Rate-optimal data-driven specification testing in regression models

Emmanuel Guerre, LSTA Paris 6

Pascal Lavergne, INRA-ESR

February 2001

Abstract

We propose a general procedure for testing that a regression function has a prescribed parametric form. We allow for multivariate regressors, non-normal errors and heteroscedasticity of unknown form. The test relies upon a nonparametric linear estimation method, such as a sieves expansion or the kernel method. The choice of the smoothing parameter is data-driven. Under the null hypothesis, the asymptotic distribution of the test statistic is the standard normal distribution. Use of bootstrap critical values is formally justified. The test is shown to be adaptive and rate-optimal in the minimax sense. Detection of Pitman-type local alternatives is also studied.

Keywords: hypothesis testing, nonparametric adaptive tests, selection methods.

AMS classification: Primary 62G10 ; Secondary 62G08.

Financial support from INRA and LSTA is gratefully acknowledged.

Corresponding address: INRA-ESR, BP 27, 31326 CASTANET-TOLOSAN Cedex FRANCE. E-mail addresses: eguerre@ccr.jussieu.fr, lavergne@toulouse.inra.fr.

1 Introduction

In recent years, a substantial amount of statistical work has focused on the search for omnibus specification tests (or lack-of-fit tests) of a parametric regression model, i.e, tests that are consistent against a large spectrum of nonparametric alternatives. Extensions to the regression setting of popular Kolmogorov-Smirnov and Cramer-von-Mises tests for a parametric density have been proposed by Andrews (1997), Buckley (1991), Bierens (1982), Bierens and Ploberger (1997), Delgado (1993), Hall and Hart (1990) and Stute (1997), among others. However, it is likely that such tests suffer from the same drawback than their analogs in the density setting, namely they cannot detect some deviations form the null hypothesis with substantial frequency, see e.g. Inglot, Kallenberg and Ledwina (1997) and the references therein. Another large strand of research relies upon a nonparametric smooth estimator that is compared to the parametric fit. Tests have been developed for specific choices of the smoothing parameter by Barry and Hartigan (1990), Chen (1994), Cleveland and Devlin (1988), Dette (1999), Ellison and Fisher Ellison (2000), Eubank and Spiegelman (1990), Eubank and Hart (1993), Guerre and Lavergne (1999), Härdle and Mammen (1993), Hong and White (1995), Li and Wang (1998) and Zheng (1996) among others. Further work acknowledges that the smoothing parameter should be data-driven. Three main methods have been proposed. The first one uses a cross-validated smoothing parameter that minimizes an estimated risk function similar to the one used in estimation, see Yanagimoto and Yanagimoto (1987) and Hart and Wehrly (1992). The second method chooses the smoothing parameter that maximizes a test statistic over a set whose cardinality increases with sample size, see Baraud, Huet and Laurent (1999), Fan and Huang (2000), Horowitz and Spokoiny (2001) and Spokoiny (1999), see also Spokoiny (1996) on the continuous-time Gaussian model. Finally, a third proposal is to use a penalization approach, which selects the model dimension through minimization of an information criterion similar to the well-known AIC (Akaike, 1974) or BIC (Schwarz, 1978) criteria, see Aerts, Claeskens and Hart (1999 and 2000), Eubank and Hart (1992). A similar method is developed by Ledwina (1994) and Inglot, Kallenberg and Ledwina (1997) for testing for a parametric density. Discussion on these competing approaches as well as other references on the topic can be found in Hart's (1997) monograph.

Some of these selection procedures are consistent under the null hypothesis, i.e. they asymptotically selects a specific smoothing parameter when the parametric model is true. While such a property is crucial in a model selection context, it is not necessary to build a consistent test. For instance, Aerts, Claskens and Hart (2000) and Eubank and Hart (1992) adopt a AIC-type selection method and derive the asymptotic distribution of the related test statistics. Horowitz and Spokoiny (2001) and Spokoiny (1999) use the second selection method and show that, while the asymptotic distribution of their test statistic is unknown, it can be approximated through simulations. However, a selection procedure that is consistent under the null hypothesis should lead to a test with a more tractable and robust asymptotic distribution. The behavior of the selection procedure under the alternative hypothesis, i.e. when the parametric model is misspecified, is more difficult to apprehend. The first method, based upon the cross-validated smoothing parameter, should asymptotically select the optimal parameter for estimation. Eubank and Hart (1992) suggest that their AIC-type criterion behaves similarly. However, the selected parameter may be not optimal for the testing problem. For instance, following Ingster (1993), Guerre and Lavergne (1999) show that the optimal smoothing parameter that allows uniform detection of smooth alternatives is different from the optimal one for estimation, as the former is determined by a different bias-variance trade-off. From the previously mentioned authors, only Baraud, Huet and Laurent (1999), Fan and Huang (2000), Horowitz and Spokoiny (2001) and Spokoiny (1996 and 1999) show that their tests are adaptive and rate-optimal in the minimax sense, that is, they detect classes of alternatives of unknown smoothness that approaches the null hypothesis at the fastest possible rate.

In this paper, we propose a new smooth test for testing the validity of a parametric regression model, that differs from previously proposed ones for it applies in quite general settings. Our framework permits multivariate regressors, nonlinear parametric models, non-normality of the errors and heteroscedasticity of unknown form. It also allows for the use of many different nonparametric methods such as sieves methods, e.g. regression on algebraic, trigonometric functions or piecewise polynomial functions, or kernel-based smoothing methods. The smoothing parameter is selected through a penalization approach applied to an increasing number of statis-

tics. Our selection procedure and the resulting test has several characteristic features. First, we use centered test statistics as estimates of the mean squared distance of the unknown regression function from the null hypothesis. Second, our penalization approach relies on a diverging sequence γ_n times an estimate of the statistics' standard deviations. This is at odds with previous proposals that use estimates of the statistics' variances in the penalization scheme, see for instance Hart (1997) for the AIC and BIC-based penalization approaches. Third, our selection procedure is consistent under the null hypothesis for a penalization sequence γ_n which is of smaller order than the one used in BIC-type procedures. Fourth, the resulting test is adaptive and rate-optimal, which has not yet been shown to hold for other penalization-based tests.

The paper is organized as follows. In Section 2, we explain the general idea of our testing procedure. In Section 3, we detail our assumptions and state our main results. We prove that (i) our test statistic has asymptotically a standard normal distribution (ii) our test is adaptive and rate-optimal in the minimax sense (ii) it is consistent against Pitman-type local alternatives of unknown smoothness that approaches the null hypothesis at a rate close to $1/\sqrt{n}$, where n is the sample size. We also show that different nonparametric methods can be used in the testing procedure. In Section 4, we detail some points of practical importance for implementation. We show how a bootstrap method can be used to derive critical values for our test. We then discuss the performances of our test compared to the test proposed by Horowitz and Spokoiny (2001). The relative advantages of our procedure is shown both theoretically and through some Monte-Carlo experiments.

The following conventions and notations are used throughout the paper. We denote any vector $[U_1, \ldots, U_n]'$ by U. The linear space $\mathbb{R}^n$ is equipped with the usual Euclidean norm, i.e. $\|U\| = (U'U)^{1/2}$. We also define the norm $|U| = \max_{i=1,\ldots,n} |U_i|$. The notation $\delta(\cdot)$ stands for a function while δ or $\delta(X)$ stands for the vector $[\delta(X_1), \ldots, \delta(X_n)]'$, where the X_i's are the regressors' observations. For any two real positive sequences a_n and b_n, $a_n \asymp b_n$ means that a_n and b_n have the same order, that is, $c.a_n \leq b_n \leq C.a_n$ for some finite and strictly positive constants c and C and n large enough. The symbol "$\asymp_{\mathbb{P}}$" is the stochastic generalization of the preceding notation. Lastly, all probabilities and expectations are conditional upon $X_1, \ldots, X_n$.

2 The testing procedure

Assume we are given n independent observations (Y_i, X_i) with $Y_i = m(X_i)+\varepsilon_i$ where the X_i's are in $\mathbb{R}^p$, $m(\cdot)$ is an unknown regression function from $\mathbb{R}^p$ to $\mathbb{R}$ and ε_i a centered regression error. To acknowledge that the (conditional) distribution of the Y_i's depend upon $m(\cdot)$, we denote probabilities, expectation and variance by $\mathbb{P}_m$, $\mathbb{E}_m$ and Var_m respectively. The null hypothesis of interest is

$$H_0 \ : \ m(.) \in \mathcal{M} = \{\mu(\cdot,\theta); \theta \in \Theta\} \ , \ \Theta \subset \mathbb{R}^d \ .$$

We consider a parametric estimator $\widehat{\theta}_n$ of θ_m, with the following properties. Under H_0, i.e. $m(\cdot) = \mu(\cdot,\theta_0)$, $\theta_m = \theta_0$. Under the alternative, $\delta_m(\cdot) = m(\cdot) - \mu(\cdot,\theta_m)$ characterizes the deviation from the null hypothesis, in a sense we define precisely in Section 3. Let $\widehat{U_i} = Y_i - \mu(X_i,\widehat{\theta}_n), i = 1,\ldots,n$, be the estimated residuals and $\widehat{U}$ be the vector with generic element $\widehat{U}_i$. For $h \in \mathcal{H}_n = \{h_0 > h_1 > \ldots > h_{D_n} > 0\}$, consider $n \times n$ matrices $W_h = [w_{ij}, i, j = 1,\ldots,n]$ that depend upon $X = (X_1,\ldots,X_n)$ and a smoothing parameter h and whose diagonal elements are all zero. We define

$$\widehat{T}_h = \widehat{U}' W_h \widehat{U} = \sum_{1\le i,j\le n} w_{ij}(h)\widehat{U}_i\widehat{U}_j \ . \tag{2.1}$$

Because $\widehat{U}' W_h \widehat{U} = \widehat{U}' \frac{W_h + W_h'}{2} \widehat{U}$, we can assume without loss of generality that W_h is symmetric.

The rationale for considering statistics of the form $\widehat{T}_h$ is as follows. If the actual U_i's, $U_i = Y_i - \mu(X_i,\theta_m) = \delta_m(X_i) + \varepsilon_i$, were known, we could estimate $\delta_m(X) = (\delta_m(X_1),\ldots,\delta_m(X_n))'$ by some nonparametric "leave-one-out" linear estimates of the form $\widehat{\delta}_m(X) = W_h U$. Hence

$$T_h = U' W_h U = \sum_{i=1}^n \delta_m(X_i)\widehat{\delta}_m(X_i) + \sum_{i=1}^n \varepsilon_i \widehat{\delta}_m(X_i) \ , \tag{2.2}$$

would be an estimator of $\|\delta_m\|^2 = \sum_{i=1}^n \delta_m^2(X_i)$, since the last term has zero expectation, for W_h has zeros on its diagonal. The statistic $\widehat{T}_h$ is expected to be asymptotically equivalent to T_h, and thus should be a valid statistic to evaluate the distance of the unknown regression function to the parametric model under consideration.

To construct a test, we choose among the statistics $\widehat{T}_h$, $h \in \mathcal{H}_n$, or equivalently among the

h's. We consider a selection procedure via penalization, i.e.,

$$\tilde{h} = \arg\max_{h \in \mathcal{H}_n} \left\{ \widehat{T}_h - \widehat{T}_{h_0} - \gamma_n \widehat{v}_{h,h_0} \right\} \ , \ \gamma_n > 0 \ , \tag{2.3}$$

where $\widehat{v}^2_{h,h_0}$, $h \in \mathcal{H}_n \setminus \{h_0\}$, is an estimator of the variance of $\widehat{T}_h - \widehat{T}_{h_0}$ to be defined in Section 3.5, and $\widehat{v}_{h_0,h_0} = 0$. The test t_n rejects H_0 if $\widehat{T}_{\tilde{h}}/\widehat{v}_{h_0}$, where $\widehat{v}_{h_0}$ is an estimate of the variance of T_{h_0}, is large. Specifically,

$$t_n = \mathbb{1}\left(\widehat{T}_{\tilde{h}} \geq \widehat{v}_{h_0} z_\alpha \right), \text{ with } \mathbb{P}(N(0,1) \leq z_\alpha) = 1 - \alpha \ . \tag{2.4}$$

The choice of the critical region of the test comes from two facts: (i) the selection procedure is consistent under H_0, that is $\widetilde{h} = h_0$ with probability approaching one under the null when γ_n diverges fast enough (ii) the test statistic $\widehat{T}_{h_0}$ is close to the centered quadratic form $T_{h_0} = U' W_{h_0} U$, which converges to a normal distribution. This simple characterization contrasts with results obtained with other selection procedures, e.g. by Baraud, Huet and Laurent (1999), Eubank and Hart (1992), Horowitz and Spokoiny (2001) and Spokoiny (1999), in which the asymptotic distribution of the test statistic under the null hypothesis is more involved. In the above critical region, $\widehat{v}_{h_0}$ can be changed into $\widehat{v}_{\tilde{h}}$, where $\widehat{v}^2_{\tilde{h}}$ is an estimate of the variance of the $T_{\tilde{h}}$. This is the usual normalization adopted for instance by Baraud, Huet, and Laurent (1999), Fan and Huang (2000), Horowitz and Spokoiny (2001) and Spokoiny (1999). However, as explained below, the variance of T_h is expected to decrease with h, so that standardizing by $\widehat{v}^2_{h_0}$ should lead to a larger rejection region and then to a more powerful test. Such a standardization together with our penalization scheme are the main characteristics of our test.

Many nonparametric methods are linear in observations. We now detail some key examples that will be subsequently studied in Section 4.

Example 1. Regression-based methods. A family of nonparametric estimation methods relies upon the expansion of the function of interest into a linear combination $\sum_{k=1}^{\infty} \beta_k \psi_k(\cdot)$ of functions $\psi_k(\cdot)$, $k = 1, \ldots$. The sum is truncated to some large order $K(h)$ and the corresponding β_k's are estimated by ordinary least squares. Such an approach is widely used in practice especially when specific alternatives are in mind, corresponding to an appropriate choice

and/or ordering of the functions $\psi_k(\cdot)$. (Clearly, one should avoid to include approximating functions that are already included in the parametric model $\mathcal{M}$.) If Ψ_h is the $n \times K(h)$ matrix $\Psi_h = [\psi_k(X), 1 \leq k \leq K(h)]$ and

$$P_h = \Psi_h(\Psi_h'\Psi_h)^{-1}\Psi_h'$$

is the projection matrix associated to Ψ_h, then W_h can be obtained from P_h by removing diagonal elements. Thus, W_h corresponds to the estimator that takes into account all observations X_i's but leaves out the i-th observation of the dependent variable. Standard leave-one-out estimators could also be used, but our definition of W_h is more convenient both for theoretical study and practical use. Note that $\widehat{U}'P_h\widehat{U}$ is easily obtained from a linear regression of $\widehat{U}$ on Ψ_h, and $\widehat{T}_h$ is obtained from the former expression by removing terms that involve the squared $\widehat{U}_i$'s.

Popular choice for the $\psi_k(\cdot)$ are trigonometric, polynomial or exponential functions, see e.g. Aerts, Claeskens and Hart (1999), Baraud, Huet and Laurent (1999), Eubank and Hart (1993), Hong and White (1995). In our analysis, we shall consider the following specific examples.

- **Example 1a: Multivariate polynomial functions**: $\psi_k(x) = \prod_{\ell=1}^p x_i^{k_i}$, for $k \in I\!N^p$ with $|k| = \max_{l=1,\ldots,p} k_l \leq 1/h$, assuming $1/h$ is an integer number. The dimension of the linear subset of $I\!R^n$ spanned by the $\psi_k(X)$'s is proportional to h^{-p} almost surely if the joint distribution of the X_i's is continuous and h does not go too fast to 0. Moreover, because the images of the P_h's are nested subspaces of $I\!R^n$, $\mathrm{Rank}(P_h - P_{h_0}) = \mathrm{Rank}(P_h) - \mathrm{Rank}(P_{h_0}) \asymp h^{-p} - h_0^{-p}$. Under suitable conditions, the variances of the quadratic forms T_h and $T_h - T_{h_0}$ should also be of order h^{-p} and $h^{-p} - h_0^{-p}$ respectively.

