The Annals of Statistics
1997, Vol. 25, No. 1, 212–243

SEMIPARAMETRIC SINGLE INDEX VERSUS FIXED LINK FUNCTION MODELLING[1]

By W. Härdle, V. Spokoiny and S. Sperlich

Humboldt Universität and
Weierstrass Institute for Applied Analysis and Stochastics

Discrete choice models are frequently used in statistical and econometric practice. Standard models such as logit models are based on exact knowledge of the form of the link and linear index function. Semiparametric models avoid possible misspecification but often introduce a computational burden especially when optimization over nonparametric and parametric components are to be done iteratively. It is therefore interesting to decide between approaches. Here we propose a test of semiparametric versus parametric single index modelling. Our procedure allows the (linear) index of the semiparametric alternative to be different from that of the parametric hypothesis. The test is proved to be rate-optimal in the sense that it provides (rate) minimal distance between hypothesis and alternative for a given power function.

1. Introduction. Discrete choice models are frequently used in statistical and econometric applications. Among them, binary response models, such as Probit or Logit regression, dominate the applied literature. A basic hypothesis made there is that the link and the index function have a known form; see McCullagh and Nelder (1989). The fixed form of the link function, for example, by logistic cdf is rarely justified by the context of the observed data but is often motivated by numerical convenience and by reference to "standard practice," say "accessible canned software."

Recent theoretical and practical studies have questioned this somewhat rigid approach and have proposed a more flexible semiparametric approach. Green and Silverman (1994) use the theory of penalizied likelihood to model nonparametric link function with splines. Horowitz (1993) gives an excellent survey on single index methods and stresses economic applications. Severini and Staniswalis (1994) use kernel methods and keep a fixed link function but allow the index to be of partial linear form. Partial linear models are semiparametric models with a parametric linear and a nonparametric index and have been studied by Rice (1986), Speckman (1988) and Engle, Granger, Rice and Weiss (1986).

These models enhance the class of generalized linear models [McCullagh and Nelder (1989)] in several ways. Here we concentrate on one generaliza-

Received March 1995; revised June 1996.
[1]Research supported by Deutsche Forschungsgemeinschaft SFB 373.
AMS 1991 *subject classifications*. Primary 62G10, 62H40; secondary 62G20, 62P20.
Key words and phrases. Semiparametric models, single index model, hypothesis testing.

tion, the single index models with link functions of unknown nonparametric form but (linear) index function. The advantage of this approach is that an interpretable linear single index, a weighted sum of the predictor variables, is still produced. The link function plays, in theoretical justifications of single index models via stochastic utility functions, an important role [Maddala (1983)]: it is the cdf of the errors in a latent variable model. Our approach enables us to interpret the results still in terms of a stochastic utility model but enhances it by allowing for an unknown cdf of the errors.

Despite the flexibility gained in semiparametric regression modelling, there is still an important gap between theory and practice, namely a device for testing between a parametric and semiparametric alternative. A first paper in bridging this gap is Horowitz and Härdle (1994). They considered for response Y and predictor X the parametric null hypothesis

$$H_0: Y = F\big(X^\top \theta_0\big) + \varepsilon, \tag{1}$$

where $x^\top \theta$ denotes the index and F is the fixed and known link function. The semiparametric alternative considered there is that the regression function has the form $f(x^\top \theta_0)$ with a nonparametric link function f and the same index $x^\top \theta_0$ as under H_0. The main drawback of that paper is that the index is supposed to be the same under the null and the alternative.

The goal of the present paper is to construct a test which has power for a large class of alternatives. We move also to a full semiparametric alternative by considering alternatives of single index type:

$$H_1: Y = f\big(X^\top \beta\big) + \varepsilon \tag{2}$$

with β possibly different from θ_0. In addition we consider a more general H_0 than in Horowitz and Härdle (1994), namely a parametric family $(F_\theta,\ \theta \in \Theta)$, thereby allowing link function and error distribution to depend on an arbitrary parameter.

