Specification tests indexed by bandwidths

Holger Dette, Benjamin Hetzler
Ruhr-Universität Bochum
Fakultät für Mathematik
44780 Bochum
Germany
email: holger.dette@ruhr-uni-bochum.de FAX: +49 234 3214 559

August 10, 2004

Abstract

In this note we consider several goodness-of-fit tests for model specification in nonparametric regression models which are based on kernel methods. In order to circumvent the problem of choosing a bandwidth for the corresponding test statistic we propose to consider the statistics as stochastic processes indexed with bandwidths proportional to the asymptotically optimal bandwidth for the estimation of the regression function. We prove weak convergence of these processes to centered Gaussian processes and suggest to use functionals of these processes as test statistics for the problem of model specification. A bootstrap test is proposed to obtain a good approximation of the nominal level. The results are illustrated by means of a simulation study and the new test is compared with some of the currently available procedures.

Keywords and Phrases: goodness-of-fit test, weak convergence, nonparametric regression, specification test, selection of smoothing parameters

1 Introduction

Recently, there has been a considerable interest in the problem of testing for a parametric model of the conditional mean $m(x) = E[\varepsilon \mid X = x]$ in a nonparametric regression model

$$Y = m(X) + \varepsilon, \tag{1.1}$$

where $Y \in \mathbb{R}, X \in \mathbb{R}^d$, $E[\varepsilon|X] = 0$ (a.s.) and $\text{Var}[\varepsilon|X = x] = \sigma^2(x) > 0$ (a.s.). Suppose we have a sequence of independent observations $\{(X_i, Y_i) \mid i = 1, \ldots, n\}$ coming from a population (X, Y) according to the model (1.1), in which the unknown regression function $m(x) = E(Y|X = x)$ is assumed to be smooth. Given a parametric function, say $g(x; \vartheta)$ for the conditional mean, the null and alternative hypotheses of a specific parametric form can be described as

$$H_0: \quad P\{m(X) = g(X; \vartheta_0)\} = 1, \text{ for some } \vartheta_0 \in \Theta, \tag{1.2}$$

$$H_1: \quad P\{m(X) = g(X; \vartheta)\} < 1, \text{ for any } \vartheta \in \Theta, \tag{1.3}$$

where $\Theta \subset \mathbb{R}^p$ denotes the parameter space. Many authors propose to compare a nonparametric with a parametric estimate of the regression function m [see e.g. Yatchew (1992), Wooldridge (1992), González-Manteiga and Cao (1993), Härdle and Mammen (1993), Zheng (1996, 1998a), Alcalá, Christóbal, González-Manteiga(1999) among many others]. Alternative test statistics have been proposed by Azzalini and Bowman (1993), Dette (1999), Fan, Zhang and Zhang (2001) and Fan and Li (2002). The various aspects of the different proposals have been carefully discussed by Zhang and Dette (2003), who demonstrated that there exist essentially three types of kernel based test statistics for the problem of testing the parametric form of the conditional mean in the regression model (1.1). These tests are attractive for practitioners, because of their easy interpretation and visualization (essentially only two curves have to be compared). However, goodness-of-fit tests using kernel based methods have been criticized by numerous authors because of their sensitivity with respect to the choice of a smoothing parameter required for the nonparametric estimation of the regression function.

In the present paper we provide a partial answer to the problem of choosing an appropriate bandwidth in these testing procedures by considering the various test statistics based on kernel methods as stochastic processes indexed by a bandwidth parameter, which is proportional to the optimal bandwidth for the nonparametric estimation of the regression function. We prove weak convergence of these processes to Gaussian processes under the null hypothesis (1.2) and under local alternatives, where the covariance structure depends on the particular testing procedure under consideration. As test statistic we propose the Kolmogorov-Smirnov or Cramér-von-Mises functional calculated over a certain range of bandwidths (proportional to the asymptotically optimal bandwidth) in order to obtain tests based on kernel methods, which are on the one hand less sensitive with respect to the specification of a bandwidth and on the other hand very powerful.

The work most similar to the spirit of the present note is the remarkable paper of Horowitz and Spokoiny (1999) who derived an adaptive rate-optimal test by considering the maximum of (standardized) test statistics of Härdle and Mammen (1993) calculated over a grid of bandwidths. While the work of theses authors is mainly motivated from a theoretical point of view (adaption of the test to the unknown smoothness of the alternative, optimal rates uniformly over Hölder classes), and leaves the problem of bandwidth choice (more precisely, the choice of the grid over which the maximum has to be taken) at least partially open, the present paper concentrates directly on the problem of bandwidth choice. In this sense the results of this paper can be considered as a complement to the work of Horowitz and Spokoiny (1999), whereas the new limiting stochastic processes obtained in Section 2 and 3 are of their own interest. Moreover, by considering only bandwidths, which are proportional to the optimal bandwidth for the nonparametric estimation of the regression function, it is intuitively clear that the resulting tests have very good power properties.

The concept is carefully described in Section 2, where we mainly discuss the test statistic introduced by Härdle and Mammen (1993) and used by Horowitz and Spokoiny (1999). The two other types of test statistics proposed by Zheng (1996) and Dette (1999) [for the latter see also Fan, Zhang and Zhang (2001) or Fan and Li (2002)] are briefly considered in Section 3, while Section 4 studies the finite sample properties of a bootstrap version of the new test. In particular we compare the new test with the test proposed by Horowitz and Spokoiny (1999) by adapting the approach used by these authors to the situation, where some preliminary knowledge regarding the smoothness of the regression function is available. This allows to consider bandwidths proportional to an optimal bandwidth (with respect to an integrated mean squared error criterion) in the adaptive procedure proposed by these authors and it is demonstrated that such additional knowledge can improve the power of the test substantially.