- **Example 1b: Trigonometric functions**: $\psi_k(\cdot) = \prod_{\ell=1}^p \cos^{\epsilon_\ell}(2\pi k_\ell x_\ell)\sin^{1-\epsilon_\ell}(2\pi k_\ell x_\ell)$, for $\epsilon_\ell \in \{0,1\}$ and $k \in I\!N^p$ with $|k| \leq 1/h$, where $1/h$ is an integer number. As in Example 1a, $\mathrm{Rank}(P_h) \asymp h^{-p}$ and $\mathrm{Rank}(P_h - P_{h_0}) \asymp h^{-p} - h_0^{-p}$, and under suitable conditions, similar conclusions hold for the variances of T_h and $T_h - T_{h_0}$, respectively.

- **Example 1c: Piecewise polynomial functions**: Assume that the support of the X_i's is a compact set, chosen as $[0,1]^p$ without loss of generality. For $h > 0$, define $I_k(h) =$

$\prod_{\ell=1}^{p}[(k_\ell - 1)h, k_\ell h)$ for $k \in I\!N^p$. The piecewise polynomial functions are defined as

$$\psi_{qk}(x) = \prod_{\ell=1}^{p} x_i^{q_i} \, 1\!\!1(x \in I_k(h)) \ , \ 0 \leq |q| \leq \bar{q}, \ |k| \leq 1/h \ ,$$

and P_h is the orthogonal projection matrix on the subspace of $I\!R^n$ defined by the $\psi_{qk}(X)$'s. The rank of P_h is $(\bar{q}+1)^p$ times the number of non-empty bins $I_k(h)$ as soon as the X_i's are continuously distributed with density bounded away from zero. This gives again $\text{Rank}(P_h) \asymp h^{-p}$ if h does not go too fast to 0. One can also obtain a nested structure for the P_h's, as in Examples 1a and 1b, by considering dyadic bandwidths $\mathcal{H}_n = \{h_0 = 2^{-j_0}, \ldots, h_{D_n} = 2^{-j_{D_n}}\}$. This will be assumed to simplify further analysis. In such a case, the test statistics are analogous to the ones derived from a wavelet expansion, see Härdle, Kerkyacharian, Picard, and Tsybakov (1998). This is clear for $\bar{q} = 0$, which corresponds to an expansion with respect to the Haar basis, a simple example of wavelet systems used by Spokoiny (1999), and is similar to the regressogram method considered by Guerre and Lavergne (1999).

Other instances of linear nonparametric methods fall into this framework. This is the case for least-squares splines with equidistant knots (see for instance Eubank, 1988), that are close in spirit to our Example 1c. The formal analysis of such an extension is left for further research.

Example 2. Kernel smoothing. When the X_i's are i.i.d. with density $f(\cdot)$, a popular alternative to projection-based techniques is the kernel method. Consider $K(\cdot)$ from $I\!R^p$ to $I\!R$ and define $K_h(x) = K(x_1/h, \ldots, x_p/h)$, $h > 0$. Ellison and Ellison (2000) study the case where

$$W_h = \left[\frac{K_h(X_i - X_j)}{nh^p \widehat{f}_h(X_i)} \, 1\!\!1(i \neq j) \right]_{1 \leq i,j \leq n} \text{ with } \ \widehat{f}_h(X_i) = \frac{1}{nh^p} \sum_{\ell=1, \ell \neq i}^{n} K_h(X_i - X_\ell) \ .$$

The leave-one-out estimate $\widehat{f}_h(X_i)$ of $f(X_i)$ can be replaced by $f(X_i)$ if $f(\cdot)$ is known. The test statistics used by Horowitz and Spokoiny (2001) are of the form $\widehat{U}' P_h' P_h \widehat{U}$, where

$$P_h = \left[\frac{K_h(X_i - X_j)}{nh^p \widetilde{f}_h(X_i)} \right]_{1 \leq i,j \leq n} \text{ with } \widetilde{f}_h(X_i) = \frac{1}{nh^p} \sum_{\ell=1}^{n} K_h(X_i - X_\ell) \ .$$

Many smoothing procedure are asymptotically equivalent to kernel regression estimation for an i.i.d. design. In particular, Fan and Gijbels (1996) mention that the local polynomial estimator

is asymptotically equivalent at first order to a kernel estimator with $\widehat{f}(\cdot)$ changed into $f(\cdot)$, suggesting that our results could apply to local polynomial estimation. Equivalence results between kernel estimators and nearest neighbor or smoothing spline estimators can be found in Härdle (1990) and the references therein, but require to consider a local design-dependent bandwidth. The study of such extensions are left for further research.

3 Asymptotic analysis

In this section, we first state our main assumptions. Then we determine the asymptotic behavior of the test under the null hypothesis. We subsequently study its behavior under two kinds of alternatives with unknown smoothness. In the first case, corresponding to the so-called minimax adaptive framework considered by Spokoiny (1996) among others, the alternatives are at a distance $\rho_n(s)$ from the null, where s is an unknown smoothness index. In the second case, we consider Pitman-type local alternatives of unknown smoothness. Then we propose some variance estimators for implementing the procedure. We conclude by giving some sufficient conditions under which regression-based methods and kernel smoothing methods can be used.

3.1 Assumptions

3.1.1 Model and estimation

The basic assumption on the data generating process is the following one.

Assumption E *The ε_i's are independent, and*

$$\forall i = 1, \ldots, n, \quad I\!E(\varepsilon_i) = 0 \ , \ 0 < \underline{\sigma}^2 \leq I\!E(\varepsilon_i^2) = \sigma^2(X_i) < \overline{\sigma}^2 < \infty \ , \ I\!E|\varepsilon_i^3| \leq \nu_3 < \infty \ .$$

For any $m(\cdot)$, not necessarily in H_0, let

$$\theta_m \equiv \theta_{m,n} = \arg \inf_{\theta \in \Theta} \|m(X) - \mu(X, \theta)\| \quad \text{and} \quad \delta_m(X) = m(X) - \mu(X, \theta_m) \ .$$

The pseudo-true value θ_m could be defined through any other norm, as for instance $I\!E(\mu(X, \theta) - m(X))^2$ if the X_i's were independently distributed as $X \in I\!R^p$. We shall assume that $\widehat{\theta}_n$, the

estimator in the parametric model, converges to the pseudo-true value θ_m under misspecification. Following Ingster (1993), we restrict our analysis to smooth regression functions. For $s \in [0,1)$, let $C_p(L,s)$ be the set of measurable functions $\delta(\cdot)$ such that

$$|\delta(x) - \delta(y)| \leq L\|x-y\|^s \quad \forall x, y \ .$$

For $s \geq 1$, let $[s]$ be the greatest integer less than or equal to s and let $C_p(L,s)$ be the set of functions $\delta(\cdot)$ almost everywhere differentiable up to order $[s]$, whose all derivatives of order $[s]$ belong to $C_p(L, s-[s])$. We consider the following admissible deviations from the null hypothesis

$$H: \ m(\cdot) \text{ with } \delta_m(\cdot) \in C_p(L,s) \text{ for some } (s,L) \in [\underline{s}, \overline{s}] \times [\underline{L}, \overline{L}] \ . \tag{3.5}$$

The two following assumptions deal with θ_m, its estimator $\widehat{\theta}_n$ and regularity conditions on the parametric model.

Assumption M1 *i. For any $m(\cdot)$, θ_m is unique.*

ii. Under Assumption E, the estimator $\widehat{\theta}_n$ is such that $\sqrt{n}(\widehat{\theta}_n - \theta_m) = O_{I\!P_m}(1)$ uniformly in $m(\cdot) \in H$, i.e.,

$$\forall \eta > 0, \exists \nu > 0 : \limsup_{n \to +\infty} \sup_{m(\cdot) \in H} I\!P_m \left(\sqrt{n} \|\widehat{\theta}_n - \theta_m\| > \nu \right) \leq \eta \ .$$

Assumption M2 *$\mu(\cdot,\cdot)$ is uniformly bounded in x and θ, is twice continuously differentiable with respect to θ, with first and second order derivatives $\mu_\theta(\cdot,\cdot)$ and $\mu_{\theta\theta}(\cdot,\cdot)$ uniformly bounded in x and $\theta \in \Theta$.*

When the parametric model is correctly specified, Assumption M1–*i* is an identification condition because θ_m is then the true value of the parameter. Under misspecification, the pseudo-true value θ_m of the parameter should be the probability limit of the estimator. This is the case when $\widehat{\theta}_n$ is the nonlinear least-squares estimator or some pseudo maximum-likelihood estimator, see White (1982) and Gourieroux, Monfort and Trognon (1984). The estimator has to be uniformly consistent with respect to the functions $m(\cdot) \in H$, as uniformity is essential for our Theorems 1 to 4 below. Birgé and Massart (1993) have shown that such a property usually holds for approximate nonlinear least-squares estimators. Assumption M2 provides regularity conditions on the parametric model $\mathcal{M}$, which are usually necessary for Assumption M1 to hold.

3.1.2 Matrix assumptions

Let us first introduce some notations. For any $n \times n$ matrix $W = [w_{ij}, i, j = 1, \ldots, n]$, let

$$\nu(W) = \left(\sum_{i,j} w_{ij}^2 \right)^{1/2} \quad \text{and} \quad r(W) = \sup_{\|u\|=1} \|Wu\| \ .$$

For any real symmetric W, $\nu(W) = \mathrm{tr}^{1/2}(W^2)$ is the Euclidean norm of the vector formed with the eigenvalues of W, and the spectral radius $r(W)$ is the largest eigenvalue in absolute value.

Assumption W *i. For each $h \in \mathcal{H}_n$, W_h is real symmetric with $w_{ii}(h) = 0, \ \forall i = 1, \ldots, n$.*

ii. $\max_{h \in \mathcal{H}_n} r(W_h) = O(1)$.

iii. $\nu^2[W_h] \asymp h^{-p}$ *and* $\nu^2[W_h - W_{h_0}] \asymp h^{-p} - h_0^{-p}$ *uniformly in* $h \in \mathcal{H}_n$.

It is immediate to see that Assumption W–*ii* and *iii* hold for the projection matrix P_h of Example 1, as $\nu^2(P_h) = \mathrm{Rank}(P_h)$ and $r(P_h) = 1$. It will remain to show that these conditions are also fulfilled when the diagonal elements of P_h are set to zero. Our discussion of regression-based methods suggests that the variance of the statistic T_h is related to $\nu^2(W_h)$, as formally shown in Lemma 1. The assumption on the spectral radius of $r(W_h)$ is used jointly with $\nu(W_h)$ to bound moments of higher order, see Section 3.6.

The next assumption is used to study $I\!E_m T_h$ under the alternative in connection with Lemma 3, which generalizes some results of Ingster (1993).

Assumption B *i. There exist some real positive semidefinite matrices P_h such that* $\max_{h \in \mathcal{H}_n} r(P_h - W_h) = o(1)$.

ii. For each $\underline{s} \leq s \leq \overline{s}$, there exist some sets of approximating functions $\Pi_{s,h}$ and $C_0 \geq 0$ such that, for each $\delta(\cdot) \in C_p(L, s)$, $\exists \ \pi_\delta(\cdot)$ in $\Pi_{s,h}$ with

$$\|\delta(x) - \pi_\delta(x)\| \leq C_0 L \sqrt{n} h^s \ .$$

iii. $\Lambda_n = \inf_{\underline{s} \leq s \leq \overline{s}} \inf_{h \in \mathcal{H}_n} \inf_{\pi(\cdot) \in \Pi_{s,h}} (\pi' P_h \pi)^{1/2} (\pi' \pi)^{-1/2}$ *is bounded away from 0.*

Assumption B–*i* ensures that the matrices W_h's are well approximated by positive semidefinite matrices P_h's, such as the projection matrices of Example 1. Assumptions B–*ii* and *iii* imply that the matrices P_h's are well behaved with respect to some family of approximating functions of smoothness index $s \in [\underline{s}, \overline{s}]$. In Examples 1a and 1b, it is easily seen from the multivariate Jackson Theorem (see Lorentz, 1966) that the latter conditions are fulfilled, where $\Pi_{s,h} = \Pi_h$ is the set of algebraic or trigonometric polynomial functions up to degree h^{-p} and $\Lambda_n = 1$.

We now introduce conditions on the variance estimators.

Assumption V *i.* $\widehat{v}_{h_0}^2 \asymp_{I\!P_m} h_0^{-p}$ *uniformly in* $m(\cdot) \in H$. $\widehat{v}_{h,h_0}^2 \asymp_{I\!P_m} h^{-p} - h_0^{-p}$ *uniformly in* $m(\cdot) \in H$ *and in* $h \in \mathcal{H}_n$.

ii. $\widehat{v}_{h_0}^2 = \mathrm{Var}_m[T_{h_0}](1 + o_{I\!P_m}(1))$ *uniformly in* $m(\cdot) \in H_0$. $\widehat{v}_{h,h_0}^2 = \mathrm{Var}_m\left[T_h - T_{h_0}\right](1 + o_{I\!P_m}(1))$ *uniformly in* $m(\cdot) \in H_0$ *and in* $h \in \mathcal{H}_n$.

As discussed in the context of Example 1, such properties are expected to hold from Assumption W–*iii*. In Section 3.5, we shall propose some variance estimators that fulfill the conditions in Assumption V.

3.2 Study of the test under the null hypothesis

A specific aspect of our selection procedure is that the penalization term involves the estimated standard deviations $\widehat{v}_{h,h_0}$ of the statistics $\widehat{T}_h - \widehat{T}_{h_0}$. This prevents us to use proof techniques reviewed in Hart (1997), where the penalization is based on the variances. Theorem 1 below shows that our test is of asymptotic level α as soon as γ_n diverges at rate $\sqrt{\ln D_n}$. When D_n is proportional to its maximum value n in Example 1, γ_n is of order $\sqrt{\ln n}$, which is smaller than the order $\ln n$ used in the BIC criterion used by Aerts, Claskens and Hart (2000), Hart (1997) and Ledwina (1994).

Theorem 1 *Under Assumptions E, M1–M2, V, W, if* $D_n^{-1} = o(1)$, $D_n \left(h_1^{-p} - h_0^{-p}\right)^{-1/2} = o(1)$, $\gamma_n \geq (1+\eta)\sqrt{2 \ln D_n}$ *for some* $\eta > 0$ *and* $h_0 = o(1)$, *then*

$$\sup_{m(\cdot) \in H_0} I\!P_m \left(t_n = 1\right) \to \alpha \ .$$

Proof of Theorem 1. Proposition 5–i gives

$$\begin{aligned} I\!P_m(\tilde{h} \neq h_0) &= I\!P_m\left[\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left(\widehat{T}_h - \widehat{T}_{h_0} - \gamma_n\widehat{v}_{h,h_0}\right) > 0\right] \\ &\leq I\!P_m\left[\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{\widehat{T}_h - \widehat{T}_{h_0}}{\widehat{v}_{h,h_0}}\right| \geq \gamma_n\right] = o(1)\ . \end{aligned}$$

Thus the asymptotic distribution of the test statistic is given by $\widehat{T}_{h_0}/\widehat{v}_{h_0}$. Now Proposition 5–$ii$ yields the desired result. □

Theorem 1 imposes the central requirement that $h_0 = o(1)$. This condition may not seem intuitive, as one may want to allow for a constant bandwidth h_0, that is for a finite number of functions in Example 1, in the spirit of Ramsey's (1969) RESET test. Though a constant bandwidth would be expected under H_0 in a nonparametric estimation context, it is not clear that this is desirable in a testing problem. Moreover, the bandwidth which is selected when the null hypothesis is not rejected is of no use for further analysis. It is only when the parametric model is rejected that the selected bandwidth can convey some information about the form of the regression function. Lastly, when h_0 is constant, the parametric estimator generally affects the behavior of the test statistic under H_0. This is noted by Hart (1997) in the case of nonlinear models. The condition $h_0 = o(1)$ aims to ensure that the asymptotic distribution of $\widehat{T}_{h_0}$ is the one of T_{h_0}, which is then a normal one, without further restrictions on the parametric model or the nonparametric method used to build the test. To see what happens when h_0 is constant, consider the simple setup where $\mu(\cdot,\theta) = \theta$, $\widehat{\theta}_n = (1/n)\sum_{i=1}^n Y_i$ and W_{h_0} has all non-diagonal elements equal to $(1/n)$ (which corresponds to the regressogram estimator of Example 1c with $h_0 = 1$). Then $\widehat{T}_{h_0} = \frac{1}{n}\sum_{i\neq j}(Y_i - \widehat{\theta}_n)(Y_j - \widehat{\theta}_n) = -\frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\theta}_n)^2$ converges in probability to minus the unconditional variance of Y and clearly behaves differently from $T_{h_0} = \frac{1}{n}\sum_{i\neq j}\varepsilon_i\varepsilon_j$, which is asymptotically $\chi^2_1 - 1$ (up to a proper normalization) under H_0. The assumption $h_0 = o(1)$ can however be avoided in particular situations, cf. Guerre and Lieberman (2000) for a test of no effect with regression-based methods, and Hart (1997) for testing linear models in a different penalization context.