The situation of our test is illustrated in Figures 1 and 2. The data is a crosssection of 462 records on apprenticeship of the German Social Economic Panel from 1984 to 1992. The dependent variable is an indicator of unemployment, ($Y = 1$ corresponds to "yes"). Explanatory variables are X_1, gross monthly earnings as an apprentice, X_2, percentage of people apprenticed in a certain occupation and X_3, unemployment rate in the state the respondent lived in during the year the apprenticeship was completed. The aim of the test is to decide between the logit model and the semiparametric model with unknown link function and possibly different index. In Härdle, Klinke and Turlach (1995) this hypothesis is tested with the Horowitz–Härdle (HH) test by Proenca and Werwatz who also prepared the dataset. They give a more detailed description of the HH test procedure which does not reject.

We measure the quality of a test by the value of minimal distance between the regression function under the null and under the alternative which is sufficient to provide the desirable power of testing. The test proposed below is shown to be rate-optimal in this sense.

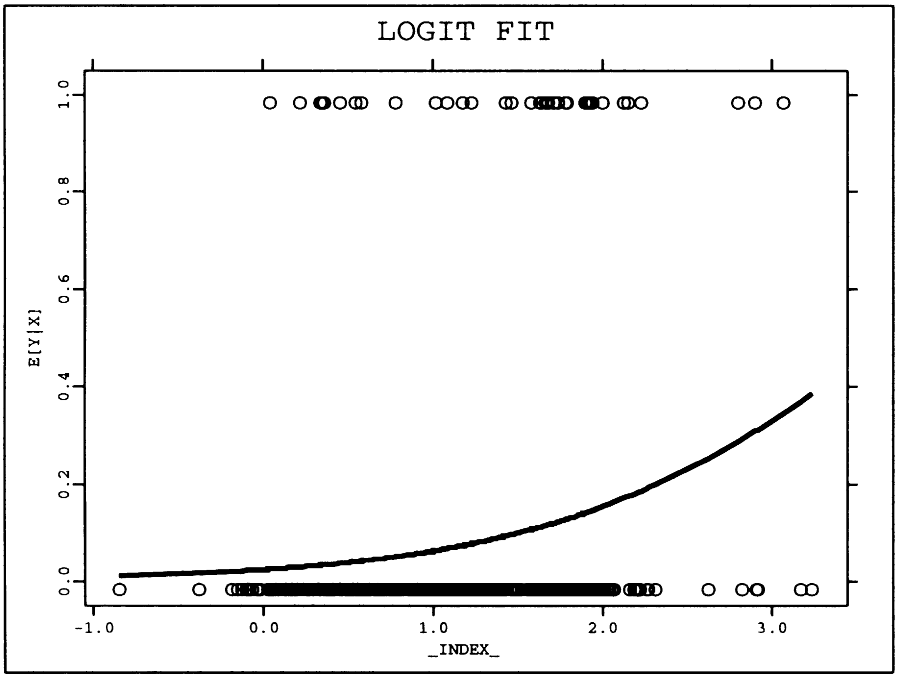

FIG. 1. *Parametric fitting.*

The paper is organized as follows. Section 2 contains the main results. The test procedure is described in Section 3. In Section 5 we present a simulation study. The proofs of main results are given in Section 4 (Theorem 2.2) and in the Appendix (Theorem 2.1).

2. Main results. We start with a brief historical background of the nonparametric hypothesis testing problem. The problem for the case of a simple hypothesis and univariate nonparametric alternative was considered by Ibragimov and Khasminskii (1977) and Ingster (1982). It was shown that the minimax rate for the distance between the null and the alternative set is of the order $n^{-2s/(4s+1)}$ where s is a measure of smoothness. Note that this rate differs from that of an estimation problem where we have $n^{-s/(2s+1)}$. In the multivariate case the corresponding rate changes to $n^{-2s/(4s+d)}$, as Ingster (1993) has shown. The problem of testing a parametric hypothesis versus a nonparametric alternative was discussed also in Härdle and Mammen (1993). Their results allow the extraction of the above minimax rate.