2 Testing parametric hypotheses by empirical processes indexed by bandwidths

A natural measure for the fit of a parametric estimate is a weighted L^2-distance between an estimate of the regression function under the null hypothesis and the alternative [see e.g. Härdle and Mammen (1993), González-Manteiga and Cao (1993), Weirather (1993) or Alcalá, Christóbal and González-Manteiga (1999) among many others]. Härdle and Mammen (1993) picked up an idea of Bickel and Rosenblatt (1973) and proposed the statistic

$$T_{n,h} = nh^{d/2} \int \left(\widehat{m}_h\left(x\right) - \mathcal{K}_{h,n} g(x; \widehat{\vartheta}) \right)^2 \pi\left(x\right)\, dx \tag{2.1}$$

as a measure for the deviation from the parametric form of the conditional mean, where $\widehat{\vartheta}$ is an estimate under the assumption of the parametric model,

$$\widehat{m}_h\left(x\right) = \frac{\sum_{i=1}^n K_h\left(x - X_i\right) Y_i}{\sum_{i=1}^n K_h\left(x - X_i\right)} \tag{2.2}$$

is the Nadaraya-Watson estimate [here we use the standard notation $K_h\left(x\right) = h^{-d} K\left(x/h\right)$]. Throughout this paper $K : \mathbb{R}^d \to \mathbb{R}$ denotes a d-dimensional kernel and we use the same bandwidth in each coordinate of X for the sake of simplicity. The operator $\mathcal{K}_{h,n}$ in (2.1) is defined by

$$\mathcal{K}_{h,n} g\left(\,\cdot\,\right) = \frac{\sum_{i=1}^n K_h\left(\,\cdot\, - X_i\right) g\left(X_i\right)}{\sum_{i=1}^n K_h\left(\,\cdot\, - X_i\right)} \tag{2.3}$$

and is used to remove a bias term of the statistic $T_{n,h}$ [see Härdle and Mammen (1993) or Bickel and Rosenblatt (1973)]. For the asymptotic inference these authors assumed that with probability 1 the random variable X_i lies in the d-dimensional cube $[0, 1]^d$ and that its marginal density $f(\cdot)$ is bounded away from 0. The regression function m and density f are assumed to be two times continuously differentiable and the conditional variance $\sigma^2(\cdot)$ is continuous. Moreover, the following conditions should be satisfied throughout this paper:

$$E[\exp(t\varepsilon)] \text{ is bounded in a neighbourhood of } 0, \tag{2.4}$$

$$g(\,\cdot\,;\widehat{\vartheta}) - g(\,\cdot\,;\vartheta_0) = (1/n)\sum_{i=1}^{n}\langle k\,(\,\cdot\,)\,,l\,(X_i)\rangle\,\varepsilon_i + o_p(\,(n\log n)^{-1/2}), \tag{2.5}$$

where k, ℓ are bounded $\mathbb{R}^d$-valued functions and $\langle\cdot,\cdot\rangle$ denotes the common inner product. Finally, it is assumed that the kernel in (2.2) and (2.3) is two times continuously differentiable. Härdle and Mammen (1993) showed that an appropriately standardized version of the statistic $T_{n,h}$ is asymptotically normal distributed and proposed a wild bootstrap test for the hypothesis (1.2). As pointed out by these authors, the testing procedure depends sensitively on the choice of the smoothing parameter in (2.2) and (2.3), and some care is necessary in the interpretation of the results of the corresponding test. A natural choice for the smoothing parameter is the asymptotically optimal bandwidth for the estimation of the regression function with respect to the integrated mean squared error criterion [see e.g. Fan and Gijbels (1996)]. It is known that in the present context this quantity is given by $h(c) = c \cdot n^{-1/(d+4)}$, but the constant c is in general unknown and depends on certain features of the regression-, variance function and design density.

In this note we propose to consider the statistic $T_{n,h}$ as function of the constant c in the asymptotically optimal bandwidth

$$h(c) = cn^{-1/(d+4)}, \tag{2.6}$$

that is

$$T_n(c) = T_{n,h(c)}, \tag{2.7}$$

where the statistic $T_{n,h}$ is defined in (2.1), $c \in [\underline{c}, \bar{c}]$, and $0 < \underline{c} < \bar{c} < \infty$ are given constants. We will prove below that an appropriate centered version of the process $(T_n(c))_{c\in[\underline{c},\bar{c}]}$ converges weakly to a Gaussian process.

Theorem 2.1. *If the assumptions stated at the beginning of this section, (2.4), (2.5) are satisfied, and the function $b_n(c)$ is defined by*

$$b_n(c) = c^{-d/2} n^{\frac{d}{2(d+4)}} \int K^2(u)du \int \frac{\sigma^2(x)\pi(x)}{f(x)}dx, \tag{2.8}$$

then the process $(T_n(c) - b_n(c))_{c\in[\underline{c},\bar{c}]}$ converges weakly to a centered Gaussian process $(G_{HM}(c))_{c\in[\underline{c},\bar{c}]}$ with covariance kernel

$$k_{HM}(c_1, c_2) = 2\Big(\frac{c_1}{c_2}\Big)^{d/2} \int \Big[\int K\,(u)\,K\Big(\frac{c_1}{c_2}u + z\Big)\,du\Big]^2 dz \int \frac{\sigma^4\,(x)}{f^2\,(x)}\pi^2(x)dx, \tag{2.9}$$

Proof. We only consider the case $d = 1$ and $\pi(x) \equiv 1$, the general case $d > 1$ and an arbitrary weight function can be obtained similarly. Following Härdle and Mammen (1993) we have

$$T_n(c) = T_{n,3}(c) + b_n(c) + o_p(1), \tag{2.10}$$

where the statistic $T_{n,3}(c)$ is defined by

$$T_{n,3}(c) = \frac{\sqrt{h(c)}}{n} \int_0^1 \frac{\sum_{i \neq j} K_{h(c)}(X_i - x) K_{h(c)}(X_j - x) \varepsilon_i \varepsilon_j}{f^2(x)} dx, \tag{2.11}$$

and the constant $b_n(c)$ is given by (2.8). Consequently, the assertion of the theorem can be established showing that

$$\{T_{n,3}(c)\}_{c \in [\underline{c}, \bar{c}]} \Rightarrow (G_{HM}(c))_{c \in [\underline{c}, \bar{c}]}, \tag{2.12}$$

where the symbol $\Rightarrow$ denotes weak convergence in the Skorohod space $D([\underline{c}, \bar{c}])$. For this purpose we define