3.3 Detection of a class of alternatives with unknown smoothness

Consider a function $\rho(\cdot) > 0$, and let $H_1(\rho(\cdot))$ be the set of regression functions $m(\cdot)$ with deviation from the null hypothesis $\delta_m(\cdot)$ of unknown smoothness s in $[\underline{s}, \overline{s}]$ and distant of $\sqrt{n}\rho(s)$ from H_0, i.e.,

$$H_1(\rho(\cdot)): \ m(\cdot) \text{ with } \delta_m(\cdot) \in C_p(L,s) \text{ for some } (s,L) \in [\underline{s},\overline{s}] \times [\underline{L},\overline{L}] \text{ and } \|\delta_m\| \geq \sqrt{n}\rho(s) \ . \tag{3.6}$$

That the nuisance smoothness s is unknown is referred in the literature as the adaptive framework. The next theorem gives conditions on $\rho_n(\cdot)$, γ_n and $\mathcal{H}_n$ under which our test is uniformly consistent against any alternative in $H_1(\rho(\cdot))$.

Theorem 2 *If for each $s \in [\underline{s}, \overline{s}]$, there is a sequence $h_n(s)$ in $\mathcal{H}_h$ with*

$$h_n(s) \asymp (\rho_n(s))^{\frac{1}{s}} \quad \text{where} \quad \rho_n(s) = \left(\frac{\gamma_n}{n}\right)^{\frac{2s}{4s+p}} \ , \tag{3.7}$$

uniformly in s, and if $\gamma_n \to \infty$, then, under Assumptions E, M1–M2, V, W, B,

$$\sup_{m(\cdot) \in H_1(\kappa\rho_n(\cdot))} I\!P_m\left(t_n = 0\right) \to 0 \quad \text{for some } \kappa > 0 \ .$$

As our test has a vanishing minimax type-II error against alternatives with regularity s lying at distance $\rho_n(s)$ from the parametric model, it is uniformly consistent against such alternatives. By mimicking the proof of Theorem 2, one can obtain a similar result for tests based upon a variance penalization of the form $\gamma_n \widehat{v}^2_{h,h_0}$, as in Hart (1997) and Ledwina (1994), with the lower rate $(\gamma_n/n)^{s/(2s+p)}$.

Because the rate $\rho_n(s)$ increases with the penalization γ_n, the best rate is obtained by choosing the smallest possible γ_n such that the asymptotic level is under control. From Theorem 1, this is obtained by choosing $\gamma_n^2 \asymp \ln D_n$, and a rate-optimal test is such that D_n is as small as possible. A simple set $\mathcal{H}_n$ can be constructed as follows. Let

$$s_d = \overline{s} + \frac{d}{D_n}\left(\underline{s} - \overline{s}\right), \quad d = 1, \ldots, D_n$$

and $\mathcal{H}_n = \{h_0, h_n(s_1), \ldots, h_n(s_{D_n})\}$, where $h_n(s)$ is defined by (3.7). Then the central condition of Theorem 2 is fulfilled if and only if $D_n \asymp \ln n$, and the smallest alternatives $H_1(\rho_n(s))$ detected by our test are of order

$$\rho_n(s) \asymp \left(\frac{\sqrt{\ln \ln n}}{n}\right)^{\frac{2s}{4s+p}} .$$

This rate corresponds to the optimal adaptive minimax rate found by Spokoiny (1996) in the continuous-time Gaussian model. This is also the rate achieved by the tests of Baraud, Huet and Laurent (1999), Horowitz and Spokoiny (2001) and Spokoiny (1999). Another choice of the admissible bandwidths, which does not depend upon $\underline{s}$ and $\overline{s}$, is given by $h_d = 2^{-(d_n+d-1)}$, $d = 1, \ldots, D_n$ where d_n diverges, see Spokoiny (1999). For such a choice, it is possible to take a diverging upper bound of the smoothness index $\overline{s} = \overline{s}_n$ in Theorem 2 with $h_n(\overline{s}_n) \asymp h_1 = 2^{-d_n}$. In any case, it is not necessary to know the constants $\underline{L}$ and $\overline{L}$.

Proof of Theorem 2. In what follows, C_i, $i = 1, \ldots, 6$, denote positive constants. Consider $\delta(\cdot)$ in $C_p(L, s)$ for some fixed $s \in [\underline{s}, \overline{s}]$ and $L \in [\underline{L}, \overline{L}]$, and let $\rho_n = \rho_n(s)$ and $h_n = h_n(s)$ as defined by (3.7). Definitions (2.3) of $\tilde{h}$ and (2.4) of t_n give, for any $m(\cdot)$,

$$I\!P_m(t_n = 0) \leq I\!P_m\left(\max_{h \in \mathcal{H}_n}\left(\widehat{T}_h - \gamma_n \widehat{v}_{h,h_0}\right) \leq z_\alpha \widehat{v}_{h_0}\right) \leq I\!P_m\left(\widehat{T}_{h_n} \leq \gamma_n \widehat{v}_{h_n,h_0} + z_\alpha \widehat{v}_{h_0}\right) .$$

By Assumption V-i and since $\sqrt{h_n^{-p} - h_0^{-p}} \leq h_n^{-p/2}$ and $h_0^{-p/2} \leq h_n^{-p/2}$,

$$I\!P_m(t_n = 0) \leq I\!P_m\left(\widehat{T}_{h_n} \leq C_1 \gamma_n h_n^{-p/2}\right) + o(1) ,$$

where the reminder term is uniform with respect to $m(\cdot) \in H$ and $h_n \in \mathcal{H}_n$. Therefore, Proposition 4 and the Tchebychev inequality yield

$$\begin{aligned} I\!P_m(t_n = 0) &\leq I\!P_m\left(T_{h_n} \leq C_1 \gamma_n h_n^{-p/2} + C_2\|\delta\|\right) + o(1) \\ &= I\!P_m\left[I\!E_m\left(T_{h_n}\right) - T_{h_n} \geq I\!E_m\left(T_{h_n}\right) - C_1 \gamma_n h_n^{-p/2} - C_2\|\delta\|\right] + o(1) \\ &\leq \frac{\mathrm{Var}_m[T_{h_n}]}{\left[I\!E_m\left(T_{h_n}\right) - C_1 \gamma_n h_n^{-p/2} - C_2\|\delta\|\right]^2} + o(1) , \end{aligned}$$

where the remainder term is uniform with respect to $m(\cdot) \in H$ and $h_n \in \mathcal{H}_n$, provided

$$I\!E_m\left(T_{h_n}\right) - C_1 \gamma_n h_n^{-p/2} - C_2\|\delta\| > 0 .$$

By Assumption B, Lemmas 1 and 3 yield that there is a $\kappa > 0$ such that, for any $m(\cdot)$ in $H_1(\kappa\rho_n)$ and n large enough,

$$\begin{aligned}\frac{I\!E_m\left(T_{h_n}\right) - C_1\gamma_n h_n^{-p/2} - C_2\|\delta\|}{\|\delta\|^2} &\geq C_3\left[1 - C_4\frac{\sqrt{n}h_n^s}{\|\delta\|}\right]^2 - C_1\frac{\gamma_n}{h_n^{p/2}\|\delta\|^2} - C_2\frac{1}{\|\delta\|} \\ &\geq C_3\left[1 - C_4\frac{h_n^s}{\kappa\rho_n}\right]^2 - C_1\frac{\gamma_n}{\kappa^2 n h_n^{p/2}\rho_n^2} - \frac{C_2}{\kappa\sqrt{n}\rho_n}\ . \quad (3.8)\end{aligned}$$

As $\rho_n \asymp h_n^s$, $nh_n^{p/2}\rho_n^2 \asymp \gamma_n$ and $\sqrt{n}\rho_n \to \infty$, this lower bound is increasing in κ and is positive for κ large enough. Also, Lemma 1, Assumptions V and W yield that for n large enough

$$\frac{\mathrm{Var}_m\left(T_{h_n}\right)}{\|\delta\|^4} \leq \frac{2\overline{\sigma}^2\nu^2(W_h)}{\|\delta\|^4} + \frac{4\overline{\sigma}^2 r^2(W_h)\|\delta\|^2}{\|\delta\|^4} \leq \frac{C_5}{h_n^p\|\delta\|^4} + \frac{C_6}{\|\delta\|^2} \leq \frac{C_5}{\kappa^4 n^2 h_n^p \rho_n^4} + \frac{C_6}{\kappa^2 n\rho_n^2}\ ,$$

and this upper bound is decreasing in κ and tends to zero as n increases. □

3.4 Detection of local alternatives of unknown smoothness

As noted by Horowitz and Spokoiny (2001), tests designed to be uniformly consistent against alternatives (3.6) are tailored to detect some extreme alternatives with narrow peaks and valleys whose widths decrease with the sample size. Such tests may be less powerful against alternatives that are less extreme. To provide some protection against this eventuality, we now study the behavior of our test against the hypothesis

$$H_1'(\rho_n):\ m(\cdot) \text{ with } \delta_m(\cdot) = \rho_n d(\cdot), d(\cdot) \in C_p(L,s) \text{ for some } (s,L) \in [\underline{s},\overline{s}]\times[\underline{L},\overline{L}] \text{ and } \|d\| \geq \sqrt{n}\ . \quad (3.9)$$

In contrast to the alternatives considered in Section 3.3, these Pitman-type local alternatives cannot have a too wildly oscillating behavior.

Theorem 3 *Under Assumptions E, M1–M2, V, W, B, and if* $h_0 = o(1)$

$$\sup_{m(\cdot)\in H_1'(\rho_n)} I\!P_m\left(t_n = 0\right) = o(1) \quad \text{whenever } \rho_n^{-1} = o\left(\sqrt{nh_0^{p/2}}\right)\ .$$

Our test is thus uniformly consistent against alternatives in $H_1'(\rho_n)$, as soon as ρ_n goes to zero slower than $\left(nh_0^{p/2}\right)^{-1/2}$. As h_0 can vanish at an arbitrary slow rate, there is little difference between our test and tests designed to be consistent against alternatives in $H_1'(\rho_n)$ with $\rho_n^{-1} = o(\sqrt{n})$, such as the Kolmogorov-Smirnov and Cramer-von-Mises type tests mentioned in the Introduction. Moreover, uniform consistency against $H_1'(\rho_n)$ is not shown for the former tests. Theorem 3 is similar to Theorem 3 of Horowitz and Spokoiny (2001), except that they require $h_0 \asymp (\ln \ln n)^{-2/p}$. Baraud, Huet and Laurent (1999) and Guerre and Lieberman (2000) propose tests that detect alternatives of order $1/\sqrt{n}$, but they restrict to testing for the specification of a linear regression model.

Proof of Theorem 3. Definition (2.3) of $\tilde{h}$ gives

$$\widehat{T}_{\tilde{h}} = \max_{h \in \mathcal{H}_n} \left(\widehat{T}_h - \widehat{T}_{h_0} - \gamma_n \widehat{v}_{h,h_0} \right) + \widehat{T}_{h_0} + \gamma_n \widehat{v}_{\tilde{h},h_0} \geq \widehat{T}_{h_0} \ ,$$

with probability converging to one, as $\gamma_n \widehat{v}_{\tilde{h},h_0} \geq 0$ with probability converging to one by Assumption V–*i*. Thus

$$I\!P_m \left(t_n = 0 \right) \leq I\!P_m \left(\widehat{T}_{h_0} \leq z_\alpha \widehat{v}_{h_0} \right) + o(1) \ ,$$

where the remainder term is uniform in $m(\cdot) \in H$ and $h_n \in \mathcal{H}_n$. The proof then follows from minor modifications of the proof of Theorem 2, changing in particular γ_n into 1 and h_n into h_0. It is also necessary to show that (3.8) holds for all $m(\cdot) \in H_1'(\rho_n)$. But Lemma 3 gives

$$I\!E_m T_{h_0} = \rho_n^2 I\!E d' W_h d \geq C_3 \rho_n^2 \left[\|d\| - C_4 \sqrt{n} h_0^s \right]^2 \ ,$$

provided $\|d\| \geq C_5 \sqrt{n} h_0^s$, which holds for n large enough since $\|d\| \geq \sqrt{n}$ and $h_0 = o(1)$. □

3.5 Variance estimation

Crucial conditions to obtain estimators that fulfill Assumption V are the ones in Assumption W. Indeed, Lemma 1 yields that

$$\mathrm{Var}_m[T_h] = 2 \sum_{1 \leq i,j \leq n} w_{ij}^2(h) \sigma^2(X_i) \sigma^2(X_j) + 4 \sum_{i=1}^n \sigma^2(X_i) \left(\sum_{j=1}^n w_{ij}(h) \delta_m(X_j) \right)^2 \ ,$$

and a similar expression holds for $\mathrm{Var}_m[T_h - T_{h_0}]$, changing $w_{ij}(h)$ into $w_{ij}(h) - w_{ij}(h_0)$. Consider nonparametric estimators $\widehat{\sigma}_n^2(\cdot)$ of $\sigma^2(\cdot)$. Variance estimators can then be defined as

$$\widehat{v}_{h_0}^2 = 2\sum_{i,j} w_{ij}^2(h_0)\widehat{\sigma}_n^2(X_i)\widehat{\sigma}_n^2(X_j) \;, \tag{3.10}$$

$$\widehat{v}_{h,h_0}^2 = 2\sum_{i,j} \left(w_{ij}(h) - w_{ij}(h_0)\right)^2 \widehat{\sigma}_n^2(X_i)\widehat{\sigma}_n^2(X_j) \;,\; \forall\, h \in \mathcal{H}_n \setminus \{h_0\} \;. \tag{3.11}$$

The next proposition shows that such estimators can be used in our procedure. Horowitz and Spokoiny (2001) and Spokoiny (1999) propose to use similar estimators, but impose stronger conditions than we do.

Proposition 1 *Under Assumptions E, W-iii, if*

$$\sup_{1\le i\le n} |\widehat{\sigma}_n^2(X_i) - \sigma^2(X_i)| = o_{I\!P_m}(1) \;, \tag{3.12}$$

uniformly with respect to $m(\cdot) \in H$, then the estimators (3.10) and (3.11) fulfill Assumption V.

Under regularity conditions on $\sigma^2(\cdot)$ and $m(\cdot)$, it is not difficult to find an uniformly consistent estimator. For instance, one can use

$$\widehat{\sigma}_n^2(X) = W_h Y^2 - (W_h Y)^2 \;,$$

for some $h \in \mathcal{H}_n$, where $Y^2 = [Y_1^2, \ldots, Y_n^2]$. Alternatively, the estimated residuals $\widehat{U}$ can be used instead of Y. Related uniform convergence results are given by Györfi, Härdle, Sarda and Vieu (1989) for kernel estimators and Newey (1997) for series estimators, among others. Estimators proposed by Horowitz and Spokoiny (2001) and Spokoiny (1999) generalize the difference-based estimator proposed by Rice (1984) to functional estimation, assuming that $\sigma^2(\cdot)$ and $I\!E(\varepsilon^4|X = \cdot)$ are Lipschitz functions.