The results of Friedman and Stuetzle (1984), Huber (1985), Hall (1989) and Golubev (1992) show that estimation of the function f under (2) can be made with the rate corresponding to the univariate case. Below we will see,

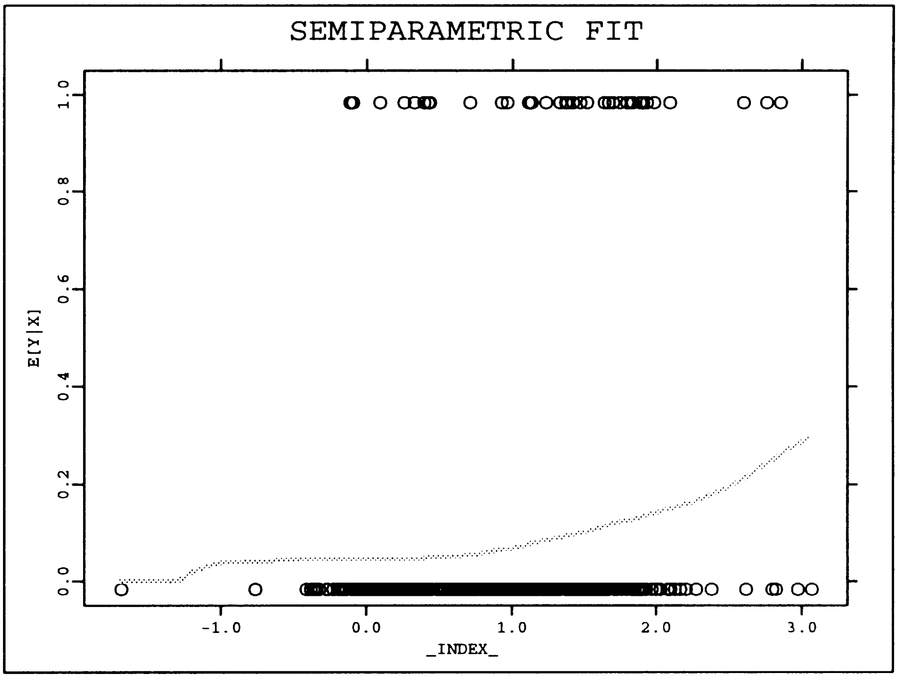

FIG. 2. *Semiparametric fitting.*

though, that for the problem of hypothesis testing the situation is slightly different. The rate for this additive alternative of single index type differs from that of a univariate alternative ($d = 1$) by an extra log factor. Nevertheless, we have a nearly univariate rate and we can therefore still expect efficiency of the test for practical applications.

We will come back to the introductory example in Section 5. Suppose we are given independent observations (X_i, Y_i), $X_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}^1$, $i = 1, \dots, n$, that follow the regression

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n. \tag{3}$$

Here $\varepsilon_i = Y_i - F(X_i)$ are mean zero error variables,

$$\mathbf{E}\varepsilon_i = 0, \qquad i = 1, \dots, n,$$

with conditional variance

$$\sigma_i^2 = \mathbf{E}\left[\varepsilon_i^2 | X_i\right], \qquad i = 1, \dots, n. \tag{4}$$

EXAMPLE 2.1. As a first example, take the above single index binary choice model. The observed response variables Y_i take two values 0, 1 and

$$\mathbf{P}(Y_i = 1 | X_i) = F(X_i),$$
$$\mathbf{P}(Y_i = 0 | X_i) = 1 - F(X_i).$$

In this case, $\sigma_i^2 = F(X_i)\{1 - F(X_i)\}$.

EXAMPLE 2.2. A second example is a nonlinear regression model with unknown transformation. An excellent introduction into nonlinear regression can be found in Huet, Jolivet and Messeau (1993). The model takes the same form as (1) but the response Y is not necessarily binary and the variance σ_i^2 may be an unknown function of the $F(X_i)$'s. Carroll and Ruppert (1988) use this kind of error structure to model fan-shaped residual structure.

