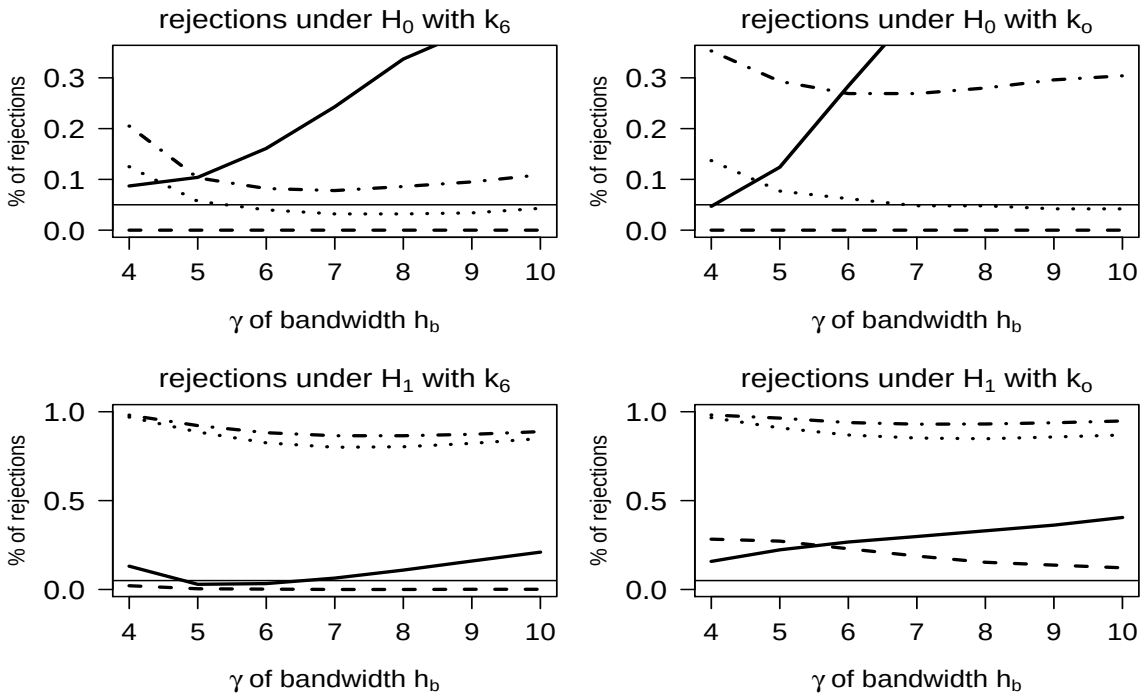

Figure 3: When $h_{-\delta} = 5h$: Real sizes (upper line) and rejection levels (lower line) over γ ($\gamma_1 = 4$ to $\gamma_7 = 10$) for k_6 and adaptive bandwidth (HSB) k_o. Nominal size was $\alpha = 5\%$. τ_1: solid line; τ_2: dashed line; τ_3: dotted line; τ_4: dots and dashes.

GL one has also to estimate the correlation of test statistics corresponding to different bandwidths which can be quite tedious even when using bootstrap. We therefore limit here our simulation comparison to the following: We estimated explicitly the required moments of τ_4 for the HS and GL procedure, i.e. the statistic for which the GL method has been derived with explicit estimators. Also the HS was developed for a modified version of τ_4 together with explicit expressions for the required moment estimators. For comparison, we present the HS and GL together with our HSB approach which we implemented for all tests. Figures 5 (for $h_{-\delta} = 5h$) and 6 (for $h_{-\delta} = 6h$) give the percentages of rejection over the range of γs considered in this study for all these k-adaptive methods. Only the GL method provides a test that holds the level for this model and sample size. As the other procedures produce much too liberal tests, it is not surprising that their power seems to be stronger.