Alternative variance estimators that do not rely upon nonparametric estimation can be defined as

$$\widehat{v}_h^2 = 2\sum_{1\le i,j\le n} w_{ij}^2(h)\widehat{U}_i^2\widehat{U}_j^2 \quad \text{and} \quad \widehat{v}_{h,h_0}^2 = 2\sum_{1\le i,j\le n} (w_{ij}(h) - w_{ij}(h_0))^2\widehat{U}_i^2\widehat{U}_j^2 \;.$$

They are computationally simpler than estimators (3.10) and (3.11), fulfill Assumption V–*ii* and, assuming that $m(\cdot)$ is bounded, also fulfill Assumption V–*i*. However, they will tend to be greater than (3.10) and (3.11) under the alternative hypothesis. It is thus expected that, though asymptotically valid, use of these simpler estimators will lead to some loss of power.

In the case of homoscedastic errors, one can use $\widehat{v}_h^2 = 2\widehat{\sigma}_n^2 \nu^2(W_h)$ and $\widehat{v}_{h,h_0}^2 = 2\widehat{\sigma}_n^2 \nu^2(W_h - W_{h_0})$ in (3.11), $\widehat{\sigma}_n^2$ being one of the several consistent estimators of the variance proposed in the literature, see e.g. Rice (1984) for a difference-based estimator and Silverman (1985) for a estimator based on nonparametric estimation of the regression, see also Lavergne and Vuong (1998) for some more recent references.

3.6 Study of the examples

When studying Examples 1 and 2, we assume

Assumption X *The X_i's are i.i.d. on $[0,1]^p$, with a continuous density $f(\cdot)$ such that $0 < \underline{f} \leq f(x) \leq \overline{f} < \infty,\ \forall x \in [0,1]^p$.*

It is easy to adapt the following propositions to regular deterministic designs such as $X_i = i/n,\ i = 0, \ldots, n$. We first deal with Examples 1a to 1c.

Proposition 2 *Assume that $D_n = O(n)$, $h_1 = o(h_0)$, and $h_{D_n}^{-p} = o\left(\frac{n}{\ln n}\right)^{1/3}$ for Example 1a, $h_{D_n}^{-p} = o\left(\frac{n}{\ln n}\right)^{1/2}$ for Example 1b, or $h_{D_n}^{-p} = o\left(\frac{n}{\ln n}\right)$ for Example 1c. Under Assumption X, Assumptions W and B hold in probability.*

The statement "in probability" means that in Assumptions W and B, $\asymp$, $O(1)$ and $o(1)$ should be changed into $\asymp_{I\!P}$, $O_{I\!P}(1)$ and $o_{I\!P}(1)$ respectively. This is sufficient to establish Theorems 1 to 3. Our condition on h_{D_n} depends upon the considered sieves family. It is less restrictive for Example 1c because the corresponding projection matrices have many 0 coefficients, so that Assumption W is easier to establish. It is more restrictive in Example 1a than in Example 1b because the algebraic polynomial functions are not standardized, and specifically $I\!E\psi_k^2(X)$ vanishes when k grows, so that the associated matrices P_h are more difficult to control. By contrast, Assumption B is a straightforward consequence of the Jackson Theorem (see Lorentz,

1966) for algebraic and trigonometric polynomials, but is less easy to establish for piecewise polynomial functions. Together with (3.7), the conditions in Proposition 2 imply that $\underline{s} > 5p/4$, $\underline{s} > 3p/4$ and $\underline{s} > p/4$ for Example 1a, 1b and 1c respectively. The penalization approach is not tailored to detect alternatives with small regularities, i.e. $\underline{s} \leq p/4$, see Guerre and Lavergne (1999) for some discussion.

We now turn to the kernel smoothing method and we assume

Assumption K *The kernel $K(\cdot)$ is symmetric, integrable with positive integral, continuous and bounded from above. Its Fourier transform $\widehat{K}(x) = \int \exp(-it'x)K(t)dt$ is strictly positive on its nonempty support. Moreover, for any $x \in [0,1]^p$, any $h \in \mathcal{H}_n$ and n large enough,*

$$f_h(x) = \frac{1}{h^p} \int K_h(x' - x) f(x') dx' \geq \underline{f}/2 \; . \tag{3.13}$$

The condition on the Fourier transform $\widehat{K}(\cdot)$ is used to show that Assumption B holds and seems typical to our minimax approach, as it is not usually required, see e.g. Zheng (1996). It clearly holds for the triangular, normal or Laplace kernel, but not for the Epachnikov or uniform kernel. The lower bound on $f_h(\cdot)$ is automatically granted if $K(\cdot)$ is positive on its support. For a general $K(\cdot)$ with positive integral, it is clear that (3.13) holds for $x \in [\eta, 1-\eta]^p$ for some $0 < \eta < 1$. For x close to the frontiers of $[0,1]^p$, it can be established under Assumption X by showing that $h^{-p} \int_{[0,1]^p} K_h(x - x')\, dx'$ as a function of x is uniformly bounded away from 0.

Proposition 3 *Assume that $D_n = O(n)$, $h_1 = o(h_0)$, $h_0 = o(1)$ and $h_{D_n}^{-p} = o\left(\frac{n}{\ln n}\right)$. Under Assumptions X and K, Assumptions W and B hold in probability.*

Propositions 2 and 3 reveal that Example 1c and Example 2 works under similar conditions, which seems sensible as regression on piecewise polynomial functions extends the regressogram method, which is close in spirit to the kernel method.

4 Practical implementation

4.1 Bootstrap critical values

There are several reasons to believe that the normal asymptotic approximation of Theorem 1 can be inaccurate for practical purposes. First, it is well known that the small sample distribution of a quadratic form is close to a mixture of chi-squared distributions. Second, it seems desirable to account for the parametric estimation error in small samples, though it is asymptotically negligible. Third, the variances of $\widehat{T}_{h_0}$ and $\widehat{T}_h - \widehat{T}_{h_0}$ must be estimated, and this may introduce some supplementary variability. Fourth, while the selection procedure selects h_0 with probability converging to one under the null hypothesis, it may select some other smoothing parameters with positive probability for a specific data set that follows the postulated parametric model. Hence, we propose a bootstrap procedure to compute critical values for our test. As Theorems 2 and 3 allow for any bounded z_α, they still hold when using a bootstrap critical value.

The wild bootstrap, initially proposed by Wu (1986), is often used in testing contexts, see e.g. Härdle and Mammen (1993). We will use a generalization of this method, the smooth conditional moments (SCM) bootstrap introduced by Gozalo (1997), that we now detail.

i. Let ε_i^*, $i = 1, \ldots, n$, be realizations of random variables independent given the initial sample $\{(Y_i, X_i), i = 1, \ldots, n\}$ and such that

$$I\!E\varepsilon_i^* = 0 \,, \tag{4.14}$$

$$I\!E\varepsilon_i^{*2} = \widehat{\sigma}_n^2(X_i) \,, \tag{4.15}$$

$$I\!E|\varepsilon_i^{*3}| < \infty \,, \tag{4.16}$$

where $I\!E^*$ denotes expectation given the initial sample.

ii. From the new sample $\left\{X_i, Y_i^* = \mu(X_i, \widehat{\theta}_n) + \varepsilon_i^*, i = 1, \ldots, n\right\}$, compute $\widehat{\theta}_n^*$ and the parametric residuals $\widehat{U}_i^* = Y_i^* - \mu(X_i, \widehat{\theta}_n^*)$. For $h \in \mathcal{H}_n$, compute $\widehat{T}_h^* = \widehat{U}^{*\prime} W_h \widehat{U}^*$ and the variance estimators $\widehat{v}_{h,h_0}^*$, as well as $\widehat{v}_{h_0}^*$.

iii. Choose the smoothing parameter as $\tilde{h}^* = \arg\max_{h\in\mathcal{H}_n} \left\{ \widehat{T}^*_h - \gamma_n \widehat{v}^*_{h,h_0} \right\}$, and compute the test statistic $\widehat{T}^*_{\tilde{h}^*}/\widehat{v}^*_{h_0}$.

The bootstrap critical value $z^*_{\alpha,n} = z^*_{\alpha,n}(m(\cdot))$ at level α is the $1-\alpha$ quantile of $\widehat{T}^*_{\tilde{h}^*}/\widehat{v}^*_{h_0}$, which in practice can be estimated by repeating Steps i to iii many times. This critical value is then compared to the initial test statistic $\widehat{T}_{\tilde{h}}/\widehat{v}_{h_0}$ to perform the test. Gozalo (1997) suggests to match the third-order moment of ε^*_i with a nonparametric estimator of $I\!E\varepsilon_i^3$, in addition to Conditions (4.14)–(4.15), and give examples of simple distribution that satisfy these conditions. Strictly speaking, our conditions are sufficient to obtain valid critical values, but Gozalo's proposal is expected to yield more precise ones. One might also consider matching higher-order moments in Step i.

Let $t^*_n = 1\!\!1\left(\widehat{T}_{\tilde{h}} \geq \widehat{v}_{h_0} z^*_\alpha\right)$. The following theorem shows the validity of this test.

Theorem 4 *Under Assumptions E, M1–M2, V, W, if* $D_n^{-1} = o(1)$, $D_n (1/h_1 - 1/h_0)^{-p/2} = o(1)$, $\gamma_n \geq (1+\eta)\sqrt{2 \ln D_n}$ *for some* $\eta > 0$, $h_0 = o(1)$ *and (3.12), (4.14)–(4.16) hold, then* $z^*_{\alpha,n}$ *converges to* z_α *uniformly in* $m(\cdot) \in H$ *and*

$$\sup_{m(\cdot)\in H_0} I\!P_m (t^*_n = 1) \to \alpha \ .$$

Proof of Theorem 4. Let $\varepsilon^* = [\varepsilon^*_1, \ldots, \varepsilon^*_n]$ and $e^* = [e^*_1, \ldots, e^*_n]$, where $e^*_i = \mu(X_i, \widehat{\theta}_n) - \mu(X_i, \widehat{\theta}^*_n)$. Let $T^*_h = \varepsilon^{*'} W_h \varepsilon^*$, $B^*_h = \varepsilon^{*'} W_h e^*$ and $R^*_h = e^{*'} W_h e^*$. We have $\widehat{T}^*_h = T^*_h + 2B^*_h + R^*_h$. Analogs of Proposition 4 (a)–(d) for R^*_h and B^*_h can be derived conditionally upon the initial sample by mimicking the proof given in Section 5.1, using (3.12) and (4.15). Similarly, a conditional version of Proposition 5 can be derived for $\widehat{T}^*_h$ using (4.14)–(4.16) and Proposition 1. The proof is then similar to Theorem 1's proof. □

4.2 Small sample behavior

We now investigate the small sample behavior of our test. We generate data through the model

$$Y = \theta_1 + \theta_2 X + \delta(X) + \varepsilon, \tag{4.17}$$

where X is distributed as $U[-0.5, 0.5]$ and ε is centered. The null hypothesis corresponds to $\delta(X) \equiv 0$. We consider different forms of alternatives as specified by $\delta(X) = r\cos(2\pi tX)$, so that $I\!E[\delta(X)] = I\!E[X\delta(X)] = 0$ and hence the pseudo-true value θ_m equals $\theta = (\theta_1, \theta_2)$. Such departures from the null are of special interest, as it is known that smooth tests of parametric specifications are sensitive to the frequency of the alternatives, see e.g. Eubank and Hart (1993) and Hart (1997).

We consider one of the simplest version of our test, corresponding to the regressogram method, cf. Example 1c. Our Monte-Carlo study also aims to compare our test with the one proposed by Horowitz and Spokoiny (2001, hereafter HS). Their test statistic is based upon the kernel method. The test by Fan and Huang (2000) also shares some similarities with HS test, but relies upon Fourier series. To avoid discrepancies related to the nonparametric estimation technique, we consider the test statistic based upon the regressogram method

$$\frac{\widehat{T}_{h_{\max}}}{\widehat{v}_{h_{\max}}} = \max_{h \in \mathcal{H}_n} \frac{\widehat{T}_h}{\widehat{v}_h} ,$$

where $\widehat{v}_h$ is defined similarly as $\widehat{v}_{h_0}$ in (3.10). To compute critical values of their test, Horowitz and Spokoiny (2001) use a bootstrap analog of the original test statistic, where the bootstrap samples are generated through the estimated parametric model with ε_i^* following a $N\left(0, \widehat{\sigma}_n^2(X_i)\right)$ distribution. In our experiment, we use bootstrap samples where the ε_i^*'s follow a two-point distribution with mean zero and variance $\widehat{\sigma}_n^2(X_i)$ to compute critical values for the different tests. Specifically, this distribution is given by

$$I\!P^*\left(\varepsilon_i^* = \frac{1-\sqrt{5}}{2}\widehat{\sigma}_n(X_i)\right) = \frac{5+\sqrt{5}}{10} \ , \ I\!P^*\left(\varepsilon_i^* = \frac{1+\sqrt{5}}{2}\widehat{\sigma}_n(X_i)\right) = \frac{5-\sqrt{5}}{10} \ , \ i = 1, \ldots, n \ ,$$

where $I\!P^*$ is the conditional distribution of the independent ε_i^*'s given the initial sample. Therefore, the only differences between tests are the selection procedure and the normalization, chosen as $\widehat{v}_{h_0}$ for our test and $\widehat{v}_{h_{\max}}$ for HS test.

We consider a small sample size of 50 observations and we run 5000 replications for each considered case. For each replication, the number of bootstrap samples is 200. The design is fixed across replications. We assume homoscedasticity and estimate the variance through Rice's

(1984) estimator

$$\widehat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n} \left(Y_{(i+1)} - Y_{(i)} \right)^2 ,$$

where the $Y_{(i)}$'s denote observations ordered in accordance with the ordered values of the X_i's. The set of binwidths is chosen as $\{h_0 = 2^{-2}, h_1 = 2^{-3}, \ldots, h_4 = 2^{-6}\}$. The penalization γ_n is set to $c\sqrt{2\ln(D_n + 1)}$ (here $D_n = 4$), where c varies into the set $\{1, 1.5, 2\}$.

In a first set of simulations, we set $\theta = (0, 0)$ and $\mathcal{M} = \{0\}$, to evaluate the tests' performances in the absence of parametric estimation error. The ε_i's are independent from the X_i's and follow a centered normal distribution with standard deviation 1/3. We choose $r = 1/3$, so that $I\!E\delta^2(X)/I\!E\varepsilon^2 = 1/2$ for integer t. Table 1 reports our Monte-Carlo results for the null hypothesis and alternatives of varying frequencies with $t = 2, 5, 10$. In each cell, the first, second and third row give empirical percentages of rejections at a 2%, 5% and 10% nominal level, respectively. Under the null hypothesis, our bootstrap method leads to accurate rejection probabilities in small samples for our test at any considered value of c, and for HS test, though no theoretical study justifies its validity in this case. Under the considered alternatives, all tests perform quite well given the small sample size. In most cases, the empirical power of our test does not vary significantly with c. As expected, power decreases for both our test and HS test when the frequency of the alternative increases. Our test always dominates HS test, that is, the empirical power of our test is significantly greater than the power of HS test in any case. The difference can be as high as 9%.

In a second set of simulations, we set $\theta = (1, 2)$ and estimate it by ordinary least-squares. Tables 2 to 4 have the same structure as Table 1. We consider three different setups where the errors (independent of the X_i's) are normal (Table 2), exponential (Table 3) or Student with five degrees of freedom (Table 4), and are centered and normalized to have a standard deviation equal to 1/3. The part of the variance of Y explained in the linear regression on X is then moderate, i.e. 75%. Results are quite comparable across cases. Under the null hypothesis, critical values are less precisely estimated than in Table 1. In particular, for Student errors, the empirical levels are always lower than the nominal ones, though the differences are not significant in most cases. It is likely that matching higher-order moments in the bootstrap samples generation would lead

to more accurate critical values. However, there is no significant difference between rejection probabilities for our test and HS test. As in Table 1, in most cases the choice of c does not affect significantly the performances of our test under the null or the alternative hypothesis. Under the lowest frequency alternative, power is quite high for all tests and, though our test leads to higher rejection probabilities, differences are not significant. Significant discrepancies in favor of our test are found for the moderate frequency alternative, in all cases at a 2% nominal level and at other nominal levels when γ_n is not too large. For the highest frequency alternative, our test significantly outperforms HS test in the case of normal and Student errors. The same feature appears in the case of exponential errors at 2% and 5% nominal levels.