We wish to test the hypothesis H_0 that the regression function $F(x)$ belongs to a prescribed parametric family $[F_\theta(x),\ \theta \in \Theta]$, where Θ is a subset in a finite-dimensional space $\mathbb{R}^m$. This hypothesis is tested versus the semiparametric alternative H_1 that the regression function $F(\cdot)$ is of the form

$$F(x) = f(x^\top \beta), \tag{5}$$

where β is a vector in $\mathbb{R}^d$ with $|\beta| = 1$, and $f(\cdot)$ is a univariate function.

EXAMPLE 2.3. Let the parametric family $[F_\theta(x),\ \theta \in \Theta]$ be of the form

$$F_\theta(x) = \frac{1}{1 + \exp(-x^\top \theta)} \tag{6}$$

and let otherwise (X, Y) have stochastic structure as in Example 2.1. This form of parametrization leads to a binary choice logit regression model. Probit or complementary log-log models have a different parametrization but still have this single index form.

Let $\mathscr{F}_0$ be the set of functions $[F_\theta(x),\ \theta \in \Theta]$ and let $\mathscr{F}_1$ be a set of alternatives of the form (5). We measure the power of a test φ_n by its power function on the sets $\mathscr{F}_0$ and $\mathscr{F}_1$: if $\varphi_n = 0$ then we accept the hypothesis H_0 and if $\varphi_n = 1$, then we accept H_1. The corresponding first and second type error probabilities are defined as usual:

$$\alpha_0(\varphi_n) = \sup_{F \in \mathscr{F}_0} \mathbf{P}_F(\varphi_n = 1),$$

$$\alpha_1(\varphi_n) = \sup_{F \in \mathscr{F}_1} \mathbf{P}_F(\varphi_n = 0).$$

Here $\mathbf{P}_F$ means the distributions of observations (X_i, Y_i) given the regression function $F(\cdot)$. When there is no risk of confusion, we write $\mathbf{P}$ instead of $\mathbf{P}_F$. Our goal is to construct a test φ_n that has power over a wide class of alternatives. The assumptions needed are made precise below. We start with assumptions on the error distribution.

(E1) There are $\lambda > 0$ and C_e such that

$$E \exp\{\lambda |\varepsilon_i|\} \le C_e, \qquad i = 1, \dots, n.$$

(E2) The conditional distributions of errors ε_i given X_i depend only on values of the regression function $F(X_i)$,

$$\mathscr{L}(\varepsilon_i | X_i) = \mathscr{L}(\varepsilon_i | F(X_i)) = P_{F(X_i)},$$

where (P_z) is a prescribed distribution family with one-dimensional parameter z.

(E3) The variance function $\sigma^2(z) = \mathbf{E}[\varepsilon_i^2 | F(X_i) = z]$ and the fourth central moment function $\kappa^4(z) = \mathbf{E}[(\varepsilon_i^2 - \mathbf{E}\varepsilon_i^2)^2 | F(X_i) = z]$ are bounded away from zero and infinity, that is,

$$0 < \alpha_* \le \sigma(z) \le \sigma^* < \infty$$
$$0 < \kappa_* \le \kappa(z) \le \kappa^* < \infty$$

with some prescribed σ_*, σ^*, κ_*, κ^* and these functions are Lipschitz: for some positive constants C_σ and C_κ one has

$$|\sigma(z) - \sigma(z')| \le C_\sigma |z - z'|,$$
$$|\kappa(z) - \kappa(z')| \le C_\kappa |z - z'|.$$

Note that (E1) through (E3) are obviously fulfilled for the single index model in Example 2.1 and 2.3.

Now we present assumptions on the design X.

(D) The predictor variables X have a design density $\pi(x)$ which is supported on the compact convex set $\mathscr{X}$ in $\mathbb{R}^d$ and is separated from zero and infinity on $\mathscr{X}$.