Certainly, the 5% rejection level is just a particular size. Holding this α level does not mean that the procedure indeed fits well the (true) distribution of the test under the

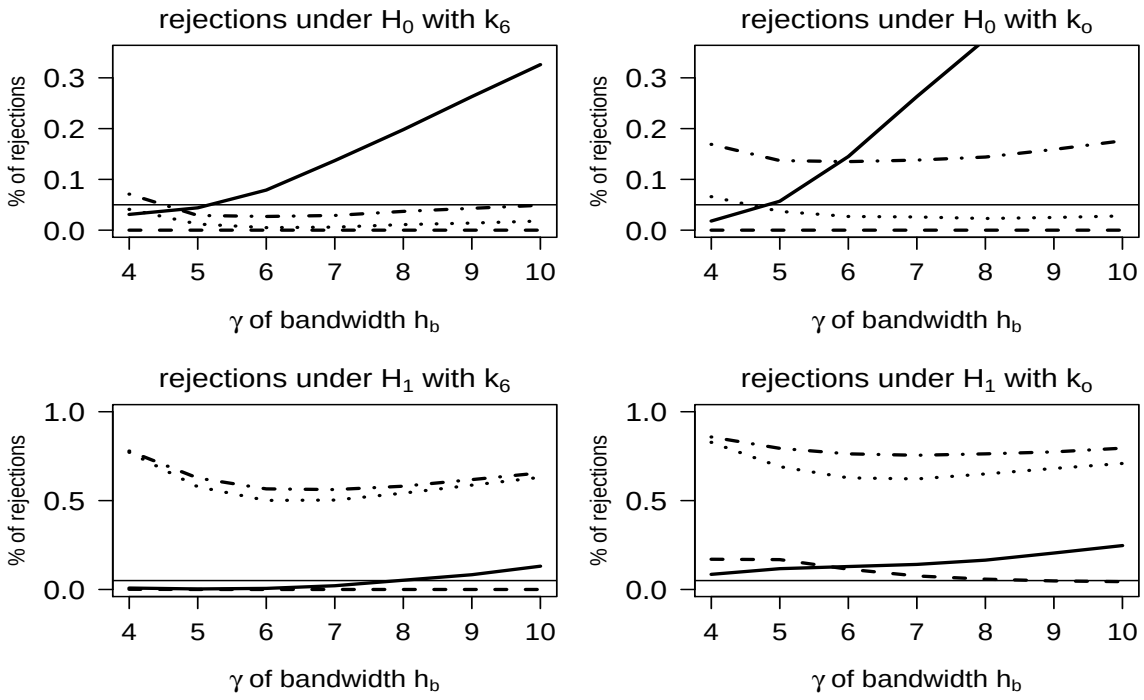

Figure 4: When $h_{-\delta} = 6h$: Real sizes (upper line) and rejection levels (lower line) over γ ($\gamma_1 = 4$ to $\gamma_7 = 10$) for k_6 and adaptive bandwidth (HSB) k_o. Nominal size was $\alpha = 5\%$. τ_1: solid line; τ_2: dashed line; τ_3: dotted line; τ_4: dots and dashes.

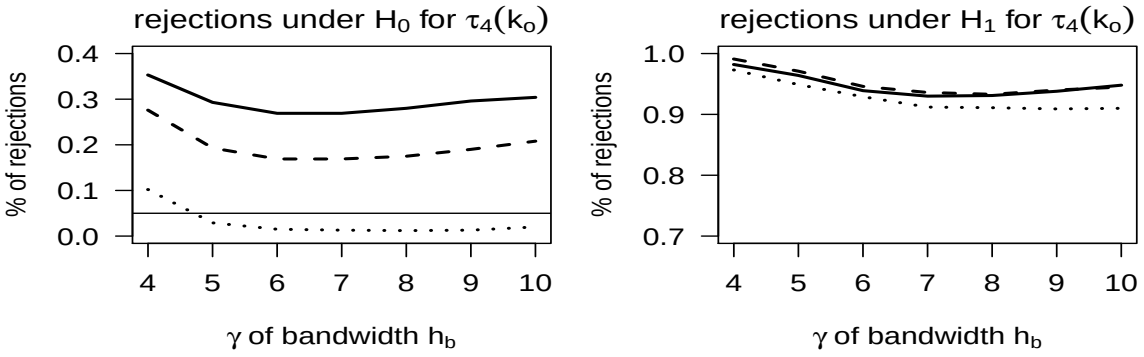

Figure 5: When $h_{-\delta} = 5h$: Real sizes (left) and rejection levels (right) of τ_4 over γ ($\gamma_1 = 4$, $\gamma_7 = 10$)with different k-adaptive versions: HSB: solid line; HS: dashed line; GL: dotted line. Nominal size was $\alpha = 5\%$.