That our test exhibits better power performances than HS test follows from the two distinctive features of our method, namely the selection procedure and the normalization by $\widehat{v}_{h_0}$. To formally compare our test and HS test, let

$$HS = \frac{\widehat{T}_{h_{\max}}}{\gamma_n \widehat{v}_{h_{\max}}} .$$

Normalization by γ_n does not affect the test, but ensures that the test statistic is bounded in probability under the null hypothesis and diverges to $+\infty$ under any fixed alternative, as shown by Fan and Huang (2000) and Horowitz and Spokoiny (2001) in different contexts. Definition (2.3) of $\tilde{h}$ yields

$$\frac{\widehat{T}_{\tilde{h}}}{\widehat{v}_{h_0}} = \frac{\widehat{T}_{\tilde{h}} - \gamma_n \widehat{v}_{\tilde{h},h_0} + \gamma_n \widehat{v}_{\tilde{h},h_0}}{\widehat{v}_{h_0}} \geq \frac{\widehat{T}_{h_{\max}} - \gamma_n \widehat{v}_{h_{\max},h_0}}{\widehat{v}_{h_0}} = \gamma_n \frac{\widehat{v}_{h_{\max}}}{\widehat{v}_{h_0}} \left(HS - \frac{\widehat{v}_{h_{\max},h_0}}{\widehat{v}_{h_{\max}}} \right) ,$$

with probability converging to 1, since the estimated standard deviations are asymptotically positive under Assumption V-i. Assume now that $\widehat{v}_h^2 \asymp_{I\!P_m} h^{-p}$ uniformly in h and $m(\cdot) \in H$, which holds for the chosen estimator under Assumptions E, W–*iii* and (3.12). This gives $\widehat{v}_{h_{\max},h_0}/\widehat{v}_{h_{\max}} = O_{I\!P_m}(1)$. As HS diverges to infinity under a fixed alternative, $HS \geq \widehat{v}_{h_{\max},h_0}/2\widehat{v}_{h_{\max}}$ with probability converging to 1. Therefore, for some constant $C > 0$,

$$\frac{\widehat{T}_{\tilde{h}}}{\widehat{v}_{h_0}} \geq C\gamma_n \left(\frac{h_0}{h_{\max}} \right)^{p/2} HS ,$$

with probability converging to 1. This shows that our test statistic diverges much faster than HS, as γ_n diverges and $h_{\max} \leq h_0$. Thus our test has greater asymptotic power than HS test

under any fixed alternative. Our simulations results indicate that such a property is already found in small samples.

5 Proofs

Recall that $U_i = Y_i - \mu(X_i, \theta_m)$, $U = (U_1, \ldots, U_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, $\delta = \delta_m(X) = (\delta_m(X_1), \ldots, \delta_m(X_n))'$, $\sigma^2 = \big(\sigma^2(X_1), \ldots, \sigma^2(X_n)\big)$, and let $e(X_i) = \mu(X, \theta_m) - \mu(X_i, \widehat{\theta})$ and $e = (e(X_1), \ldots, e(X_n))'$. For any $u \in I\!R^n$, $D(u)$ denotes the $n \times n$ diagonal matrix with diagonal u. Finally, C denotes a positive constant that may vary from line to line.

5.1 Intermediate results for Theorems 1 to 4

Let $T_h = U'W_hU$, $A_h = \delta'W_he$, $B_h = \varepsilon'W_he$, $R_h = e'W_he$. As $\widehat{U} = U + e = \delta + \varepsilon + e$,

$$\widehat{T}_h = T_h + 2(A_h + B_h) + R_h \ , \tag{5.18}$$

Note that $A_h = 0$ under H_0 for any h.

Proposition 4 *Under Assumptions E, M1–M2, W–i and ii, then*

$$\text{(a) } h^{p/2}B_h = o_{I\!P_m}(1) \ , \ \forall h \to 0 \qquad \text{(b) } \max_{h \in \mathcal{H}_n \setminus \{h_0\}} \left|(h^{-p} - h_0^{-p})^{-1/2}(B_h - B_{h_0})\right| = o_{I\!P_m}(1) \ ,$$

$$\text{(c) } h^{p/2}R_h = o_{I\!P_m}(1) \ , \ \forall h \to 0 \qquad \text{(d) } \max_{h \in \mathcal{H}_n \setminus \{h_0\}} \left|(h^{-p} - h_0^{-p})^{-1/2}(R_h - R_{h_0})\right| = o_{I\!P_m}(1) \ ,$$

uniformly for $m(\cdot) \in H$, if $D_n \left(h_1^{-p} - h_0^{-p}\right)^{-1} = o(1)$ for (b) and (d). Moreover, for any $h \to 0$, $A_h = O_{I\!P_m}(\|\delta\|)$ uniformly for $m(\cdot) \in H$.

Proof of Proposition 4. We limit to the case of an unidimensional θ, the extension to a multidimensional one is straightforward. We first prove (b), (a) is similarly shown. Under Assumptions M1 and M2, a standard Taylor expansion yields

$$e(X_i) = (\theta_m - \widehat{\theta})\mu_1(X_i) + (\theta_m - \widehat{\theta})^2\mu_2(X_i) \ , \tag{5.19}$$

where $\mu_1(X_i) = \mu_\theta(X_i, \theta_m)$ depends only on X_i and $\mu_2(X_i)$ depends on X_i and $\widehat{\theta}_n$. Therefore $B_h - B_{h_0} = (\theta_m - \widehat{\theta})B_{1h} + (\theta_m - \widehat{\theta})^2 B_{2h}$, where $B_{1h} = \varepsilon'(W_h - W_{h_0})\mu_1$ and $B_{2h} = \varepsilon'(W_h - W_{h_0})\mu_2$. Now

$$I\!E_m(B_{1h}^2) = \mu_1'(W_h - W_{h_0})D(\sigma^2)(W_h - W_{h_0})\mu_1 \leq \overline{\sigma}^2\|\mu_1\|^2 r^2(W_h - W_{h_0}) \leq n\overline{\sigma}^2\overline{\mu}_1^2 r^2(W_h - W_{h_0}) \ ,$$

by Assumptions E and M2, where $|\mu_1(\cdot,\cdot)| \leq \overline{\mu}_1 < \infty$. Also, $I\!E_m(|B_{2h}|) \leq \sqrt{n}\overline{\sigma}\ \overline{\mu}_2 r(W_h - W_{h_0})$, where $|\mu_2(\cdot,\cdot)| \leq \overline{\mu}_2 < \infty$. Now

$$\begin{aligned} & I\!P_m\left[\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|(h^{-p}-h_0^{-p})^{-1/2}B_{1h}\right| \geq t\left(D_n n(h_1^{-p}-h_0^{-p})^{-1}\right)^{1/2}\right] \\ \leq & \sum_{h\in\mathcal{H}_n\setminus\{h_0\}} I\!P_m\left[\left|(h^{-p}-h_0^{-p})^{-1/2}B_{1h}\right| \geq t\left(D_n n(h_1^{-p}-h_0^{-p})^{-1}\right)^{1/2}\right] \\ \leq & \frac{\overline{\sigma}^2\overline{\mu}_1}{t^2 D_n n(h_1^{-p}-h_0^{-p})^{-1}} \sum_{h\in\mathcal{H}_n\setminus\{h_0\}} n(h^{-p}-h_0^{-p})^{-1} r^2(W_h) \leq \frac{C\overline{\sigma}^2\overline{\mu}_1}{t^2}\ , \end{aligned}$$

so that $\max_{h\in\mathcal{H}_n\setminus\{h_0\}}|(h^{-p}-h_0^{-p})^{-1/2}B_{1h}| = O_{I\!P_m}\left(D_n n(h_1^{-p}-h_0^{-p})^{-1}\right)^{1/2}$ uniformly in $m(\cdot)$. Similarly, $\max_{h\in\mathcal{H}_n\setminus\{h_0\}}|(h^{-p}-h_0^{-p})^{-1/2}B_{2h}| = O_{I\!P_m}\left(D_n\sqrt{n}(h_1^{-p}-h_0^{-p})^{-1/2}\right)$ uniformly in $m(\cdot)$, and then, from Assumption M1, we have uniformly in $m(\cdot)$

$$\max_{h\in\mathcal{H}_n\setminus\{h_0\}}|(h^{-p}-h_0^{-p})^{-1/2}B(h)| = O_{I\!P_m}\left(D_n(h_1^{-p}-h_0^{-p})^{-1}\right)^{1/2} = o_{I\!P_m}(1)\ .$$

We now prove (c), (d) is similarly shown. For any h, $|R_h| \leq r^2(W_h)\|e\|^2$ and by Assumptions M1 and M2, $\|e\|^2 = \sum_{i=1}^n e_i^2 = O_{I\!P_m}(1)$. Hence, $|h^{p/2}R_h| = O_{I\!P_m}(h^{p/2}) = o_{I\!P_m}(1)$ if $h = o(1)$.
From (5.19), we have $A_h = (\theta_m - \widehat{\theta})A_{1h} + (\theta_m - \widehat{\theta})^2 A_{2h}$, where

$$\begin{aligned} |A_{1h}| &= |\mu_1' W_h\delta| \leq \overline{\mu}_1\|\delta\| r(W_h)\ , \\ |A_{2h}| &= |\mu_2' W_h\delta| \leq \overline{\mu}_2\|\delta\| r(W_h)\ , \end{aligned}$$

and Assumption M1 implies the last result. □

Proposition 5 *Under Assumptions E, M1–M2, V and W, if* $D_n^{-1} = o(1)$ *and* $D_n\left(h_1^{-p}-h_0^{-p}\right)^{-1/2} = o(1)$*, then under* H_0*,*

i. For any sequence $\gamma_n \geq (1+\eta)\sqrt{2\ln D_n}$*,* $\eta > 0$*, we have, uniformly in* $m(\cdot)$*,*

$$I\!P_m\left(\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{\widehat{T}_h - \widehat{T}_{h_0}}{\widehat{v}_{h,h_0}}\right| \geq \gamma_n\right) = o(1)\ .$$

ii. If $h_0 = o(1)$*,* $\widehat{T}_{h_0}/\widehat{v}_{h_0}$ *converges in distribution to a* $N(0,1)$ *uniformly in* $m(\cdot)$*.*

Proof of Proposition 5. Part *i.* Lemma 1 implies that, uniformly with respect to $m(\cdot)$ in H_0

$$\mathrm{Var}_m\left[T_h - T_{h_0}\right] = 2\nu^2\left[D(\sigma)\left(W_h - W_{h_0}\right)D(\sigma)\right] \asymp (1/h - 1/h_0)^p\ ,$$

by Assumptions E and W–*iii*. Then, by Proposition 4 and Assumption V, for any $\eta > 0$,

$$\begin{aligned} I\!P_m\left(\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{\widehat{T}_h-\widehat{T}_{h_0}}{\widehat{v}_{h,h_0}}\right|\geq\gamma_n\right) &= I\!P_m\left(\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{T_h-T_{h_0}+o_{I\!P_m}((h^{-p}-h_0^{-p})^{1/2})}{\mathrm{Var}_m^{1/2}[T_h-T_{h_0}](1+o_{I\!P_m}(1))}\right|\geq\gamma_n\right) \\ &\leq I\!P_m\left(\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{T_h-T_{h_0}}{\mathrm{Var}_m^{1/2}[T_h-T_{h_0}]}\right|\geq\frac{\gamma_n}{1+\eta/2}\right)+o(1)\ , \end{aligned}$$

uniformly with respect to $m(\cdot)\in H_0$, since γ_n diverges. Now, Lemma 2–*i*, Assumptions E, V, W give that for n large enough,

$$\begin{aligned} & I\!P_m\left(\max_{h\in\mathcal{H}_n\setminus\{h_0\}}\left|\frac{T_h-T_{h_0}}{\mathrm{Var}_m^{1/2}[T_h-T_{h_0}]}\right|\geq\frac{\gamma_n}{1+\eta/2}\right) \\ \leq & \sum_{h\in\mathcal{H}_n\setminus\{h_0\}} I\!P_m\left(\left|\frac{T_h-T_{h_0}}{\mathrm{Var}_m^{1/2}[T_h-T_{h_0}]}\right|\geq\sqrt{2\ln D_n}+\eta\right) \\ \leq & D_n\frac{\sqrt{2}}{\sqrt{\pi}\sqrt{2\ln D_n}}\exp\left(-\ln D_n\right)+C\max_{h\in\mathcal{H}_n}r(W_h)\sum_{h\in\mathcal{H}_n\setminus\{h_0\}}\frac{1}{\nu\left(W_h-W_{h_0}\right)} \\ = & \frac{1}{\sqrt{\pi\ln D_n}}+O(D_n\left(h_1^{-p}-h_0^{-p}\right)^{-1/2})=o(1)\ . \end{aligned}$$

Part *ii*. By Lemma 1, under H_0, $\mathrm{Var}_m T_{h_0}=2\nu^2\left[D(\sigma)W_{h_0}D\left(\sigma\right)\right]\asymp h_0^{-p}$, by Assumptions E and W-*iii*. Hence, $T_{h_0}/\mathrm{Var}_{H_0}T_{h_0}$ is asymptotically $N(0,1)$ by Lemma 2-*ii* and Assumption V. Proposition 4 and Assumption V allow to conclude. □

5.2 Proof of Proposition 1

We prove the result for $\widehat{v}^2_{h,h_0}$, $h\in\mathcal{H}_n\setminus\{h_0\}$, the proof being similar for $\widehat{v}^2_{h_0}$.
Let $\Omega_h=\left[\omega_{ij}(h)=\left(w_{ij}(h)-w_{ij}(h_0)\right)^2,\ i,j=1,\ldots,n\right]$, $\widehat{\sigma}^2_n=[\widehat{\sigma}^2_n(X_i),\ i=1,\ldots,n]$. Define $v^2_{h,h_0}=2\sigma^{2'}\Omega_h\sigma^2$. By Lemma 1, $v^2_{h,h_0}=\mathrm{Var}_m\left[T_h-T_{h_0}\right]$ uniformly in $m(\cdot)\in H_0$ and then $v^2_{h,h_0}\asymp(h^{-p}-h_0^{-p}$ uniformly in $m(\cdot)\in H$ and $h\in\mathcal{H}_h$ from Assumptions E and W-*iii*. Since

$$\begin{aligned} \widehat{\sigma}^2_n(X_i)\widehat{\sigma}^2_n(X_j)-\sigma^2(X_i)\sigma^2(X_j) &= \left[\widehat{\sigma}^2_n(X_i)-\sigma^2(X_i)\right]\left[\widehat{\sigma}^2_n(X_j)-\sigma^2(X_j)\right]+\sigma^2(X_i)\left[\widehat{\sigma}^2_n(X_j)-\sigma^2(X_j)\right] \\ &\quad +\sigma^2(X_j)\left[\widehat{\sigma}^2_n(X_i)-\sigma^2(X_i)\right]=o_{I\!P_m}(1)\ , \end{aligned}$$

using Assumption E and that $\widehat{\sigma}^2_n$ is uniformly consistent, we get

$$\widehat{v}^2_{h,h_0}-v^2_{h,h_0}=2\sum_{i,j}\omega_{ij}(h)\left[\widehat{\sigma}^2_n(X_i)\widehat{\sigma}^2_n(X_j)-\sigma^2(X_i)\sigma^2(X_j)\right]=o_{I\!P_m}(1)\nu^2\left[W_h-W_{h_0}\right]=o_{I\!P_m}\left(h^{-p}-h_0^{-p}\right)\ ,$$

uniformly in $h\in\mathcal{H}_n$, using Assumption W-*iii*. This shows that $\widehat{v}^2_{h,h_0}/v^2_{h,h_0}$ converges to 1 in probability under W-*iii* uniformly in $m(\cdot)$ and $h\in\mathcal{H}_n$. □

5.3 Study of the examples

Proof of Proposition 2. We first show that Assumption W holds. The triangular inequality yields

$$r(W_h) \leq r(W_h - P_h) + r(P_h) \ .$$

Moreover,

$$\begin{aligned} \nu^2(W_h) &= \nu^2(P_h) - \nu^2(W_h - P_h) \\ \nu^2(W_h - W_{h_0}) &= \nu^2(P_h - P_{h_0}) - \nu^2\left(W_h - W_{h_0} - (P_h - P_{h_0})\right) \ , \end{aligned}$$