Assumption (D) is quite common in nonparametric regression analysis. It is apparently fulfilled for the above example on apprenticeship and youth unemployment. As an alternative to (D), one may assume that the design density π is continuous and positive on some compact subset D of $\mathscr{X}$. In the last case, only observations on D are to be taken into account and this allows reducing the problem to the situation described in the condition; see, for example, Härdle, Hall and Ichimura (1993).

We now specify the hypothesis and alternative.

(H0) The parameter set Θ is a compact subset in $\mathbb{R}^m$.

For some positive constant C_Θ the following holds:

$$|F_\theta(x) - F_{\theta'}(x)| \le C_\Theta |\theta - \theta'| \qquad \forall\, x \in \mathscr{X},\ \theta, \theta' \in \Theta.$$

All functions $F_\theta(\cdot)$ belong to the Hölder class $\Sigma_d(s, L)$ of functions in $\mathbb{R}^d$.

(H1) The univariate link function $f(\cdot)$ from (5) belongs to the Hölder class $\Sigma(s, L)$. The function $F(x) = f(x^\top \beta)$ is bounded away from the parametric family $\mathscr{F}_0$, that is,

$$\inf_{\theta \in \Theta} \|F - F_\theta\| \ge c_n \tag{7}$$

with a given $c_n > 0$. Here $\|F - F_\theta\| = \int |F(x) - F_\theta(x)|^2 \pi(x)\, dx$.

For the definition of a Hölder smoothness class in the context of statistical nonparametric problems we refer, for example, to Ibragimov and Khasminskii (1981). In the case of an integer s, the class $\Sigma(s, L)$ can be defined as the class of s-time differentiable functions with the sth derivative bounded in absolute value by L. Assumption (H0) is certainly fulfilled for

Example 2.3 but also in Probit and other generalized linear regression models such as the log linear models.

The main results are given below. We compute first the optimal rate of convergence of the distance c_n distinguishing the null from the alternative. The second theorem states the existence of an optimal test. The test will be given more explicitly in the next section where we also apply it to the above concrete examples. Theorem 2.2 is proved in Section 4 and the proof of Theorem 2.1 is given in the Appendix.

THEOREM 2.1. *Let $c_n = (a(\sqrt{\log n})/n)^{2s/(4s+1)}$. If a is small enough then for any sequence of tests φ_n one has*

$$\liminf_{n\to\infty} \alpha_0(\varphi_n) + \alpha_1(\varphi_n) \geq 1.$$

THEOREM 2.2. *For any constant a^* large enough there is a sequence of tests φ_n^* which distinguish consistently the hypothesis H_0 from the alternative $H_1 = H_1(c_n^*)$ with $c_n^* = (a^*(\sqrt{\log n})/n)^{2s/(4s+1)}$, that is,*

$$\lim_{n\to\infty} \alpha_0(\varphi_n^*) = 0$$

and

$$\lim_{n\to\infty} \alpha_1(\varphi_n^*) = 0.$$

3. The test procedure. Before we describe the test procedure let us introduce some notation. Given functions $F(x)$ and $G(x)$ we denote by

$$\langle F, G\rangle = \frac{1}{n}\sum_{i=1}^{n} F(X_i)G(X_i) \tag{8}$$

the scalar product of the functions F and G. We write also $\langle F\rangle$ instead of $\langle F, F\rangle$ and identify the sequences (Y_i), (ε_i) with the functions $Y(X_i)$ and $\varepsilon(X_i)$. We construct the tests φ_n^* from Theorem 2.2 in several steps.

First, we shall do a preliminary parametric pilot estimation $\tilde{F}_0$ under the null. Second, we estimate the d-dimensional nonparametric regression $\tilde{F}_1$ of $E(Y|x)$ by a kernel estimation. These estimators are used in the approximating of expected value and the variance of the proposed test statistics. In the third step, we estimate for each feasible value of β the corresponding link function f under (H1) as in (2). Finally we compute the test statistic based on comparison of residuals under H_0 and H_1.