$$W_{ijn}(c) = \begin{cases} \dfrac{\sqrt{h(c)}}{n} \displaystyle\int_0^1 \frac{K_{h(c)}(X_i - x) K_{h(c)}(X_j - x)}{f^2(x)} dx \, \varepsilon_i \varepsilon_j & \text{if } i \neq j, \\ 0 & \text{else,} \end{cases} \tag{2.13}$$

introduce the notation $h_j = h(c_j)$ $(j = 1, 2)$ and obtain for the asymptotic covariance

$$\begin{aligned}
\operatorname{Cov}(T_{n,3}(c_1), T_{n,3}(c_2)) &= 2n(n-1) \, \mathrm{E}[W_{ijn}(c_1) \cdot W_{ijn}(c_2)] \\
&= 2 \, \mathrm{E}\Big[\sqrt{h_1 h_2} \int \frac{K_{h_1}(x - X_i) K_{h_1}(x - X_j)}{f^2(x)} dx \\
&\quad \times \int \frac{K_{h_2}(y - X_i) K_{h_2}(y - X_j)}{f^2(y)} dy \; \varepsilon_i^2 \varepsilon_j^2\Big] \\
&= 2\sqrt{h_1 h_2} \int\!\!\int\!\!\int\!\!\int \frac{1}{h_2^2} \frac{K(u) K(v) K\left(\frac{h_1}{h_2} u + \frac{y-x}{h_2}\right) K\left(\frac{h_1}{h_2} v + \frac{y-x}{h_2}\right)}{f^2(x) f^2(y)} \\
&\quad \times \sigma^2(x - h_1 u) \sigma^2(x - h_1 v) f(x - h_1 u) f(x - h_1 v) \, du \, dv \, dx \, dy \\
&= 2\sqrt{r} \int \left[\int K(u) K(ru + z) \, du\right]^2 dz \cdot \int \frac{\sigma^4(x)}{f^2(x)} dx + o(1),
\end{aligned}$$

where we used the continuity of the conditional variance and the notation $r = c_1/c_2$. This proves the representation of the covariance kernel in Theorem 2.1 (for the case $\pi(x) \equiv 1$ and $d = 1$).

For a proof of the claimed weak convergence in (2.12) we show convergence of the finite dimensional distributions and that there exists an $M \in \mathbb{R}$ such that

$$\mathrm{E}\left[|T_{n,3}(c_1) - T_{n,3}(c_2)|^2 |T_{n,3}(c_2) - T_{n,3}(c_3)|^2\right] \leq M (c_1 - c_3)^2 \;\; \forall c_1 \leq c_2 \leq c_3 \tag{2.14}$$

[see Billingsley (1968), Theorem 15.6]. The convergence of the finite dimensional distributions follows along the lines of Härdle and Mammen (1993) using the Cramér-Wold device and is omitted for the sake of brevity. For a proof of the estimate (2.14) we write

$$\begin{aligned} T_{n,3}(c_1) - T_{n,3}(c_2) &= \sum_{i\neq j} a_{ij}, \\ T_{n,3}(c_2) - T_{n,3}(c_3) &= \sum_{i\neq j} b_{ij}, \end{aligned} \tag{2.15}$$

with $h_j = h(c_j)$ $(j = 1, 2, 3)$,

$$\begin{aligned} a_{ij} &:= \Big(\frac{\sqrt{h_1}}{n}\int_0^1 \frac{K_{h_1}(X_i - x)\,K_{h_1}(X_j - x)}{f^2(x)}dx - \frac{\sqrt{h_2}}{n}\int_0^1 \frac{K_{h_2}(X_i - x)\,K_{h_2}(X_j - x)}{f^2(x)}\,dx\Big)\varepsilon_i\varepsilon_j, \\ b_{ij} &:= \Big(\frac{\sqrt{h_2}}{n}\int_0^1 \frac{K_{h_2}(X_i - x)\,K_{h_2}(X_j - x)}{f^2(x)}dx - \frac{\sqrt{h_3}}{n}\int_0^1 \frac{K_{h_3}(X_i - x)\,K_{h_3}(X_j - x)}{f^2(x)}\,dx\Big)\varepsilon_i\varepsilon_j, \end{aligned}$$

and a straightforward calculation shows

$$\begin{aligned} k_n(c_1, c_2, c_3) &= \mathrm{E}\left[|T_{n,3}(c_1) - T_{n,3}(c_2)|^2 |T_{n,3}(c_2) - T_{n,3}(c_3)|^2\right] \\ &= \mathrm{E}\sum_{i\neq j} a_{ij} \sum_{i'\neq j'} a_{i'j'} \sum_{k\neq l} b_{kl} \sum_{k'\neq l'} b_{k'l'} \\ &= 8\,\mathrm{E}\sum_{\neq} a_{ij}^2 b_{kl}^2 + 8\,\mathrm{E}\sum_{\neq} a_{ij} b_{ij} a_{kl} b_{kl} + o(1) \qquad (2.16) \\ &= 8\sum_{i\neq j} \mathrm{E}\left[a_{ij}^2\right] \sum_{k\neq l} \mathrm{E}\left[b_{kl}^2\right] + 8\sum_{i\neq j} \mathrm{E}\left[a_{ij} b_{ij}\right] \sum_{k\neq l} \mathrm{E}\left[a_{kl} b_{kl}\right] + o(1) \\ &= 8\sum_{i\neq j} \mathrm{E}\left[a_{ij}^2\right] \sum_{k\neq l} \mathrm{E}\left[b_{kl}^2\right] + 8\left(\sum_{i\neq j} \mathrm{E}\left[a_{ij} b_{ij}\right]\right)^2 + o(1) \end{aligned}$$

uniformly with respect to c_1, c_2, c_3, where the symbol $\sum_{\neq}$ means the summation over only all pairwise different indices. Now it is easy to see that