As the P_h's are orthogonal projection matrices on linear subspaces of $\mathbb{R}^n$, $r(P_h) = 1$, $\mathrm{Rank}(P_h) = \nu^2(P_h) \asymp 1/h^p$ and, because the image of P_{h_0} is a linear subspace of the image of P_h, $\mathrm{Rank}(P_h - P_{h_0}) = \nu^2(P_h - P_{h_0}) \asymp h^{-p} - h_0^{-p}$. Such properties hold almost surely in Examples 1a and 1b, and almost surely for n large enough under the condition $h_{D_n}^{-p} = o(n/\ln n)$ in Example 1b, which ensures that the number of X_i's in each bins $I_k(h)$ diverge. Thus, it remains to check that

$$\left.\begin{aligned} \max_{h\in\mathcal{H}_n} r(W_h - P_h) &= o_{I\!P}(1) \ , \\ \max_{h\in\mathcal{H}_h} h^p \nu^2(W_h - P_h) &= o_{I\!P}(1) \ , \\ \max_{h\in\mathcal{H}_n\setminus\{h_0\}} \left(h^{-p} - h_0^{-p}\right)^{-1} \nu^2\left(W_h - W_{h_0} - (P_h - P_{h_0})\right) &= o_{I\!P}(1) \ . \end{aligned}\right\} \tag{5.20}$$

Observe that the functions $\psi_k(\cdot)$ can be changed into an orthonormal system of $L_2([0,1]^p, dx)$ to compute the P_h's. This leads to consider the functions

$$\left.\begin{aligned} \phi_k(x) &= \prod_{\ell=1}^{p} \sqrt{2k_\ell + 1} Q_{k_\ell}(x_\ell) \mathbb{1}(x \in [0,1]^p) && \text{for Example 1a,} \\ \phi_k(x) &= \tfrac{1}{2^{p/2}} \prod_{\ell=1}^{p} \cos^{\epsilon_\ell}(2\pi k_\ell x_\ell) \sin^{1-\epsilon_\ell}(2\pi k_\ell x_\ell) && \text{for Example 1b,} \\ \phi_{qk}(x) &= h^{p/2} \prod_{\ell=1}^{p} \sqrt{2k_\ell + 1} Q_{q_\ell}(k_\ell h - x_\ell) \mathbb{1}(x \in I_k(h)) && \text{for Example 1c,} \end{aligned}\right\} \tag{5.21}$$

where the $Q_k(\cdot)$'s are the Legendre polynomials of degree k on $[0,1]$, with $\sup_{t\in[0,1]} |Q_k(t)| \leq 1$, $\int_0^1 Q_k^2(t)dt = 1/(2k+1)$, $\int_0^1 Q_k(t)Q_{k'}(t)dt = 0$ for $k \neq k'$, see e.g. Davis (1975). If Φ_h is the $n \times K(h)$ matrix $\Phi_h = [\phi_k(X), 1 \leq |k| \leq K(h)]$ (alternatively the $n \times (\bar{q}+1)K(h)$ matrix $\Phi_h = [\phi_{kq}(X), 1, \leq |q| \leq \bar{q}, 1 \leq |k| \leq K(h)]$ in Example 1c) , we have $P_h = \Phi_h(\Phi_h'\Phi_h)^{-1}\Phi_h'$. Lemma 4 shows that the eigenvalues of $n^{-1}\Phi_{h_{D_n}}'\Phi_{h_{D_n}}$ are in $[1/C, C]$ for some $C > 1$ with probability tending to 1. As the symmetric matrices $n^{-1}\Phi_h'\Phi_h$ are North West sub-matrices of $n^{-1}\Phi_{h_{D_n}}'\Phi_{h_{D_n}}$, this implies that uniformly in $h \in \mathcal{H}_n$

$$\frac{1}{C.n}\Phi_h\Phi_h' \prec P_h = \frac{1}{n}\Phi_h\left(\frac{1}{n}\Phi_h'\Psi_h\right)^{-1}\Phi_h' \prec \frac{C}{n}\Phi_h\Phi_h' \ ,$$

with probability tending to 1, where $\prec$ is the ordering of symmetric matrices. Hence,

$$\begin{aligned}
|p_{ii}(h)| &\leq \frac{C}{n} \sum_{|k|\leq 1/h} \phi_k^2(X_i) \leq \frac{C}{n.h^p} \,, \\
\max_{h\in\mathcal{H}_n} r(W_h - P_h) &= \max_{h\in\mathcal{H}_n} \max_{i=1,\ldots,n} |p_{ii}(h)| = O(1/(n.h_{D_n}^p)) = o(1) \,, \\
\max_{h\in\mathcal{H}_n} h^p \nu^2(W_h - P_h) &= \max_{h\in\mathcal{H}_n} h^p \sum_{i=1}^n p_{ii}^2(h) = O(1/(n.h_{D_n}^p)) = o(1) \,,
\end{aligned}$$

since $n.h_{D_n}^p \to \infty$. The last equation of (5.20) is proven observing that the above bounds are uniform with respect to $h \in \mathcal{H}_n$ and using

$$\begin{aligned}
&\max_{h\in\mathcal{H}_n\setminus\{h_0\}} \left(h^{-p} - h_0^{-p}\right)^{-1} \nu^2 \left(W_h - W_{h_0} - (P_h - P_{h_0})\right) \\
&\leq 2\left(h_1^{-p} - h_0^{-p}\right)^{-1} \left(\nu^2\left(W_{h_0} - P_{h_0}\right) + \max_{h\in\mathcal{H}_n} \nu^2\left(W_h - P_h\right)\right) \\
&= h_0^p \left(\left(\frac{h_0}{h_1}\right)^p - 1\right)^{-1} \left(\frac{1}{nh_0^{2p}} + \frac{1}{nh_{D_n}^{2p}}\right) = o(1) \,,
\end{aligned}$$

since $h_1 = o(h_0)$ and $nh_{D_n}^p$ diverges.

We now deal with Assumption B. Note that the first equation in (5.20) implies B–i. Assumptions B-ii, iii holds for Examples 1a and 1b as explained in the discussion following Assumption B. For Example 1c, it suffices to establish B–ii and iii for the regressogram, since the regressogram corresponds to polynomials of order 0. This can be established as in Guerre and Lavergne (1999) or Guerre and Lieberman (2000). □

Proof of Proposition 3. Let us study the leave-one-out density estimator. We have

$$\left|\frac{1}{h^p} K_h(x - X_j)\right| \leq \frac{C}{h^p} \quad \text{and Var}\left[\frac{1}{h^p} K_h(x - X_j)\right] \leq \frac{1}{h^{2p}} \int K_h^2(x - x') f(x') dx' \leq \frac{C}{h^p} \,. \tag{5.22}$$

The Bernstein inequality then yields, for any $\epsilon > 0$ large enough,

$$\begin{aligned}
&I\!P\left(\max_{h\in\mathcal{H}_n} \max_{1\leq i\leq n} \left(\frac{nh^p}{\ln n}\right)^{1/2} \left|\widehat{f}_h(X_i) - f_h(X_i)\right| \geq \epsilon\right) \\
&\leq \sum_{h\in\mathcal{H}_n} \sum_{i=1}^n I\!P\left(\left|\widehat{f}_h(X_i) - f_h(X_i)\right| \geq \left(\frac{\ln n}{nh^p}\right)^{1/2} \epsilon\right) \\
&\leq 2 \sum_{h\in\mathcal{H}_n} \sum_{i=1}^n \exp\left(-\frac{\epsilon^2 \ln n}{C}\right) \leq 2\exp\left(-\frac{\epsilon^2 \ln n}{C} + \ln(nD_n)\right) \to 0 \,.
\end{aligned}$$

Hence, under Assumption X and K, as $\ln n / nh^{D_n} = o(1)$,

$$\max_{h\in\mathcal{H}_n} \max_i \left|\frac{1}{f_h(X_i)} \frac{\sum_{j\neq i} K_h(X_i - X_j)}{nh^p} - 1\right| = o_{I\!P}(1) \,. \tag{5.23}$$

We now prove W-*ii*. We shall consider the matrices $(W_h + W_h')/2$, where W_h is defined in Section 1, Example 2. Note that $r((W_h + W_h')/2) \leq r(W_h)$. For $u \in I\!R^n$ with $\|u\| = 1$, we have

$$\begin{aligned}
\|W_h u\|^2 &= \sum_{i=1}^n \left(\sum_{j=1}^n w_{ij}(h) u_j \right)^2 \leq \sum_{i=1}^n \left(\sum_{j=1}^n |w_{ij}(h)| \right)^2 \left(\sum_{j=1}^n |u_j| \frac{|w_{ij}(h)|}{|\sum_{\ell=1}^n |w_{i\ell}(h)||} \right)^2 \\
&\leq \sum_{i=1}^n \left(\sum_{j=1}^n |w_{ij}(h)| \right) \sum_{j=1}^n u_j^2 |w_{ij}(h)| \leq \max_i \sum_{j=1}^n |w_{ij}(h)| \left(\sum_{j=1}^n u_j^2 \sum_{i=1}^n |w_{ij}(h)| \right) \\
&\leq \max_i \sum_{j=1}^n |w_{ij}(h)| \max_j \sum_{i=1}^n |w_{ij}(h)| \qquad (5.24) \\
&\leq \max_i \left(\frac{\sum_{j=1, j\neq i}^n |K_h(X_i - X_j)|}{nh^p |\widehat{f}_h(X_i)|} \right) \max_j \left(\frac{\sum_{i=1, i\neq j}^n |K_h(X_i - X_j)|}{nh^p |\widehat{f}_h(X_i)|} \right) .
\end{aligned}$$

Then by (5.23) and Assumption X, $\max_{h \in \mathcal{H}_n} r(W_h) = O_{I\!P}(1)$.

Let us now check W-*iii*. We have

$$\begin{aligned}
\nu^2 \left(\frac{W_h + W_h'}{2} \right) &= \sum_{i,j} \left(\frac{w_{ij}(h) + w_{ji}(h)}{2} \right)^2 \leq (1/2) \sum_{i,j} \left(w_{ij}^2(h) + w_{ji}^2(h) \right) = \nu^2(W_h) , \\
\nu^2 \left(\frac{W_h + W_h'}{2} \right) &= (1/2) \left[\nu^2(W_h) + \sum_{i,j} w_{ij}(h) w_{ji}(h) \right] = (1/2) \left[\nu^2(W_h) + \sum_{i \neq j} \frac{K_h^2 (X_i - X_j)}{\widehat{f}_h(X_i) \widehat{f}_h(X_j)} \right] ,
\end{aligned}$$

and the last term is positive uniformly in $h \in \mathcal{H}_n$ with probability converging to one by (5.23) and Assumption X, so that $\nu^2 \left((W_h + W_h')/2 \right) \asymp_{I\!P} \nu^2(W_h)$. ¿From the triangular inequality and for $h \leq h_1 < h_0$,

$$\begin{aligned}
\left(h^{-p} - h_0^{-p} \right)^{-1/2} \nu(W_h - W_{h_0}) &\leq \left(h^{-p} - h_0^{-p} \right)^{-1/2} \left(\nu(W_h) + \nu(W_{h_0}) \right) \\
&\leq \left(1 - \left(\frac{h_1}{h_0} \right)^p \right)^{-1/2} \left(h^{p/2} \nu(W_h) + h_0^{p/2} \nu(W_{h_0}) \right) , \\
\left(h^{-p} - h_0^{-p} \right)^{-1/2} \nu(W_h - W_{h_0}) &\geq \left(h^{-p} - h_0^{-p} \right)^{-1/2} \left(\nu(W_h) - \nu(W_{h_0}) \right) \\
&\geq h^{p/2} \nu(W_h) - \left(\left(\frac{h_0}{h_1} \right)^p - 1 \right)^{-1/2} h_0^{p/2} \nu(W_{h_0}) .
\end{aligned}$$

Hence, as $h_1 = o(h_0)$, W-*iii* holds if $h^{p/2} \nu(W_h) \asymp_{I\!P} 1$ uniformly in $h \in \mathcal{H}_n$. From (5.23),

$$h^p \nu^2(W_h) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\widehat{f}_h^2(X_i)} \frac{1}{nh^p} \sum_{j=1, j\neq i}^n K_h^2(X_i - X_j) \asymp_{I\!P} 1 ,$$

under Assumption X, uniformly in $h \in \mathcal{H}_n$.

We now check that Assumption B–i holds. Let P_h, $h \in \mathcal{H}_n$ be the symmetric matrices with entries

$$p_{ij}(h) = \frac{K_h\left(X_i - X_j\right)}{nh^p\sqrt{|\widehat{f}_h(X_i)||\widehat{f}_h(X_j)|}} \ , \ 1 \leq i,j \leq n \ .$$

For any $a \in I\!R^n$,

$$\begin{aligned} a'P_h a &= \frac{1}{nh^p}\sum_{1\leq p,q\leq n}\frac{a_p}{\sqrt{|\widehat{f}_h(X_p)|}}\frac{a_q}{\sqrt{|\widehat{f}_h(X_q)|}}\int \exp(it'(X_p - X_q))\widehat{K}(t)dt \\ &= \frac{1}{nh^p}\int\left|\sum_{p=1}^n \frac{a_p}{\sqrt{|\widehat{f}_h(X_p)|}}\exp(it'X_p)\right|^2 \widehat{K}(t)dt \end{aligned}$$

is nonnegative since $\widehat{K}(\cdot)$ is nonnegative and continuous. Therefore the P_h's are positive semidefinite. As $r((W_h + W_h')/2 - P_h) \leq r(W_h - P_h)$, the bound (5.24) yields

$$\begin{aligned} r^2\left(\frac{1}{2}(W_h + W_h') - P_h\right) &\leq \max_i\left(\sum_{j=1}^n |w_{ij}(h) - p_{ij}(h)|\right)\max_j\left(\sum_{i=1}^n |w_{ij}(h) - p_{ij}(h)|\right) \\ &\leq \max_i\left(\frac{|K(0)|}{nh^p|\widehat{f}_h(X_i)|} + \left|\sum_{j\neq i}\frac{K_h(X_i - X_j)}{nh^p}\left[\frac{1}{\widehat{f}_h(X_i)} - \frac{1}{\sqrt{|\widehat{f}_h(X_i)||\widehat{f}_h(X_j)|}}\right]\right|\right)^2 \\ &\leq \max_i\left(\frac{|K(0)|}{nh^p f_h(X_i)} + C\sum_{j\neq i}\frac{|K_h(X_i - X_j)|}{nh^p}|f_h(X_i) - f_h(X_j)|\right)^2 + o_{I\!P}(1) \ , \end{aligned}$$

where the remaining term is uniform with respect to $h \in \mathcal{H}_n$, using (5.23) and (3.13). Observe now that the expectation of $|K_h(x - X_j)||f_h(x) - f_h(X_j)|$ goes to 0 uniformly in x and h, since $h_0 = o(1)$ and $f(\cdot)$ is continuous on $[0,1]^p$ by Assumption X. Then applying the Bernstein inequality as in (5.23) shows that $\max_{h\in\mathcal{H}_n} r((W_h + W_h')/2 - P_h) = o_{I\!P}(1)$.