null and therefore would work well in general. Tables 1 and 2 provide the p-values for all adaptive tests we have studied so far. We let γ of h_b run from 5 to 9. One can see that only $\tau_3^{max}(HSB)$ and $\tau_4(k_o - GL)$ can compete, giving p-values of about 0.5 or more under H_0 and reasonable power (i.e. small p-values) under the alternative H_1.

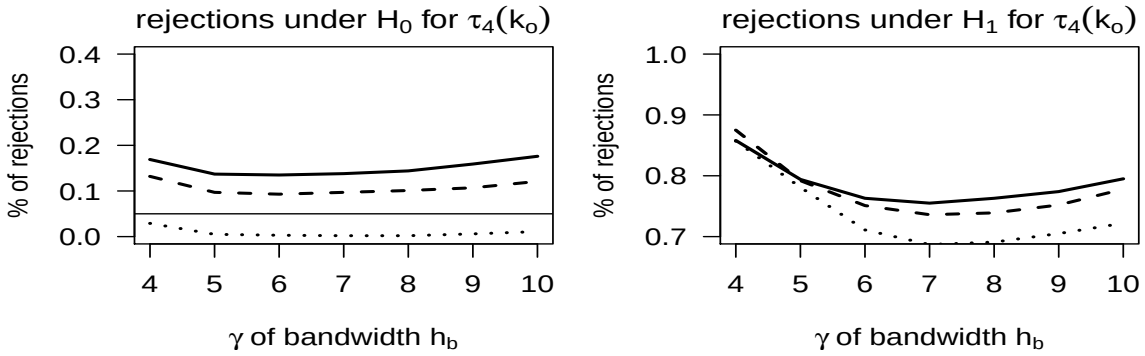

Figure 6: When $h_{-\delta} = 6h$: Real sizes (left) and rejection levels (right) of τ_4 over γ ($\gamma_1 = 4$, $\gamma_7 = 10$) with different k-adaptive versions: HSB: solid line; HS: dashed line; GL: dotted line. Nominal size was $\alpha = 5\%$.

	under H_0 *a=0.0*					*under* H_1 *a=2.0*				
$\gamma =$	5	6	7	8	9	5	6	7	8	9
$\tau_4(k_0 - GL)$	.485	.558	.577	.567	.544	.014	.019	.022	.022	.022
$\tau_4^{max}(HS)$	.269	.304	.309	.299	.285	.009	.014	.017	.017	.016
$\tau_4^{max}(HSB)$	.243	.275	.278	.269	.254	.009	.013	.016	.016	.016
$\tau_1^{max}(HSB)$	.350	.270	.199	.151	.120	.191	.191	.193	.189	.180
$\tau_2^{max}(HSB)$	.663	.746	.828	.880	.912	.196	.220	.256	.289	.315
$\tau_3^{max}(HSB)$	.350	.425	.476	.503	.515	.016	.025	.030	.032	.030

Table 1: P-values of k-adaptive tests under H_0 and H_1 when $h_{-\delta} = 5h$.

Next we turn to the automatic choice of h_b along the procedure introduced at the end

	under H_0 *a=0.0*					*under* H_1 *a=2.0*				
$\gamma =$	5	6	7	8	9	5	6	7	8	9
$\tau_4(k_o - GL)$	.712	.773	.779	.761	.732	.095	.116	.123	.121	.116
$\tau_4^{max}(HS)$	.495	.527	.521	.499	.472	.080	.103	.107	.101	.091
$\tau_4^{max}(HSB)$	.482	.514	.507	.484	.456	.079	.100	.104	.097	.088
$\tau_1^{max}(HSB)$	.548	.464	.364	.285	.231	.348	.383	.398	.391	.369
$\tau_2^{max}(HSB)$	.756	.839	.901	.938	.958	.328	.397	.457	.500	.527
$\tau_3^{max}(HSB)$	.520	.600	.639	.655	.655	.070	.102	.108	.099	.088