$$\begin{aligned} A_n &= 2\sum_{i\neq j} \mathrm{E}\left[a_{ij}^2\right] = 2n(n-1)\,\mathrm{E}\left[a_{12}^2\right] \qquad (2.17) \\ &= 2n(n-1)\,\mathrm{E}\left\{\frac{h_1}{n^2}\left(\int_0^1 \frac{K_{h_1}(X_1 - x)\,K_{h_1}(X_2 - x)}{f^2(x)}\,dx\right)^2 \varepsilon_1^2\varepsilon_2^2\right. \\ &\quad - \frac{2\sqrt{h_1 h_2}}{n^2}\left(\int_0^1 \frac{K_{h_1}(X_1 - x)\,K_{h_1}(X_2 - x)}{f^2(x)}\,dx\right. \\ &\quad \times \left.\int_0^1 \frac{K_{h_2}(X_1 - y)\,K_{h_2}(X_2 - y)}{f^2(y)}\,dy\right)\varepsilon_1^2\varepsilon_2^2 \\ &\quad + \left.\frac{h_2}{n^2}\left(\int_0^1 \frac{K_{h_2}(X_1 - x)\,K_{h_2}(X_2 - x)}{f^2(x)}\,dx\right)^2 \varepsilon_1^2\varepsilon_2^2\right\} \\ &= 2\,\mathit{Var}\;(T_{n,3}(c_1)) - 2\,\mathrm{Cov}\,(T_{n,3}(c_1), T_{n,3}(c_2)), \end{aligned}$$

and similarly

$$B_n = 2\sum_{k\neq l} \mathrm{E}\left[b_{kl}^2\right] = 2\mathrm{Var}\ \left(T_{n,3}\left(c_2\right)\right) - 2\,\mathrm{Cov}\left(T_{n,3}\left(c_2\right), T_{n,3}\left(c_3\right)\right). \tag{2.18}$$

For the remaining term we obtain by the same arguments

$$\begin{aligned}
c_n &= 2\sum_{i\neq j} \mathrm{E}\left[a_{ij}b_{ij}\right] = 2n(n-1)\,\mathrm{E}\left[a_{12}b_{12}\right] \\
&= 2n(n-1)\,\mathrm{E}\left\{\frac{\sqrt{h_1h_2}}{n^2}\left(\int_0^1 \frac{K_{h_1}\left(X_1-x\right)K_{h_1}\left(X_2-x\right)}{f^2\left(x\right)}dx\right.\right. \\
&\qquad\qquad \left.\times\int_0^1 \frac{K_{h_2}\left(X_1-y\right)K_{h_2}\left(X_2-y\right)}{f^2\left(y\right)}dy\right)\varepsilon_1^2\varepsilon_2^2 \\
&- \frac{\sqrt{h_1h_3}}{n^2}\int_0^1 \frac{K_{h_1}\left(X_1-x\right)K_{h_1}\left(X_2-x\right)}{f^2\left(x\right)}dx\int_0^1 \frac{K_{h_3}\left(X_1-y\right)K_{h_3}\left(X_2-y\right)}{f^2\left(y\right)}dy\varepsilon_1^2\varepsilon_2^2 \\
&- \frac{h_2}{n^2}\left(\int_0^1 \frac{K_{h_2}\left(X_1-x\right)K_{h_2}\left(X_2-x\right)}{f^2\left(x\right)}dx\right)^2\varepsilon_1^2\varepsilon_2^2 \\
&+ \left.\frac{\sqrt{h_2h_3}}{n^2}\int_0^1 \frac{K_{h_2}\left(X_1-x\right)K_{h_2}\left(X_2-x\right)}{f^2\left(x\right)}dx\int_0^1 \frac{K_{h_3}\left(X_1-y\right)K_{h_3}\left(X_2-y\right)}{f^2\left(y\right)}dy\varepsilon_1^2\varepsilon_2^2\right\},
\end{aligned}$$

which yields

$$\begin{aligned}
C_n = 2c_n^2 &= 2\left[\mathrm{Cov}\left(T_{n,3}\left(c_1\right), T_{n,3}\left(c_2\right)\right) - \mathrm{Cov}\left(T_{n,3}\left(c_1\right), T_{n,3}\left(c_3\right)\right)\right. \\
&\quad \left. -\mathrm{Var}\ \left(T_{n,3}\left(c_2\right)\right) + \mathrm{Cov}\left(T_{n,3}\left(c_2\right), T_{n,3}\left(c_3\right)\right)\right]^2.
\end{aligned} \tag{2.19}$$

Now the same arguments as given at the beginning of this proof show

$$\mathrm{Cov}\left(T_{n,3}\left(c_i\right), T_{n,3}\left(c_j\right)\right) \leq k_{\mathrm{HM}}\left(c_i, c_j\right)\left(1+o(1)\right) \tag{2.20}$$

uniformly with respect to c_i, c_j, and it follows (using the differentiability of the kernel k_{HM}) that

$$\begin{aligned}
A_nB_n &\leq C_1|c_1-c_2||c_2-c_3| \leq C_1|c_1-c_3|^2 \\
C_n &\leq C_2|c_1-c_3|^2
\end{aligned} \tag{2.21}$$

for some constants $C_1, C_2 > 0$. This completes the proof of the estimate (2.14) and the weak convergence in (2.12) follows directly from Theorem 15.6 in Billingsley (1968). Finally, the assertion of Theorem 2.1 follows from the decomposition (2.10).

□

Example 2.2. The covariance kernel k_{HM} in Theorem 2.1 depends on the kernel K used in the nonparametric regression estimate. For example, if $d=1$ and $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ is the Gaussian kernel we obtain by a straightforward but tedious calculation

$$k_{\mathrm{HM}}(c_1, c_2) = \sqrt{\frac{c_1c_2}{\pi(c_1^2+c_2^2)}}\int \frac{\sigma^4\left(x\right)}{f^2\left(x\right)}\pi^2(x)dx. \tag{2.22}$$

If $K(u) = \frac{1}{2} I_{[-1,1]}(u)$ is the rectangular kernel, it follows that for $c_1 \geq c_2$

$$k_{\mathrm{HM}}(c_1, c_2) = \sqrt{\frac{c_2}{c_1}} (1 - \frac{c_2}{3c_1}) \int \frac{\sigma^4(x)}{f^2(x)} \pi^2(x) dx \tag{2.23}$$

while for the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2) I_{[-1,1]}(u)$ we obtain

$$k_{\mathrm{HM}}(c_1, c_2) = \sqrt{\frac{c_2}{c_1}} \left\{ \frac{6}{5} - \frac{3}{5} \Big(\frac{c_2}{c_1}\Big)^2 + \frac{2}{7} \Big(\frac{c_2}{c_1}\Big)^3 - \frac{1}{55} \Big(\frac{c_2}{c_1}\Big)^5 \right\} \int \frac{\sigma^4(x)}{f^2(x)} \pi^2(x) dx. \tag{2.24}$$

The remaining cases $c_1 \leq c_2$ are obtained by symmetry.