3.1. *Parametric pilot estimation.* Let Θ_n be a grid in the parametric set Θ with the step $\log n/\sqrt{n}$. Put

$$\tilde{\theta}_n = \operatorname*{arginf}_{\theta\in\Theta_n}\langle Y - F_\theta\rangle = \operatorname*{arginf}_{\theta\in\Theta_n}\frac{1}{n}\sum_{i=1}^{n}|Y_i - F_\theta(X_i)|^2. \tag{9}$$

Denote also

$$\tilde{F}_0(\cdot) = F_{\tilde{\theta}_n}(\cdot). \tag{10}$$

Note that $\tilde{\theta}_n$ is not necessarily an efficient estimator under the null since we do not correct for the variance function.

REMARK. We define $\tilde{\theta}_n$ as a discretized minimum contrast estimator. The discretization of the pilot parametric estimator is a standard device in this kind of problem. It allows replacing in the proof the random point $\tilde{\theta}_n$ by a nonrandom point θ_n which is the point of the grid closest to θ; see Lemma 4.8.

3.2. *Nonparametric pilot estimation.* For the nonparametric estimation of the expected value and the variance of the test statistic, we shall use the standard kernel technique; see, for example, Härdle (1990) or Müller (1987). More precisely, we use a one-dimensional kernel satisfying the following conditions.

(K1) $K(\cdot)$ is compactly supported.
(K2) $K(\cdot)$ is symmetric.
(K3) $K(\cdot)$ has s continuous derivatives.
(K4) $\int K(t)\,dt = 1$.
(K5) $\int K(t)t^k\,dt = 0$, $k = 1,\dots,s-1$.

Recall from (H0) and (H1) that s denotes the degree of smoothness of the regression function. Note also that (K5) ensures that K is orthogonal to polynomials of order 1 to $s-1$. For a list of kernels satisfying (K1)–(K5), we refer to Müller (1987). A d-dimensional product kernel K_1 is defined as

$$K_1(u_1,\dots,u_d) = \prod_{j=1}^{d} K(u_j). \tag{11}$$

Take now

$$h_1 = n^{-1/(2s+d)}, \tag{12}$$

the rate optimal smoothing bandwidth in d-dimensions, and put

$$\tilde{F}_1(x) = \frac{\sum_{i=1}^n Y_i K_1((x-X_i)/h_1)}{\sum_{i=1}^n K_1((x-X_i)/h_1)}. \tag{13}$$

The nonparametric kernel smoother $\tilde{F}_1$ is the well-known multidimensional Nadaraya–Watson kernel estimator.

3.3. *Estimation under H_1.* We will use the following bandwidth:

$$h = \left(\frac{\sqrt{\log n}}{n}\right)^{2/(4s+1)} \tag{14}$$

for estimation of f in the semiparametric model. Note that this bandwidth is different from the one used for the nonparametric pilot estimation. It is also different from the bandwidth rate $n^{-2/(4s+d)}$ used in obtaining the purely nonparametric rate of testing, $n^{-2s/(4s+d)}$ even for $d = 1$; see Ingster (1993). The additional log factor comes from the semiparametric structure of our alternative and can be viewed as extra payment for reducing the multivariate structure to a univariate one.

In practice, the smoothness parameter s is typically unknown and the bandwidth h is to be chosen in a data-driven way. Unfortunately, the theory for adaptive nonparametric testing is not sufficiently developed and discussion of the possibility of a data-driven bandwidth choice lies beyond the scope of this paper. Recent results from Spokoiny (1996) indicate that the described testing procedure can be performed in an adaptive way without significant loss of power.

Let S_d be the unit sphere in $\mathbb{R}^d$. Denote by $S_{n,d}$ a discrete grid in S_d with the step $b_n = h^{2s+2}$. Let N be the cardinality of $S_{n,d}$

$$N = \#S_{n,d}. \tag{15}$$

For each $\beta \in S_{n,d}$ define

$$K_{h,\beta}(x) = K\left(\frac{x^\top \beta}{h}\right), \qquad x \in \mathbb{R}^d \tag{16}$$

and introduce the smoothing operator $\mathscr{K}_\beta$ with

$$\mathscr{K}_\beta Y(X_i) = \Pi_\beta(X_i) \sum_{j \neq i} Y_j K_{h,\beta}(X_i - X_j), \tag{17}$$

where

$$\Pi_\beta(X_i) = \left(\sum_{j \neq i} K_{h,\beta}(X_i - X_j)\right)^{-1}. \tag{18}$$

Similarly we define $\mathscr{K}_\beta \varepsilon$ and $\mathscr{K}_\beta F$. Note that given β, the values $\mathscr{K}_\beta Y$ estimate f in (2).