Table 2: P-values of k-adaptive tests under H_0 and H_1 when $h_{-\delta} = 6h$.

of Section 3.4. Let $\{Y_i^*, x_i^*\}_{i=1}^n := \mathcal{Y}^*$ be a member of the pseudo sequence of samples drawn from the original sample and following the same distribution. Then, to test $H_0^* : \ m(x) - m_S(x) = \hat{m}(x) - \hat{m}_S(x)$, the analogs to τ_1 and τ_3 would be

$$\frac{1}{n}\sum_{i=1}^{n}\left[\{\hat{m}(X_i) - \hat{m}_S(X_i)\} - \{\hat{m}^*(X_i) - \hat{m}_S^*(X_i)\}\right]^2, \quad (16)$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\mathbf{K}_k(X_i - X_j^*)\{Y_j^* - \hat{m}_S^*(X_j^*)\} - \mathbf{K}_k(X_i - X_j)\{Y_j - \hat{m}_S(X_j)\}\right]^2, \quad (17)$$

respectively, always neglecting the possible weighting by $w(\cdot)$ for brevity. Here, $\hat{m}^*$, $\hat{m}_S^*$ and $\hat{e}_i^*$ refer to estimates obtained from sample $\mathcal{Y}^*$. As this exercise was computationally rather expensive, we draw only $L = 100$ samples $\mathcal{Y}^*$ and repeated the bootstrap test as before but with the statistics (16) and (17) for all γ, and fixed $k = k_6 := k_1 + 5(k_P - k_1)/(P-1)$. We found that this procedure worked reasonably well for test τ_3. However, for both τ_1 and τ_3 this method somewhat underestimates the real rejection level. Interestingly, this method indicates that the real rejection level decreases for increasing γ; this is consistent with asymptotic theory but unfortunately not always with the practice. In our simulations for $h_{-\delta} = 5h$ this method recommends for τ_3 (given k_6) to take $\gamma \geq 5$ when the real data generating process is the null model, and $\gamma \geq 8$ when we draw the data from the alternative - recall that in practice we do not dispose of any information whether we are in H_0 or H_1. The predicted rejection level for τ_1 is below 1%. The real rejection levels (respectively the power) for the different γ can be seen in Figure 3; they confirm that these recommendations are not bad. For $h_{-\delta} = 6h$ the method gives no clear recommendation as it predicts for $\tau_3(k_6)$ a rejection level of about 2% for all γ, and below 1% for $\tau_1(k_6)$. Again, comparing these predictions with Figure 4 gives some hope that the procedure may work. We guess that also for τ_2 and τ_4 some analogs could be constructed.

Finally let us comment on our findings when doing subsampling instead of bootstraping. Still focusing on the nominal level $\alpha = 5\%$ we found that τ_1 never rejected except for k_{max} and $N \leq 0.4n$ but without having power. Also τ_2 was much too conservative and rejected under H_0 even more often than under H_1, whereas τ_3 held the level only for very large k, depending on subsample size N. Finally, τ_4 showed reasonable performance for $0.6n \leq N \leq 0.6n$ and $k_{max} \geq k \geq k_7$. Here, none of the k-adaptive versions worked

well. To summarize, the interplay between the choice of k and N seems to be even more crucial than the interplay between the choice of k and h_b in the bootstrap.

5 Conclusions

We discuss the choices of all "parameters" a practitioner has to choose when facing a smoothing based specification test, in particular when the null hypothesis is non- or semiparametric. We have set *parameters* in quotation marks because we refer here also to questions like how to generate bootstrap errors, etc. For illustration we have chosen the problem of testing additivity as this topic is well studied and many statistics have been proposed. We concentrate here on the kernel based methods, but it is clear that for other smoothers the choice and size problems are similar.

Typically, for the null model the practitioner has either a clear idea about the smoothness he wants to impose or he simply will go for a data-driven bandwidth h, cross-validation being then the most popular. However, when bootstrap methods are used to estimate the critical value, then the bandwidth h_b to pre-estimate the null model is crucial for the correct size and reasonable power of the test. In most cases the asymptotic theory says, that h/h_b has to go to zero, but our simulations reveal that this choice problem is much more involved in practice.