Remark 2.3. If local alternatives of the form

$$m(x) = g(x; \vartheta_0) + c^{-\frac{d}{4}} n^{-\frac{d+8}{4(d+4)}} r(x)$$

for some fixed $\vartheta_0 \in \Theta$ are considered, the assertion of Theorem 2.1 is still correct, where the centered Gaussian process G_{HM} has to be replaced by a Gaussian process with mean

$$\int r^2(x) \pi(x) dx$$

and the same covariance kernel k_{HM}. This follows by a careful inspection of the proof of the decomposition (2.10) [see also Härdle and Mammen (1993) for more details].

A closely related test statistic was considered by Horowitz and Spokoiny (1999), who derived an adaptive rate-optimal test by considering the maximum of a studentized version of the statistic $T_{n,h}$, calculated on a grid of bandwidths. In contrast to our work these authors did not assume knowledge about the smoothness of the regression function and (as a consequence) did not use bandwidths proportional to the optimal bandwidth, which allows to consider a process indexed by bandwidths. In the notation of the present paper the statistic discussed by these authors is asymptotically first order equivalent to the statistic

$$\widetilde{T}_{n,h} = \frac{T_{n,h} - b_h}{v}, \tag{2.25}$$

where the bias b_h and the variance v^2 are defined by

$$b_h = h^{-d/2} \int K^2(u) du \int \frac{\sigma^2(x) \pi(x)}{f(x)} dx \tag{2.26}$$

$$v^2 = 2 \int (K * K)^2(u) du \int \frac{\sigma^4(x)}{f^2(x)} \pi^2(x) dx, \tag{2.27}$$

respectively, and $K * K$ denotes the convolution of K with itself. For bandwidths of the form $h(c) = cn^{-1/(d+4)}$ we obtain the following corollary from Theorem 2.1.

Corollary 2.4. *If the assumptions of Theorem 2.1 are satisfied and $\widetilde{T}_n(c) = \widetilde{T}_{n,h(c)}$ denotes the statistic $\widetilde{T}_{n,h}$ defined in (2.25) for the bandwidth $h = h(c) = cn^{-1/(d+4)}$, then*

$$(\widetilde{T}_n(c))_{c\in[\underline{c},\bar{c}]} \Rightarrow (\widetilde{G}(c))_{c\in[\underline{c},\bar{c}]},$$

where $\widetilde{G}$ denotes a centered Gaussian process with covariance kernel

$$k_{HS}(c_1, c_2) = \frac{\left(\frac{c_1}{c_2}\right)^{d/2} \int \left[\int K(u)K(\frac{c_1}{c_2}u + z)du\right]^2 dz}{\int (K * K)^2(u)du}.$$

3 Related processes

In this section we briefly discuss two related processes corresponding to kernel based methods, which have been recently proposed in the literature for testing the parametric form of the conditional mean. For the sake of brevity we restrict ourselves to the case $d = 1$ and $\pi(x) \equiv 1$, but similar results for the case $d > 1$ and general weight functions are readily available. Zheng (1996) proposed to reject the null hypothesis (1.2) for large values of the statistic

$$V_n = \frac{\sqrt{h}}{n-1} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{h} K\left(\frac{X_i - X_j}{h}\right) e_i e_j, \tag{3.1}$$

where

$$e_i = Y_i - g(X_i; \widehat{\vartheta}) \quad (i = 1, \dots, n) \tag{3.2}$$

denote the residuals based on the parametric fit. Note that it is heuristically clear that V_n attains large values if the hypothesis (1.2) is not satisfied. For example, if the least squares technique is used for the estimation of the parameter ϑ, V_n is a consistent estimate of

$$E[(Y_i - g(X_i; \vartheta_0))E[Y_i - g(X_i; \vartheta_0) \mid X_i]], \tag{3.3}$$

[see e.g. Zhang and Dette (2003)] where

$$\vartheta_0 := \arg\min_{\vartheta\in\Theta} E[(Y - g(X; \vartheta))^2], \tag{3.4}$$

and the expression in (3.3) vanishes if and only if the null hypothesis (1.2) is satisfied. A rather different approach for testing the parametric form of the conditional mean was proposed by Dette (1999) which is based on the difference of variance estimators under the null hypothesis and alternative [see also Fan, Zhang and Zhang (2001) or Fan and Li (2002) who used a similar method]. This author suggested to reject the null hypothesis for large values of the statistic

$$U_n = n\sqrt{h}(\widehat{\sigma}^2_{H_0} - \widehat{\sigma}^2_{H_1}), \tag{3.5}$$

where $\widehat{\sigma}^2_{H_0}$ is the common estimate of the variance under the null hypothesis (1.2), i.e.

$$\widehat{\sigma}^2_{H_0} = \frac{1}{n-p} \sum_{j=1}^{n} e_j^2, \tag{3.6}$$

and $\widehat{\sigma}^2_{H_1}$ is the variance estimator based on the sum of squared residuals from a nonparametric fit with the Nadaraya-Watson estimate [see Hall and Marron (1990) or Dette (1999) for more details]. For least squares estimation it was shown by Dette (1999) that the statistic U_n is a consistent estimate of $E[(m(X) - g(X;\vartheta_0))^2]$ and similar properties can be derived for other types of estimators. The following theorem gives the analogue of Theorem 2.1 for processes based on V_n and U_n. The proof is very similar to the proof of Theorem 2.1 and omitted for the sake of brevity.