3.4. *The test statistic.* We start with some discussion explaining the applied test statistics. First we treat the case with a simple null hypothesis corresponding to the trivial regression function $F \equiv 0$. We assume also that β is known. In this situation, the obvious goodness-of-fit test can be based on $\langle \mathscr{K}_\beta Y \rangle$ or better on $T_\beta^0 = 2(Y, \mathscr{K}_\beta Y\rangle - \langle \mathscr{K}_\beta Y \rangle$. (The latter test statistic has smaller variance.) Under the null, both these statistics are quadratic forms in the errors ε_i and, therefore, U-type statistics. Standard calculation for U-statistics show that T_β^0 has under the null the expectation of order $(nh)^{-1}$ and the variance of order $n^{-2}h^{-1}$, and, being centered and normalized, it is asymptotically normal; see, for example, Hall (1984). The proper test could be taken in the form $\varphi^* = \mathbf{1}\ [(\mathrm{Var}\, T_\beta^0)^{-1/2}(T_\beta^0 - ET_\beta^0) > \chi_\alpha]$ where χ_α is $(1-\alpha)$-quantile of the standard normal law. If a considered model has heteroskedastic noise with variance depending on the unknown regression function (see, e.g., Example 2.1 for the binary response model), then the value ET_β^0 and $\mathrm{Var} T_\beta^0$ are to be estimated using a pilot nonparametric estimator.

In the case of the parametric null, the natural idea could be to construct a pilot parametric estimator $\tilde{F}_0$ and then to substruct it from the observation Y, as was done in Härdle and Mammen (1993). This leads to the test based on $2\langle Y - \tilde{F}_0, \mathscr{K}_\beta(Y - \hat{F}_0)\rangle - \langle \mathscr{K}_\beta(Y - \tilde{F}_0)\rangle$. Unfortunately, this method cannot be applied to the case of the semiparametric alternative since the null hypothesis can have another structure (e.g., with a different index) and hence the value of the bias term in $\mathscr{K}_\beta \tilde{F}_0$ can be too large. To bypass this problem, we modify slightly the latter test statistics, changing $\mathscr{K}_\beta(Y - \tilde{F}_0)$ to $\mathscr{K}_\beta Y - \tilde{F}_0$.

Finally, to proceed with an unknown β, we calculate a statistic T_β for each β from the grid $S_{n,d}$ and the resulting test statistic is based on the supremum of all these quantities. If now β be the value of the index for the unknown alternative, then, at least for one $\beta' \in S_{n,d}$, the difference $\beta' - \beta$ will be small and the corresponding $T_{\beta'}$ will be large enough to detect this alternative. From the other side, considering the maximum of a growing number of test statistics creates a problem with the error probability of the first kind. To provide a prescribed level for this error probability, we have to take a logarithmically growing test level that results in the extra log factor in the rate of testing. The result of Theorem 2.1 shows that this log factor is an unavoidable payment for the choice of an unknown index for the alternative of semiparametric structure.