We now deal with Assumption B–*ii* and *iii*. For $k = (k_1, \ldots, k_p) \in I\!N^p$ and $h \in \mathcal{H}_n$, let $x_k'(h) = ((k_1 + 1/2), \ldots, (k_p + 1/2))\, h$, $I_k(h) = x_k(h) + h[-1/2, 1/2]^p$, with $x_k(h) \in (0,1)^p$, which ensures $[0,1]^p \subset \cup_k I_k(h)$, and $(x)_1^q = x^q\, I\!I(x \in [-1/2, 1/2]^p)$. Let $\Pi_{s,h} = \Pi_h$ be the set of piecewise polynomial functions over the $I_k(h)$'s, of degree smaller than or equal to $\overline{s}$. Then a standard Taylor expansion shows that B–*ii* holds. Let $\pi_{k,h}(\cdot)$ be a polynomial function over $I_k(h)$, i.e.

$$\pi_{k,h}(x) = \sum_{|q|\leq\overline{s}} a_q\left(\frac{x - x_k(h)}{h}\right)_1^q \ .$$

Consider the $[\overline{s}]^p \times [\overline{s}]^p$ matrix $\bar{P}_{k,h}$ defined by $\pi'_{k,h} P_h \pi_{k,h} = nh^p a' \overline{P}_{k,h} a$. Then, as in Ingster (1993), Guerre and Lavergne (1999), Guerre and Lieberman (2000), Assumption B–*iii* holds if the limits of the $\overline{P}_{k,h}$'s when h goes to 0 and n grows are positive definite. The entries of these matrices write

$$\overline{p}_{q_1 q_2}(h,k) = \frac{1}{n^2 h^{2p}} \sum_{1 \le i,j \le n} \left(\frac{X_j - x_k(h)}{h} \right)_1^{q_1} \frac{K_h(X_i - X_j)}{\sqrt{|\widehat{f}_h(X_i)\widehat{f}_h(X_j)|}} \left(\frac{X_j - x_k(h)}{h} \right)_1^{q_2} .$$

Using (5.22) and applying the Bernstein inequality for U-statistics in Arcones and Giné (1993, Prop. 2.3) yields a bound similar to (5.23). This shows that the limits of these coefficients are, uniformly with respect to q_1, q_2, k and h, given by

$$\begin{aligned}
&\frac{1}{h^{2p}} I\!E \left[\left(\frac{X_j - x_k(h)}{h} \right)_1^{q_1} \frac{K_h(X_i - X_j)}{\sqrt{f_h(X_i) f_h(X_j)}} \left(\frac{X_j - x_k(h)}{h} \right)_1^{q_2} \right] \\
&= \frac{1}{h^{2p}} \int\int \left(\frac{x_1 - x_k(h)}{h} \right)_1^{q_1} \frac{K_h(x_1 - x_2)}{\sqrt{f_h(x_1) f_h(x_2)}} \left(\frac{x_2 - x_k(h)}{h} \right)_1^{q_2} f(x_1) f(x_2)\, dx_1 dx_2 \\
&= \int\int (t_1)_1^{q_1} K(t_1 - t_2) (t_2)_1^{q_2} \frac{f(x_k(h) + ht_1) f(x_k(h) + ht_2)}{\sqrt{f(x_k(h) + ht_1) f(x_k(h) + ht_2)}}\, dt_1 dt_2 \\
&\sim f(x_k(h)) \int\int (t_1)_1^{q_1} K(t_1 - t_2) (t_2)_1^{q_2}\, dt_1 dt_2 .
\end{aligned}$$

Let $\overline{P}$ be the $[\overline{s}]^p \times [\overline{s}]^p$ with generic elements $\int\int (t_1)_1^{q_1} K(t_1 - t_2)(t_2)_1^{q_2}\, dt_1 dt_2$. We conclude the proof by showing that $\overline{P}$ is definite positive. Indeed, for $\pi_1(t) = \sum_{|q| \le \overline{s}} a_q (t)_1^q$ and $\widehat{\pi}_1(x) = \int \exp(itx')\pi_1(t) dt$,

$$a' \overline{P} a = \int\int \pi_1(t_1) \left(\int \exp(ix'(t_1 - t_2)) \widehat{K}(x) dx \right) \pi_1(t_2)\, dt_1 d_2 = \int |\widehat{\pi}_1(x)|^2 \widehat{K}(x)\, dx ,$$

which, under Assumption K, is nonnegative and vanishes iff $\widehat{\pi}_1(\cdot) = 0$, that is $a = 0$. □

5.4 Auxiliary lemmas

Lemma 1 *Under Assumption E,* $I\!E_m[T_h] = \delta' W_h \delta$,

$$\begin{aligned}
\mathrm{Var}_m[T_{h_0}] &= 2\nu^2 [D(\sigma) W_{h_0} D(\sigma)] + 4\delta' W_{h_0} D(\sigma^2) W_{h_0} \delta \\
&\le 2\overline{\sigma}^2 \nu^2 (W_{h_0}) + 4\overline{\sigma}^2 r^2 (W_{h_0}) \|\delta\|^2 , \\
\mathrm{Var}_m[T_h - T_{h_0}] &= 2\nu^2 [D(\sigma) (W_h - W_{h_0}) D(\sigma)] + 4\delta' (W_h - W_{h_0}) D(\sigma^2) (W_h - W_{h_0}) \delta \\
&\le 2\overline{\sigma}^2 \nu^2 (W_h - W_{h_0}) + 4\overline{\sigma}^2 r^2 (W_h - W_{h_0}) \|\delta\|^2 .
\end{aligned}$$

Proof of Lemma 1. As $T_h = \delta' W_h \delta + 2\delta' W_h \varepsilon + \varepsilon' W_h \varepsilon$, where W_h has zeros on its diagonal, the mean formula is immediate. For the variance, as W_h is symmetric with zeros on its diagonal,

$$\begin{aligned} I\!E_m \left[\varepsilon' W_h \varepsilon \varepsilon' W_h \varepsilon\right] &= \sum_{i_1,j_1} \sum_{i_2,j_2} w_{i_1 j_1}(h) w_{i_2,j_2}(h) I\!E\left[\varepsilon_{i_1}\varepsilon_{j_1}\varepsilon_{i_2}\varepsilon_{j_2}\right] \\ &= 2\sum_{i,j} w_{ij}^2(h)\sigma^2(X_i)\sigma^2(X_j) = 2\nu^2\left[D(\sigma)W_h D(\sigma)\right] . \end{aligned}$$

$$\begin{aligned} \text{Hence,} \qquad \mathrm{Var}_m[T_h] &= I\!E_m\left[\left(2\delta' W_h \varepsilon + \varepsilon' W_h \varepsilon\right)^2\right] = I\!E_m\left[\varepsilon' W_h \varepsilon \varepsilon' W_h \varepsilon\right] + 4 I\!E_m\left[\delta' W_h \varepsilon \varepsilon' W_h \delta\right] \\ &= 2\nu^2 [D(\sigma) W_h D(\sigma)] + 4\delta' W_h D(\sigma^2) W_h \delta , \end{aligned}$$

This gives the result for T_{h_0} directly, and for $T_h - T_{h_0}$ by replacing W_h by $W_h - W_{h_0}$. □

Under H_0, consider

$$T_h = U' W_h U = \varepsilon' W_h \varepsilon \ , \ v_h^2 = \mathrm{Var}_{H_0}(T_h) = 2\nu^2 [D(\sigma) W_h D(\sigma)] = \sum_{i,j} w_{ij}^2 \sigma^2(X_i)\sigma^2(X_j) \ ,$$

see Lemma 1 for this expression.

Lemma 2 *Under Assumption E,*

i. Under H_0, there exists $C_\eta > 0$ depending only upon η, $\underline{\sigma}^2$, $\overline{\sigma}^2$ and ν_3, such that

$$I\!P\left(\left|\frac{T_h}{v_h}\right| \geq \gamma_n\right) \leq \frac{\sqrt{2}}{\sqrt{\pi}(\gamma_n - \eta)} \exp\left(-\frac{(\gamma_n - \eta)^2}{2}\right) + C_\eta \frac{r(W_h)}{\nu(W_h)} , \quad \forall \gamma_n > \eta > 0$$

ii. Under H_0, if $r(W_h) < \infty$ and $\nu(W_h) \to \infty$, then T_h / v_h converges in distribution to a $N(0,1)$.

Proof of Lemma 2. The proof works by establishing that for some constant $C > 0$,

$$\left|I\!E\psi(T_h/v_h) - I\!E\psi(N(0,1)\right| \leq C \sup_{t \in I\!R} \left|\psi^{(3)}(t)\right| \frac{r(W_h)}{\nu(W_h)} , \tag{5.25}$$

for any three times continuously derivable $\psi(\cdot)$ with bounded third derivative. Part *i* follows by taking $\psi(\cdot) = \exp(ix\cdot)$. For the second part, observe that there exists a three times differentiable function $\psi_\eta(\cdot)$ with $1\!\!1(|\cdot| \geq \gamma_n) \leq \psi_\eta(|\cdot| - \gamma_n) \leq 1\!\!1(|\cdot| \geq \gamma_n - \eta)$ for any $\gamma_n > \eta$. Applying Mill's ratio inequality, (5.25) implies the desired result.

The bound (5.25) is obtained in two steps, which relies on some standard arguments to establish the central limit theorem, see e.g. the proof of Theorem 7.2 in Billingsley (1968). Horowitz and Spokoiny (2001) also use arguments similar to our Step 1.

Step 1. Let ε^g be a $(1 \times n)$ vector with typical element distributed as $N(0, \sigma^2(X_i))$, independent of ε and of other elements of ε^g. Define

$$T_h^g = \varepsilon^{g'} W_h \varepsilon^g \ , \ T_h^i(\tau) = T_h(\varepsilon_1, \ldots, \varepsilon_{i-1}, \tau, \varepsilon_{i+1}^g, \ldots, \varepsilon_n^g) \ .$$

We have

$$\begin{aligned} \left| I\!E\left[\psi\left(\frac{T_h}{v_h}\right) - \psi\left(\frac{T_h^g}{v_h}\right)\right]\right| &= \left| I\!E\left[\psi\left(\frac{T_h^n(\varepsilon_n)}{v_h}\right) - \psi\left(\frac{T_h^1(\varepsilon_1^g)}{v_h}\right)\right]\right| \\ = \left|\sum_{i=1}^n I\!E\left[\psi\left(\frac{T_h^i(\varepsilon_i)}{v_h}\right) - \psi\left(\frac{T_h^i(\varepsilon_i^g)}{v_h}\right)\right]\right| &\leq \sum_{i=1}^n \left| I\!E\left[\psi\left(\frac{T_h^i(\varepsilon_i)}{v_h}\right) - \psi\left(\frac{T_h^i(\varepsilon_i^g)}{v_h}\right)\right]\right| \ . \end{aligned} \tag{5.26}$$

Since $w_{ii} = 0$,

$$\frac{\partial T_h^i(\tau)}{\partial \tau} = \sum_{j<i} w_{ij}(h)\varepsilon_j + \sum_{j>i} w_{ij}(h)\varepsilon_j^g \ ,$$

and the derivatives of higher order are all 0. This gives, for $k = 1, 2, 3$,

$$\frac{\partial^k \psi(T_h^j(\tau)/v_h)}{\partial \tau^k} = \left(\frac{1}{v_h}\frac{\partial T_h^i(\tau)}{\partial \tau}\right)^k \psi^{(k)}(T_h^j(\tau)/v_h) \ .$$

Observe that $(\varepsilon_i, \varepsilon_i^g)$ is independent of $\partial T_h^i(\tau)/\partial\tau$ and of $T_h^i(0)$. Therefore, a third order Taylor expansion of $\psi(T_h^i(\tau)/v_h)$ around $\tau = 0$ yields

$$\begin{aligned} &\left| I\!E\left[\psi\left(\frac{T_h^i(\varepsilon_i)}{v_h}\right) - \psi\left(\frac{T_h^i(\varepsilon_i^g)}{v_h}\right)\right]\right| \\ &\leq \left| I\!E\left[(\varepsilon_i - \varepsilon_i^g)\frac{\partial \psi(T_h^j(0)/v_h)}{\partial \tau} + \frac{1}{2}\left((\varepsilon_i)^2 - (\varepsilon_i^g)^2\right)\frac{\partial^2 \psi(T_h^j(0)/v_h)}{\partial^2 \tau}\right]\right| \\ &\quad + \frac{\sup_{t \in I\!R}|\psi^{(3)}(t)|}{6v_h^3}\left[\left(I\!E|\varepsilon_i|^3 + I\!E|\varepsilon_i^g|^3\right) I\!E\left|\frac{\partial T_h^i(\tau)}{\partial \tau}\right|^3\right] \\ &\leq \frac{\sup_{t \in I\!R}|\psi^{(3)}(t)|}{6v_h^3}\left[\left(\nu_3 + I\!E|N(0,1)|^3\right) I\!E\left|\sum_{j<i} w_{ij}(h)\varepsilon_j + \sum_{j>i} w_{ij}(h)\varepsilon_j^g\right|^3\right] \ . \end{aligned} \tag{5.27}$$

Using the Marcinkiewicz-Zygmund and Minkowski inequalities, see e.g. Chow and Teicher (1988),

$$\begin{aligned} \left(I\!E^{1/3}\left|\sum_{j<i} w_{ij}(h)\varepsilon_j + \sum_{j>i} w_{ij}(h)\varepsilon_j^g\right|^3\right)^3 &\leq C\left(I\!E^{2/3}\left|\sum_{j<i} w_{ij}^2(h)(\varepsilon_j)^2 + \sum_{j>i} w_{ij}^2(h)(\varepsilon_j^g)^2\right|^{3/2}\right)^{3/2} \\ &\leq C\left(\sum_{j<i} w_{ij}^2(h) I\!E^{2/3}|\varepsilon_j|^3 + \sum_{j>i} w_{ij}^2(h) I\!E^{2/3}|\varepsilon_j^g|^3\right)^{3/2} \end{aligned}$$

$$\leq \quad C \left(\sum_{j=1}^{n} w_{ij}^2(h) \right)^{3/2} \quad , \tag{5.28}$$

where C varies from line to line. Let $\{u_i, i = 1, \ldots, n\}$ be the canonical basis of $I\!R^n$ with $\|u_i\| = 1$. Then

$$\left(\sum_{j=1}^{n} w_{ij}^2(h) \right)^{3/2} \leq \sup_{j=1,\ldots,n} \|W_h u_j\| \times \sum_{j=1}^{n} w_{ij}^2(h) \leq r(W_h) \times \sum_{j=1}^{n} w_{ij}^2(h) \; .$$

Combining the last inequality with (5.26)–(5.28) yields

$$\left| I\!E \left[\psi \left(\frac{T_h}{v_h} \right) - \psi \left(\frac{T_h^g}{v_h} \right) \right] \right| \leq C \sup_{t \in I\!R} \left| \psi^{(3)}(t) \right| \frac{r(W_h)}{v_h^3} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^2(h) \leq C \sup_{t \in I\!R} \left| \psi^{(3)}(t) \right| \frac{r(W_h)}{\nu(W_h)} \; , \tag{5.29}$$

as $v_h^2 \asymp \nu^2(W_h)$, by Assumptions E, W–*iii* and Lemma 1.