Theorem 3.1. *Let $V_n(c)$ and $U_n(c)$ denote the statistics V_n and U_n defined in (3.1) and (3.5), respectively, with the bandwidth $h(c) = cn^{-1/5}$. If the assumptions of Theorem 2.1 are satisfied, then under the null hypothesis (1.2) we have*

$$\begin{aligned} (V_n(c))_{c\in[\underline{c},\bar{c}]} &\Rightarrow (G_Z(c))_{c\in[\underline{c},\bar{c}]} \\ (U_n(c) - b_D(c))_{c\in[\underline{c},\bar{c}]} &\Rightarrow (G_D(c))_{c\in[\underline{c},\bar{c}]} \end{aligned}$$

where $(G_Z(c))_{c\in[\underline{c},\bar{c}]}$ and $(G_D(x))_{c\in[\underline{c},\bar{c}]}$ are centered Gaussian processes with covariance kernels

$$k_Z(c_1,c_2) = 2\sqrt{\frac{c_1}{c_2}} \int K(u)\, K\left(\frac{c_1}{c_2}u\right) du \int \sigma^4(x)\, f^2(x)\, dx, \tag{3.7}$$

$$\begin{aligned} k_D(c_1,c_2) &= 2\sqrt{\frac{c_1}{c_2}} \int \sigma^4(v)\, dv \int [2K(u) - K * K(u)] \\ &\quad \times \left[2K\left(\frac{c_1}{c_2}u\right) - \frac{c_1}{c_2}\int K\left(\frac{c_1}{c_2}(u-z)\right) K\left(\frac{c_1}{c_2}z\right) dz\right] du, \end{aligned} \tag{3.8}$$

respectively, and the bias term $b_D(c)$ is defined by

$$\begin{aligned} b_D(c) &= -\frac{c^{9/2}n^{1/10}}{4}\Big(\int u^2K(u)du\Big)^2 \cdot \int_0^1 \Big\{(mf)^{(2)}(u) - mf^{(2)}(u)\Big\}\frac{du}{f(u)} \\ &\quad -\frac{n^{1/10}}{c^{1/2}}\Big(2K(0) - \int K^2(u)du\Big)\Big(\int_0^1 \sigma^2(t)(f(t)-1)dt\Big) \end{aligned} \tag{3.9}$$

Example 3.2. For the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$ the covariance kernels corresponding to the Gaussian processes $(G_Z(c))_{c\in[\underline{c},\bar{c}]}$ and $(G_D(c))_{c\in[\underline{c},\bar{c}]}$ are given by

$$\begin{aligned} k_Z(c_1,c_2) &= \sqrt{\frac{2c_1c_2}{\pi\left(c_1^2+c_2^2\right)}}\int \sigma^4(x)f^2(x)dx, \\ k_D(c_1,c_2) &= \left\{\left(4\sqrt{2}+1\right)\sqrt{\frac{c_1c_2}{\pi\left(c_1^2+c_2^2\right)}} - 2\sqrt{\frac{c_1c_2}{\pi\left(c_1^2+\frac{c_2^2}{2}\right)}} - 2\sqrt{\frac{c_1c_2}{\pi\left(\frac{c_1^2}{2}+c_2^2\right)}}\right\}\int \sigma^4(x)dx, \end{aligned} \tag{3.10}$$

respectively. For the rectangular kernel $K(u) = \frac{1}{2}I_{[-1,1]}(u)$ we obtain

$$k_Z(c_1, c_2) \;=\; \sqrt{\frac{c_1}{c_2}} \int \sigma^4(x) f^2(x) dx \quad (c_1 \le c_2) \tag{3.11}$$

$$k_D(c_1, c_2) \;=\; \begin{cases} \sqrt{\frac{c_1}{c_2}} \left\{ \frac{1}{6}\frac{c_1}{c_2} - 2\frac{c_1}{c_2} + \frac{1}{2}\left(\frac{c_1}{c_2}\right)^2 + 3 \right\} \int \sigma^4(x) dx & if \;\; \frac{c_1}{c_2} \in [1,2] \\ \sqrt{\frac{c_1}{c_2}} \left\{ \frac{1}{6}\frac{c_1}{c_2} + 1 \right\} \int \sigma^4(x) dx & if \;\; \frac{c_1}{c_2} \ge 2 \end{cases} \tag{3.12}$$

(the other case is obtained by symmetry).

4 Finite sample comparison

In this section we briefly investigate the finite sample performance of tests based on functionals of the stochastic processes discussed in Section 3. In principle all processes considered in this section will yield a test for the parametric form of the conditional mean in the nonparametric regression model and for the sake of comparison (with the results of Horowitz and Spokoiny (1999)) and brevity we restrict ourselves to a version of the process $\widetilde{T}_n(c)$ considered in Corollary 2.4. To be precise, we write the statistic $T_{n,h}$ considered in (2.1) as a quadratic form, i.e.

$$T_{n,h} = \sum_{i,j=1}^{n} a_{ij}(h)(Y_j - g(X_j; \widehat{\vartheta}))(Y_i - g(X_i; \widehat{\vartheta})),$$

and consider the estimates of the bias

$$\widehat{b}_h = \sum_{i=1}^{n} a_{ii}(h)\widehat{\sigma}_i^2, \tag{4.1}$$

and variance

$$\widehat{v}_h^2 = 2\sum_{i,j=1}^{n} a_{ij}^2(h)\widehat{\sigma}_i^2\widehat{\sigma}_j^2, \tag{4.2}$$

where $\widehat{\sigma}_i^2$ denotes an estimate of the conditional variance at the point X_i $(i = 1, \ldots, n)$. For bandwidths of the form (2.6) we investigate the stochastic process

$$\bar{T}_n(c) = \frac{T_{n,h(c)} - \widehat{b}_{h(c)}}{\widehat{v}_{h(c)}}, \tag{4.3}$$

which was also discussed by Horowitz and Spokoiny (1999) and is asymptotically first order equivalent to the process considered in Corollary 2.4 under appropriate assumptions on the variance estimates. Therefore it is easy to see that this process converges weakly to the process