The choice of smoothness of the alternative and test, respectively, depends on k. There exists some literature on adaptive testing tackling exactly the problem of choosing k by maximizing the power. We have reviewed different procedures but implemented only for τ_4 three k-adaptive tests, namely GL, HS, and its poorly bootstrap based version HSB. For τ_1, τ_2, and τ_3 we implemented only the always easily available HSB version of the statistics. As we found the GL method to be successful, it would be interesting how the GL method can be more easily extended to a broad variety of statistics. It could also be interesting to study extensions of Gao & Gijbels (2008) to non- and semi-parametric null hypotheses.

As already mentioned in the context of choosing h_b, a main problem is the bootstrap and its size distortion in practice when the sample size is small or moderate. Concern-

ing the residuals and wild bootstrap procedure, we concentrated on bootstrap residuals taken under the null hypothesis for reasons discussed in detail. Further, our findings confirmed the ones of Davidson & Flachaire (2008) saying that often a simpler procedure outperforms theoretically more efficient ones. We only got reasonable results for a simple wild bootstraps where the new bootstrap residual is the product of a $N(0,1)$ random variable times the residual under H_0.

It is obvious that the proper choice of the different smoothing parameters in nonparameteric specification testing is essential for size and power. These various choice problems are complex, and most of the so far known methods are computationally rather expensive. Unfortunately, this problem is typically not addressed when a new test is proposed in the econometric literature. Moreover, in the existing literature little attention has been spent on this crucial problem at all. The proposal of applying an idea borrowed from subsampling to obtain a h_b that holds the nominal size seems to be a promising one. It entails, however, another computational expensive outer loop.

References

Barrientos, J. and Sperlich, S. (2010) The size problem of bootstrap tests when the null is non- or semiparametric. *Revista Colombiana de Estadística*, **33**, 307-319.

Cao-Abad, R. and González-Manteiga, W. (1993) Bootstrap methods in regression smoothing. *Nonparametric Statistics*, **2**, 379-388.

Claeskens, G. and Van Keilegom, I. (2003) Bootstrap confidence bands for regression curves and their derivatives. *Annals of Statistics*, **6**, 1852-1884.

Davidson, R. and Flachaire, E. (2008) The Wild Bootstrap, Tamed at Last. *Journal of Econometrics*, **146**, 162-169.

Delgado, M.A., Rodriguez-Poó, J.M. and Wolf, M. (2001) Subsampling Cube Root Asymptotics with an Application to Manski's MSE. *Economics Letters*, **73**, 241-250.

Dette, H., von Lieres und Wilkau, C. and Sperlich, S. (2005) A Comparison of Different Nonparametric Method for Inference on Additive Models. *Nonparametric Statistics*, **17**, 57-81.

Duong, T. and Hazelton, M. (2003) Plug-in bandwidth matrices for bivariate kernel density estimation. *Nonparametric Statistics*, **15**, 17-30.

Fahrmeir, L., Kneib, T., Lang, S. (2004) Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, **14**, 731761.

Fahrmeir, L., Kneib, T. and Konrath, S. (2010) Bayesian Regularisation in Structured Additive Regression: A Unifying Perspective on Shrinkage, Smoothing and Predictor Selection. *Statistics and Computing*, **20**, 203-219.

Gao, J. and Gijbels, I. (2008) Bandwidth Selection in Nonparametric Kernel Testing, *Journal of the American Statistical Association*, **484**, 1584-1594.

Gao, J., Gijbels, I. and Van Bellegem, S. (2008) Nonparametric simultaneous testing for structural breaks. *Journal of Econometrics*, **143**, 123-142.

González-Manteiga, W., Martínez-Miranda, M.D. and Pérez-González, A. (2004) The choice of smoothing parameter in nonparametric regression through Wild Bootstrap. *Computational Statistics & Data Analysis*, **47**, 487-515.

Gozalo, P.L. and Linton, O.B. (2001) Testing additivity in generalized nonparametric regression models with estimated parameters. *Journal of Econometrics*, **104**, 1-48.