Step 2. We now show that the distribution of T_h^g / v_h is close to the standard normal. Let $\{b_i, i = 1, \ldots, n\}$ be an orthonormal system of eigenvectors of the symmetric W_h associated with eigenvalues $\lambda_1, \ldots, \lambda_n$. Because T_h^g is centered, we have

$$T_h^g = \sum_{i=1}^{n} \lambda_i \left((b_i' \varepsilon_g)^2 - I\!E (b_i' \varepsilon_g)^2 \right) \; ,$$

the $b_i' \varepsilon_g$ being independent $N(0, b_i' D(\sigma^2) b_i)$ variables. Therefore, the distribution of T_h^g is the distribution of $\sum_{i=1}^{n} \lambda_i b_i' D(\sigma^2) b_i \zeta_i$, where the ζ_i's are centered chi-squared with one degree of freedom. Reasoning as in Step 1, we can approximate T_h^g by $\sum_{i=1}^{n} \lambda_i b_i' D(\sigma^2) b_i \tilde{\zeta}_i$, where the $\tilde{\zeta}_i$'s are i.i.d. $N(0, 2)$. This yields

$$\left| I\!E \left[\psi \left(\frac{T_h^g}{v_h} \right) - \psi(N(0,1) \right] \right| \leq \sup_{t \in I\!R} \left| \psi^{(3)}(t) \right| \frac{C}{v_h^3} \sum_{i=1}^{n} |\lambda_i|^3 \left(b_i' D(\sigma^2) b_i \right)^3 \; ,$$

with $|\lambda_i| \leq r(W_h)$, $|b_i' D(\sigma^2) b_i| \leq \overline{\sigma}^2$, $v_h^2 = 2 \sum_{i=1}^{n} \lambda_i^2 \left(b_i' D(\sigma^2) b_i \right)^2 \asymp \nu^2(W_h)$. Therefore

$$\left| I\!E \left[\psi \left(\frac{T_h^g}{v_h} \right) - \psi(N(0,1) \right] \right| \leq C \sup_{t \in I\!R} \left| \psi^{(3)}(t) \right| \frac{r(W_h)}{\nu(W_h)} \; .$$

Combining this last inequality with (5.29) implies (5.25) . □

Lemma 3 *Define $r_n = \max_{h \in \mathcal{H}_n} r(P_h - W_h)$ and $R_n = \max_{h \in \mathcal{H}_n} r(P_h)$. Under Assumptions E and B, if $r_n \geq \Lambda_n$, then for any $\delta(\cdot) \in C_p(L, s)$,*

$$\delta' W_h \delta \geq \left[(\Lambda_n - r_n) \|\delta\| - C_0 L (\Lambda_n + R_n) \sqrt{n} h^s \right]^2 \; \textit{whenever} \; \|\delta\| \geq C_0 L \frac{\Lambda_n + R_n}{\Lambda_n - r_n} \sqrt{n} h^s \; .$$

Proof of Lemma 3. Assumption B–i gives

$$\delta' W_h \delta = \delta' P_h \delta - \delta'(P_h - W_h)\delta \geq \delta' P_h \delta - r_n^2 \|\delta\|^2 \ .$$

Let $\eta(\cdot) = \delta(\cdot) - \pi_\delta(\cdot)$, and note that $\|\eta\| \leq C_0 L \sqrt{n} h^s$ by Assumption B-ii. As P_h is positive semidefinite,

$$(\delta' P_h \delta)^{1/2} \geq (\pi' P_h \pi)^{1/2} - (\eta' P_h \eta)^{1/2} \geq \Lambda_n \|\delta - \eta\| - R_n \|\eta\| \geq \Lambda_n \|\delta\| - (\Lambda_n + R_n)\|\eta\| \ .$$

Since $a^2 - b^2 = (a-b)(a+b) \geq (a-b)^2$, for $a > b > 0$, then if $\|\delta\| \geq C_0 L \frac{\Lambda_n + R_n}{\Lambda_n - r_n} \sqrt{n} h^s$,

$$\begin{aligned} \delta' W_h \delta &\geq [\Lambda_n \|\delta\| - (\Lambda_n + R_n)\|\eta\|]^2 - r_n^2 \|\delta\|^2 \geq [(\Lambda_n - r_n)\|\delta\| - (\Lambda_n + R_n)\|\eta\|]^2 \\ &\geq \left[(\Lambda_n - r_n)\|\delta\| - C_0 L (\Lambda_n + R_n)\sqrt{n} h^s\right]^2 \ . \square \end{aligned}$$

Lemma 4 *Let the matrices Φ_h be as defined from the functions in (5.21). Then, under Assumption X, there exists a constant $C > 1$ such that*

$$\max \left[r\left(\left(n^{-1} \Phi'_{h_{D_n}} \Phi_{h_{D_n}} \right)^{-1} \right), r\left(n^{-1} \Phi'_{h_{D_n}} \Phi_{h_{D_n}} \right) \right] \leq C \ ,$$

with probability tending to 1, provided $h_{D_n}^{-p} = o((n/\ln n)^{1/3})$ in Example 1a, $h_{D_n}^{-p} = o((n/\ln n)^{1/2})$ in Example 1b and $h_{D_n}^{-p} = o(n/\ln)$ in Example 1c, with $D_n = O(n)$.

Proof of Lemma 4. The case of Examples 1b and 1c is studied in Baraud (1997), and we restrict to Example 1a. Observe that $n^{-1} \Phi'_{h_{D_n}} \Phi_{h_{D_n}}$ is a Gram matrix with entries $\sum_{i=1}^n \phi_k(X_i)\phi_{k'}(X_i)/n$, $0 \leq |k|, |k'| < 1/h_{D_n}$. Moreover

$$\begin{aligned} |\phi_k(X_i)\phi_{k'}(X_i)| &\leq \prod_{\ell=1}^p \sqrt{2k_\ell + 1}\sqrt{2k'_\ell + 1} \leq C h_{D_n}^{-p} \ , \\ \mathrm{Var}(\phi_k(X_i)\phi_{k'}(X_i)) &\leq I\!E \phi_k^2(X_i)\phi_{k'}^2(X_i) \leq I\!E^{1/2}\phi_k^4(X_i) I\!E^{1/2}\phi_{k'}^4(X_i) \\ &\leq \sup_{x \in [0,1]^p} |\phi_k(x)| \sup_{x \in [0,1]^p} |\phi_{k'}(x)| I\!E^{1/2}\phi_k^2(X_i) I\!E^{1/2}\phi_{k'}^2(X_i) \leq C h_{D_n}^{-p} \ , \end{aligned}$$

under Assumption X and $\phi_k^2(\cdot)$ has integral one for any k. Therefore the Bernstein inequality yields, for any $t > 0$,

$$I\!P\left(\sqrt{\frac{n h_{D_n}^p}{\ln n}} \sup_{0 \leq |k|, |k'| < 1/h_{D_n}} \left| \frac{1}{n} \sum_{i=1}^n \phi_k(X_i)\phi_{k'}(X_i) - I\!E \phi_k(X)\phi_{k'}(X) \right| \geq t \right) \leq O(h_{D_n}^{-2p}) \exp\left(-\frac{t^2 \ln n}{C} \right) \ .$$

This shows that

$$n^{-1}\Phi'_{h_{D_n}}\Phi_{h_{D_n}} = n^{-1}I\!E\Phi'_{h_{D_n}}\Phi_{h_{D_n}} + R_{h_{D_n}} ,$$

where $R_{h_{D_n}}$ is a $h_{D_n}^{-p} \times h_{D_n}^{-p}$ matrix whose elements are uniformly $O_{I\!P}\left(\sqrt{\ln n/nh_{D_n}^p}\right)$. Thus the spectral radius of $R_{h_{D_n}}$ is such that

$$r\left(R_{h_{D_n}}\right) = O_{I\!P}\left(\frac{1}{h_{D_n}^p}\sqrt{\frac{\ln n}{nh_{D_n}^p}}\right) = o_{I\!P}(1) ,$$

as $h_{D_n}^{-p} = o((n/\ln n)^{1/3})$. Hence the eigenvalues of $n^{-1}\Phi'_{h_{D_n}}\Phi_{h_{D_n}}$ are between the smallest and largest eigenvalues of $n^{-1}I\!E\Phi'_{h_{D_n}}\Phi_{h_{D_n}}$ with probability tending to one. But, under Assumption X, for any $a \in I\!R^{h_{D_n}^{-p}}$,

$$n^{-1}a'I\!E\Phi'_{h_{D_n}}\Phi_{h_{D_n}}a = I\!E\left(\sum_{0\leq|k|<[1/h_{D_n}]} a_k\phi_k(X)\right)^2 \asymp \int_{[0,1]^p}\left(\sum_{0\leq|k|<[1/h_{D_n}]} a_k\phi_k(x)\right)^2 dx = a'a ,$$

since the $\phi_k(\cdot)$'s are orthonormal in $L_2([0,1]^p, dx)$. Therefore the eigenvalues of the symmetric matrix $n^{-1}I\!E\Phi'_{h_{D_n}}\Phi_{h_{D_n}}$ are bounded away from 0 and infinity when n grows. □

REFERENCES

AERTS, M., G. CLAESKENS and J.D. HART (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association* 94 869–879.

AERTS, M., G. CLAESKENS and J.D. HART (2000). Testing lack of fit in multiple regression. *Biometrika* 87 (2) 405–4242.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki eds., Akademiai Kiodo, Budapest, pp. 267–281.

ANDREWS (1997). A conditional Kolmogorov test. *Econometrica* 65 (5) 1097–1128.

ARCONES, M.A. and E. GINE (1993). Limit theorems for U-processes. *Annals of Probability* 21 1494–1542.

BARAUD, Y. (1997). Model selection for regression on random design. Ecole Normale Supérieure, Paris. Available at `www.dma.ens.fr/~baraud/`.

BARAUD, Y., S. HUET and B. LAURENT (1999). Adaptive tests of linear hypotheses by model selection. Ecole Normale Supérieure, Paris. Available at `www.dma.ens.fr/~baraud/`.

BARRY, D. and J.A. HARTIGAN (1990). An omnibus test for departures from constant mean. *Annals of Statistics* 18 1340–1357.

BIERENS, H.J. (1982). Consistent model specification tests. *Journal of Econometrics* 20 105–134.

BIERENS, H.J. and W. PLOBERGER (1997). Asymptotic theory of integrated conditional moment tests. *Econometrica* 65 (5) 1129–1151.

BIRGE, L. and P. MASSART (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields* 97 113-150.

BUCKLEY, M.J. (1991). Detecting a smooth signal: optimality of cusum based procedures. *Biometrika* 78 253–262.

CHEN, J.C. (1994). Testing goodness-of-fit of polynomial models via spline smoothing techniques. *Statistics and Probability Letters* 19 65–76.

CHOW, Y.S. and H. TEICHER (1988). *Probability Theory: Independence, Interchangeability, Martingales*. Springer-Verlag, New-York.

CLEVELAND, W.S. and S.J. DEVLIN (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83 (403) 596–610.

DAVIS, P.J (1975). *Interpolation and approximation*. Dover.

DELGADO, M.A. (1993). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters* 17 199–204.

DETTE, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics* 27(3) 1012–1040.

ELLISON, G. and S. FISHER ELLISON (2000). A simple framework for nonparametric specification testing. *Journal of Econometrics* 96 1–23.

EUBANK, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New-York.

EUBANK, R.L. and J.D. HART (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* 20 (3) 1412–1425.

EUBANK, R.L. and J.D. HART (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* 80 89–98.

EUBANK, R.L. and C.H. SPIEGELMAN (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association* 85 (410) 387–392.

FAN, J. and I. GIJBELS (1996). *Local Polynomial Modelling and its Applications*. Chapman et Hall, London.

FAN, J. and L.S. HUANG (2000). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association* forthcoming.

GOURIEROUX, C., A. MONFORT and A. TROGNON (1984). Pseudo-maximum likelihood methods: theory. *Econometrica* 52 681–700.

GOZALO, P.L. (1997). Nonparametric bootstrap anamysis with applicatins to demographic effects in demand functions. *Journal of Econometrics* 81 357–393.

GUERRE, E. and P. LAVERGNE (1999). Optimal minimax rates for nonparametric specification testing in regression models. LSTA, Univ. Paris 6. Available at `www.toulouse.inra.fr/centre/esr/CV/lavergne/lavergnefr.html`.

GUERRE, E. and O. LIEBERMAN (2000). α-level adaptive testing in nonparametric regression via selection criteria. LSTA, Univ. Paris 6.

GYÖRFI, L., W. HÄRDLE, P. SARDA and P. VIEU (1989). *Nonparametric Curve Estimation From Time Series*. Springer-Verlag, Berlin.

HALL, P. and J.D. HART (1990). Bootstrap test for difference bewteen means in nonparametric regression. *Journal of the American Statistical Association* 85 1039–1049.

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

HÄRDLE, W., G. KERKYACHARIAN, D. PICARD and A. TSYBAKOV (1998). *Wavelets, Approximation and Statistical Applications*. Lecture notes in Statistics, 129. Springer-Verlag, Berlin.

HÄRDLE, W. and E. MAMMEN (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21 (4) 1926–1947.

HART, J.D. and T.E. WEHRLY (1992). Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. *Journal of the American Statistical Association* 87 1018–1024.

HART, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Verlag, New-York.

HONG, Y. and H. WHITE (1995). Consistent specification testing via non-parametric series regressions. *Econometrica* 63 1133–1160.

HOROWITZ, J.L. and V.G. SPOKOINY (2001). An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica* forthcoming.

INGLOT, T., W.C.M. KALLENBERG and T. LEDWINA (1997). Data-driven smooth tests for composite hypotheses. *Annals of Statistics* 25 (3) 1222–1250.

INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. (Part I, II and III) *Mathematical Methods of Statistics* 2 85-114 171–189 and 249–268.

LAVERGNE, P. and Q.H. VUONG (1998). An integral estimator of residual variance and a measure of explanatory power of covariates in nonparametric regression. *Journal of Nonparametric Statistics* 9(4) 363–380.

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

LI, Q. and S. WANG (1998). A simple consistent bootstrap test for a parametric regression functional form. *Journal of Econometrics* 87 145–165.

LORENTZ, G.G. (1966). *Approximation of functions*. Holt, Rinehart, and Winston.

NEWEY, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79 147–168.

RAMSEY, J.B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31, 350–371.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* 12 1215–1230.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2) 461–464.

SILVERMAN B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* 47 (1) 1–52.

SPOKOINY, V.G. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics* 24 (6) 2477–2498.

SPOKOINY, V.G. (1999). Data-driven testing the fit of linear models. Weierstrass Institute, Berlin. Available at `www.wias.berlin.de/private/spokoiny`.

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

WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76 419–433.

WU, C. F. J. (1986). Jacknife, Bootstrap and other Resampling Methods in Regression Analysis (with discussion). *Annals of Statistics* 14 1261–1350.

YANAGIMOTO, T. and M. YANAGIMOTO (1987). The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model. *Technometrics* 29 95–101.

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

Table 1: White-noise Gaussian model

(r,t)	HS test	Our test		
		$c=1$	$c=1.5$	$c=2$
$(0,\cdot)$	2.1	2.0	2.0	2.0
	5.0	4.8	4.9	4.7
	10.6	10.3	10.1	10.0
$(1/3,2)$	78.4	82.1	82.2	82.4
	89.6	91.2	91.3	91.6
	95.0	96.1	96.2	96.6
$(1/3,5)$	34.9	43.4	43.8	43.9
	53.7	60.7	61.0	60.3
	69.8	74.3	74.4	72.7
$(1/3,10)$	25.5	31.9	32.2	32.9
	43.5	49.5	50.5	51.8
	61.1	65.9	66.6	66.6

Each cell contains empirical rejection percentages at 2%, 5% and 10% nominal levels from top to bottom.

Table 2: Linear model — Normal errors

(r,t)	HS test	Our test		
		$c=1$	$c=1.5$	$c=2$
$(0,\cdot)$	1.9	2.3	2.3	2.3
	5.0	5.0	5.0	4.8
	10.3	10.3	10.3	9.6
$(1/3,2)$	85.9	86.5	86.4	86.4
	93.0	93.3	93.3	93.2
	96.8	97.0	97.0	97.3
$(1/3,5)$	46.8	50.6	50.6	50.0
	63.9	66.6	66.2	65.0
	77.2	78.9	78.3	76.3
$(1/3,10)$	35.1	37.3	37.6	38.4
	53.1	55.3	56.1	56.7
	68.8	70.5	71.0	71.1

Each cell contains empirical rejection percentages at 2%, 5% and 10% nominal levels from top to bottom.

Table 3: Linear model — Exponential errors

(r,t)	HS test	Our test		
		$c=1$	$c=1.5$	$c=2$
$(0,\cdot)$	2.5	2.4	2.5	2.3
	5.2	5.4	5.4	5.4
	9.4	9.4	9.5	9.6
$(1/3,2)$	84.6	85.5	85.3	85.3
	90.8	91.4	91.5	91.4
	94.7	95.0	95.0	95.3
$(1/3,5)$	49.8	53.7	53.7	53.5
	66.1	68.9	68.5	67.6
	77.9	79.5	78.6	77.1
$(1/3,10)$	38.9	41.3	41.7	42.3
	56.8	59.0	59.5	60.2
	72.0	73.6	74.0	73.3

Each cell contains empirical rejection percentages at 2%, 5% and 10% nominal levels from top to bottom.

Table 4: Linear model — Student errors

(r,t)	HS test	Our test		
		$c=1$	$c=1.5$	$c=2$
$(0,\cdot)$	1.8	1.8	1.7	1.8
	4.2	4.5	4.3	4.3
	9.4	9.1	9.0	8.9
$(1/3,2)$	84.7	85.8	85.7	85.7
	91.9	92.5	92.3	92.5
	96.0	96.2	96.2	96.4
$(1/3,5)$	48.8	53.0	52.9	52.7
	65.5	68.2	67.8	66.7
	77.6	79.5	79.0	77.0
$(1/3,10)$	37.3	39.8	40.1	40.6
	54.7	57.1	57.8	58.3
	70.5	72.4	72.8	72.6

Each cell contains empirical rejection percentages at 2%, 5% and 10% nominal levels from top to bottom.