$(\widetilde{G}(c))_{c\in[\underline{c},\bar{c}]}$ defined in Corollary 2.4 and the null hypothesis should be rejected for large values of a functional of $|\bar{T}_n(c)|$ such as

$$T_{\text{KS}} = \max_{c\in[\underline{c},\bar{c}]} |\bar{T}_n(c)| \quad \text{or} \quad T_{CM} = \int_{\underline{c}}^{\bar{c}} \bar{T}_n^2(c)dc. \tag{4.4}$$

For the sake of comparison with the work of Horowitz and Spokoiny (1999) we consider the Kolmogorov-Smirnov-statistic (4.4) in a homoscedastic regression model. These authors recommend a bootstrap procedure, which is also used in the present simulation study, where the conditional variances in (4.1) and (4.2) are estimated by the difference estimate of Rice (1984). To be precise, we generated data from the parametric model

$$Y_i^* = g(X_i; \widehat{\vartheta}) + \varepsilon_i^* \tag{4.5}$$

where $\varepsilon_1^*, \ldots, \varepsilon_n^*$ are i.i.d. with a $\mathcal{N}(0, \widehat{\sigma}^2_{H_0})$ law and $\widehat{\vartheta}$ is the least squares estimate from the original sample. The bootstrap statistic T^*_{KS} is calculated from (1.2) with the data $Y_1^*, \ldots, Y_n^*$ and the null hypothesis is rejected at level $\alpha \in (0,1)$ if T_{KS} is larger than the corresponding quantile of the bootstrap distribution.
Note that the essential difference of the method proposed in this paper to the work of Horowitz and Spokoiny (1999) is the specific choice of smoothing parameters proportional to the asymptotically optimal bandwidth. This allows us to consider the statistic $\bar{T}_n(c)$ as a stochastic process of the parameter c, which converges weakly in the Skorohod space $D([\underline{c}, \bar{c}])$. In our simulation we replace the bandwidth (2.6) by

$$h(c) = c \cdot \Big(\frac{\widehat{\sigma}^2_{H_0}}{n}\Big)^{1/(d+4)}; \; c \in [\underline{c}, \bar{c}] \tag{4.6}$$

in order to reflect different standard deviations in the errors. For the range $[\underline{c}, \bar{c}]$ (over which the stochastic process has to be considered) we choose the interval

$$[\underline{c}, \bar{c}] = [1, 9], \tag{4.7}$$

which covers a rather broad area of bandwidths.

Example 4.1. Our first example gives a comparison with the results of Horowitz and Spokoiny (1999) and demonstrates the advantages of choosing bandwidths proportional to the asymptotic optimal bandwidth. These authors investigated the model

$$Y_i = 1 + X_i + \alpha\Big\{\frac{5}{\tau}\varphi\Big(\frac{X_i}{\tau}\Big)\Big\} + \varepsilon_i \quad i = 1, \ldots, n, \tag{4.8}$$

where φ denotes the density of the standard normal distribution, $\tau = \frac{1}{4}$ or $\tau = 1$, X_i is a univariate random variable sampled from a $\mathcal{N}(0, 25)$ distribution, which is truncated at the 5th and 95th percentile. Three distributions were considered for the errors, that is

$$\varepsilon_i \sim \mathcal{N}(0, 4), \tag{4.9}$$

$$\varepsilon_i \sim I\{B = 1\}\mathcal{N}(0, 1.56) + I\{B = 0\}\mathcal{N}(0, 25), \tag{4.10}$$

$$\varepsilon_i \sim 2\sqrt{6}\frac{G - \gamma}{\pi}, \tag{4.11}$$

where B denotes a Bernoulli random variable with parameter 0.9, G is a Gumbel distribution with distribution function $F(x) = \exp(-\exp(-x)); x \in \mathbb{R}$ and γ denotes Euler's constant. Note that the variance of the error is (approximately) 4 in all cases under consideration. The parameter α indicates the null hypothesis ($\alpha = 0$) and alternatives ($\alpha = 1, \tau = 0.25; \alpha = 1, \tau = 1$).

Horowitz and Spokoiny (1999) used a preliminary simulation study to determine the bandwidth h such that the power of Härdle and Mammen's (1993) test becomes maximal and obtained $h = 3.5$. For their test they therefore used a grid of bandwidths $h \in \{2.5, 3.0, \ldots, 4.5\}$ and calculated the maximin of $\bar{T}_{n,h}$ on this grid. The critical values for the corresponding test statistic are obtained by the bootstrap method described in the previous paragraph. Obviously, this choice of the grid is only possible in a simulation study, because in practice the alternative is unknown. The test statistic $\bar{T}_n(c)$ uses bandwidths proportional to the asymptotically optimal bandwidth for the estimation of the regression function over a certain range of c and is therefore less sensitive with respect to the specification of a range for the bandwidths. The results for the simulated power based on 250 simulation runs in the situation considered by Horowitz and Spokoiny (1999) are depicted in Table 4.1. Comparing these results with the corresponding values obtained by Horowitz and Spokoiny (1999) we observe that the test based on the Kolmogorov-Smirnov statistic of the empirical process indexed with bandwidths yields significantly larger rejection probabilities. We note that this improvement is obtained by assuming additional knowledge about the smoothness of the regression function, which allows to consider an empirical process indexed by (asymptotically) optimal bandwidths. On the other hand the procedure proposed by Horowitz and Spokoiny (1999) leaves the problem of choosing the grid for the different bandwidths at least partially open.

n	$\alpha = 0$				$\alpha = 1, \tau = 1$				$\alpha = 1, \tau = 0.25$				
	20 %	10 %	5 %	2.5 %	20 %	10 %	5 %	2.5 %	20 %	10 %	5 %	2.5 %	error
	.195	.098	.047	.028	.989	.957	.931	.898	1.000	1.000	1.000	1.000	(4.9)
250	.181	.092	.043	.020	1.000	.996	.952	.894	1.000	1.000	1.000	1.000	(4.10)
	.196	.104	.052	.021	.980	.956	.946	.898	1.000	1.000	1.000	1.000	(4.11)

Table 4.1: *Rejection probabilities of the bootstrap test based on the statistic T_{KS} defined in (4.4). The regression model is given by (4.8) with $\alpha = 0$ corresponding to the null hypothesis and three error distributions are considered.*