Guerre, E. and Lavergne, P. (2005) Data-driven rate-optimal specification testing in regression models. *Annals of Statistics*, **33**, 840-870.

Härdle, W. and Mammen, E. (1993) Comparing Nonparametric Versus Parametric Regression Fits. *Annals of Statistics*, **21**, 1926-1947.

Härdle, W and Marron, J.S. (1991) Bootstrap Simultaneous Bars For Nonparametric Regression. *Annals of Statistics*, **19**, 778-796.

Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer.

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

Hengartner, N.W. and Sperlich, S. (2005) Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates. *Journal of Multivariate Analysis*, **95**, 246-272.

Horowitz, J.L. and Spokoiny, V. (2001) An Adaptive, Rate-optimal Test of Parametric Mean-Regression Model Against A Nonparametric Alternative. *Econometrica*, **69**, 599-631.

Ibragimov, I.A. and Khasminski, R.Z. (1981) *Statistical Estimation; Asymptotic Theory*. Springer.

Ingster, Yu.I. (1982) Minimax nonparametric detection of signals in white Gaussian noise, *Problems of Information Transmission*, **18**, 130-140.

Ingster, Yu.I. (1993) Asymptotically minimax hypothesis testing for nonparametric alternatives I, II, III, *Mathematical Methods of Statistics*. **2**, 85-114.

Kallenberg, W.C.M. and Ledwina, T. (1995) Consistency and Monte-Carlo simulatioins of a data driven version of smooth goodness-of-fit tests. *Annals of Statistics*, 23, 1594-1608.

Kim, W., Linton, O.B. and Hengartner, N. (1999) A computationally efficient oracle estimator of additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**, 278-297.

Köhler, M., Schindler, A. and Sperlich, S. (2012) A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. *International Statistical Review*, revised and resubmitted.

Kulasekera, K.B., and Wang, J. (1997) Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves. *Journal of the American Statistical Association*, **438**, 500-511.

Ledwina, T. (1994) Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, **89**, 1000-1005.

Linton, O.B. and Nielsen, J.P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93101.

Mammen, E., Linton, O.B. and Nielsen, J.P. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics*, **27**, 14431490.

Mammen, E. and Park, B. (2005) Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, **33**, 1260-1294.

Neumann, M.H. and Polzehl, J. (1998) Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, **9**, 307-333.

Neumeyer, N. and S. Sperlich, S. (2006) Comparison of Separable Components in Different Samples. *Scandinavian Journal of Statistics*, **33**, 477-501.

Nielsen, J.P. and S. Sperlich (2005) Smooth backfitting in practice. *Journal of the Royal Statistical Society, B*, **67**, 43-61.

Politis, D.N., Romano, J.P., and Wolf, M. (1999) *Subsampling*. Springer Series in Statistics. Springer.

Roca-Pardiñas, J. and Sperlich, S. (2007) Testing the link when the index is semiparamtric – A comparison study. *Computational Statistics & Data Analysis*, **12**, 6565-6581.

Rodriguez-Póo, J.M., Sperlich, S. and Vieu, P. (2012) And Adaptive Specification Test For Semiparametric Models. *Econometric Theory*, revised and resubmitted.

Rosenblatt, M. (1969) Conditional Probability Density and Regression estimators. *Multivariate Analysis*, **II**, 25-31.

Severance-Lossin, E. and Sperlich, S. (1999) Estimation of derivatives for additive separable models. *Statistics*, **33**, 241265.

Sperlich, S., Linton, O.B. and Härdle, W. (1999) Integration and backfitting methods in additive models: finite sample properties and comparison. *Test*, **8**, 419458.

Spokoiny, V. (1998) Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Mathematical Methods of Statistics*, **7**, 245-273.

Vieu, P. (1993) Bandwidth selection for kernel regression: a survey. In: Härdle, W., and Simar, L. (Eds.), *Computer Intensive Methods in Statistics, Statistics and Computing*, **1**, Physica, Berlin, 13-149.

Wood, S.N. (2008) Fast stable direct fitting and smoothness selection for Generalized Additive Models. *Journal of the Royal Statistical Society, B*, **70**, 495-518.