Example 4.2. In our second example we compare the new test based on the empirical process indexed by (asymptotically) optimal bandwidths with the test proposed by Dette (1999) [see also Fan, Zhang and Zhang (2001) and Fan and Li (2002) for a similar proposal]. In his Example 4.1 Dette (1999) considered the model

$$Y_i = 5X_i + aX_i^2 + \varepsilon_i \ , \quad i = 1, \ldots n \tag{4.12}$$

where X follows a uniform distribution on the interval $[0, 1]$, ε has a centered normal distribution with variance $\sigma^2 > 0$. The case $a = 0$ corresponds to the null hypothesis (straightline through the origin) and two alternatives are considered. In Table 4.2 we display the level and power of the new bootstrap test for sample sizes $n = 50$ and $n = 100$ with a 2.5%, 5%, 10% and 20% level. These results are based on 1000 simulation runs and for the 5% level directly comparable with the values of Table 1 in Dette (1999).

n		50				100			
σ^2	a	20%	10%	5%	2.5%	20%	10%	5%	2.5%
	0	.216	.117	.061	.034	.187	.095	.056	.030
1	1	.296	.190	.114	.066	.355	.237	.150	.091
	2	.530	.381	.264	.171	.696	.554	.442	.329
	0	.193	.107	.055	.024	.214	.105	.058	.037
2	1	.254	.145	.081	.045	.306	.195	.127	.072
	2	.365	.229	.145	.089	.485	.350	.248	.169
	0	.184	.092	.039	.024	.200	.106	.061	.023
3	1	.210	.120	.058	.024	.266	.147	.091	.061
	2	.302	.182	.118	.071	.428	.296	.208	.116

Table 4.2: *Rejection probabilities of the bootstrap test based on the statistic T_{KS} defined in (4.4). The regression model is given by (4.12) and the case $a = 0$ corresponds to the null hypothesis of a straightline through the origin.*

We observe a good approximation of the nominal level ($a = 0$), while the power increases with an increasing parameter a and decreases with an increasing variance σ^2. Comparing the power (for the nominal level of 5%) with the corresponding results of Table 1 in Dette (1999) we observe substantial improvements. For example, in the case $a = 2$ the new test has approximately 50% more power than the test proposed by Dette (1999), and the superiority in the other cases is very similar.

5 Conclusion

In this paper a method is proposed to choose the bandwidth for goodness-of-fit tests for model specification in nonparametric regression models which are based on kernel methods. The main idea is to consider the corresponding test statistic as a stochastic process indexed by a bandwidth proportional to the asymptotically optimal bandwidth for the estimation of the regression function. Weak convergence of the corresponding process is established, which allows to consider Kolmogorov-Smirnov or Cramér von Mises type statistics of the corresponding processes. On the one hand this approach leads to a procedure which is less sensitive with respect to the choice of a smoothing parameter, on the other hand the methodology usually increases the power of the kernel based methods because optimal bandwidths for regression estimation are used. The finite sample properties of a bootstrap version of the corresponding test are investigated by means of a small simulation study and compared with some of the currently available procedures. It is observed empirically that the new method yields significantly larger power in the considered examples.

Although the present paper deals with the specific problem of testing for the parametric form of regression function, the basic methodology is broadly applicable to all goodness-of-fit testing problems, where kernel based methods are involved. This includes such important problems of testing for additivity [see e.g. Gozalo and Linton (2001)], testing for omitted variables and semiparametric functional forms [see e.g. Fan and Li (1996)] or for symmetry of the error distribution [see Ahmad and Li (1997) or Zheng (1998b)].

Acknowledgements. The authors would like to thank Isolde Gottschlich, who typed parts of the paper with considerable technical expertise The work of the first author was supported by the Deutsche Forschungsgemeinschaft (SFB 475, Komplexitätsreduktion in multivariaten Datenstrukturen).

References

Ahmad, I.A., Li, Q. (1997). Testing symmetry of an unknown density function by kernel method. Nonparam. Statistics 7, 279-293.

Alcalá, J.T., Christóbal, J.A. , González-Manteiga, W. (1999). Goodness-of-fit test for linear models based on local polynomials. Statist. & Probab. Letters 42, 39-46.

Azzalini, A., Bowman, A. (1993). On the use of nonparametric regression for checking linear relationships. J. Roy. Statist. Soc., Ser. B, 55, 549-559.

Bickel, P., Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. Ann. Statist. 1, 1071-1095.

Billingsley, P. (1968). Convergence of Probability Measures. Wiley, New York-London-Sydney-Toronto.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. Ann. Statist. 27, 1012-1050.

Fan, J., and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications, London: Chapman and Hall.

Fan, Y., Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric function forms. Econometrica 64, 865-890.

Fan, J., Zhang, C.M. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. Ann. Statist. 29, 153-193.

Fan, Y., Li, Q. (2002). A consistent model specification test based on the kernel sum of squares of residuals. Econometric Reviews 21, 337-352.

González-Manteiga, W., Cao, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. Test 2, 161-188.

Gozalo, P.L. and O.B.Linton (2001) A nonparametric test of additivity in generalized nonparametric regression with estimated parameters. J. Econometrics, 104, 1-48.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. Ann. Statist. 21, 1926-1947.

Horowitz, J.L. and Spokoiny, V. (1999). An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. Econometrica 69, 599-631.

Weirather, G. (1993). Testing a linear regression model against nonparametric alternatives. Metrika 40, 367-379.

Wooldridge, J.M. (1992). A test for functional form against nonparametric alternatives. Econometric Theory, 8, 452-475.

Yachew, A.J. (1992). Nonparametric regression tests based on least squares. Econometric Theory, 8, 435-451.

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

Zheng, J.X. (1998a). A consistent nonparametric test of parametric regression models under conditional quanile restriction. Econometric Theory, 14, 263-289.

Zheng, J.X. (1998b). Consistent specification testing for conditional symmetry. Econometric Theory, 14, 139-149.

Zhang, C., Dette, H. (2003). A power comparison between nonparametric regression test. Statist. 8, Probability Letters 66, 289-301.