Now we present the test statistics. For each $\beta \in S_{n,d}$ we calculate T_β as follows:

$$T_\beta = \frac{n\sqrt{h}}{\tilde{V}_\beta}\Big[2\langle Y - \tilde{F}_0, \mathscr{K}_\beta Y - \tilde{F}_0\rangle - \langle \mathscr{K}_\beta Y - \tilde{F}_0\rangle + \tilde{E}_\beta\Big]. \tag{19}$$

Here $\langle \cdot \rangle$ is defined by (8), h by (14), $\tilde{F}_0$ by (10). We use the following notation:

$$\tilde{E}_\beta = \frac{1}{n}\sum_i \sum_{j\neq i} \tilde{\sigma}_j^2 \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j), \tag{20}$$

where $\Pi_\beta(X_i)$ is from (18),

$$\tilde{\sigma}_j^2 = \sigma^2\Big(\tilde{F}_1(X_j)\Big), \qquad j = 1, \dots, n, \tag{21}$$

the function $\sigma^2(\cdot)$ being defined in the model assumptions of $\tilde{F}_1(x)$ being the nonparametric pilot estimator. Finally,

$$\begin{aligned}\tilde{V}_\beta^2 = {} & h\sum_i \sum_{j\neq i} \tilde{\sigma}_i^2 \tilde{\sigma}_j^2 \Pi_\beta^2(X_i)\Big|2K_{h,\beta}(X_i - X_j) - K_{h,\beta}^{(2)}(X_i, X_j)\Big|^2 \\ & + h\sum_i \tilde{\kappa}_i^4 \Big|\sum_{j\neq i} \Pi_\beta^2(X_j) K_{h,\beta}^2(X_i - X_j)\Big|^2\end{aligned}$$

with $\tilde{\kappa}_i = \kappa(\tilde{F}_1(X_i))$, $i = 1, \dots, n$, $\kappa(\cdot)$ being from (E3) and

$$K_{h,\beta}^{(2)}(X_i, X_j) = \frac{1}{\Pi_\beta(X_i)} \sum_{k\neq i,j} \Pi_\beta^2(X_k) K_{h,\beta}(X_k - X_i) K_{h,\beta}(X_k - X_j). \tag{22}$$

Put now

$$T_n^* = \sup_{\beta \in S_{n,d}} T_\beta \tag{23}$$

and

$$\varphi_n^* = \mathbf{1}\big(T_n^* > \sqrt{4 \log N}\big). \tag{24}$$

Here $\mathbf{1}(\cdot)$ is the indicator function of the corresponding event and N is the cardinality of $S_{n,d}$; see (15).

REMARK. Similarly to the case of parametric pilot estimation, we optimize the parameter β over some grid of the unit sphere. This restriction allows simplifying the proof by reducing the random field $(T_\beta, \beta \in S_d)$ to a family of random variables $(T_\beta, \beta \in S_{n,d})$. It may be possible to define the test statistic T_n^* as the supremum of T_β over the whole sphere S_d but the proof in this case seems to be much more involved.

4. Proof of Theorem 2.2. We start with a decomposition of the test statistics T_β. Denote by $B_\beta(x)$ the bias function for the smoothing operator $\mathscr{K}_\beta$ from (17):

$$B_\beta(X_i) = \mathscr{K}_\beta F(X_i) - F(X_i), \qquad i = 1, \dots, n. \tag{25}$$

Fix some $\beta \in S_{n,d}$ and $F \in \mathscr{F}_0 \cup \mathscr{F}_1$.

LEMMA 4.1.

$$\begin{aligned} T_\beta = \frac{n\sqrt{h}}{\tilde{V}_\beta} \Big[&\langle F - \tilde{F}_0 \rangle - \langle B_\beta \rangle + 2\langle \mathscr{K}_\beta \varepsilon, \varepsilon \rangle - \langle K_\beta \varepsilon \rangle + \tilde{E}_\beta \\ &+ 2\langle F - \tilde{F}_0, \varepsilon \rangle + 2\langle B_\beta, \varepsilon \rangle - 2\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle \Big]. \end{aligned} \tag{26}$$

PROOF. By definition, $Y = F + \varepsilon$ and therefore

$$\mathscr{K}_\beta Y = \mathscr{K}_\beta F + \mathscr{K}_\beta \varepsilon = F + B_\beta + \mathscr{K}_\beta \varepsilon.$$

Now

$$\begin{aligned} 2\langle Y - \tilde{F}_0, \mathscr{K}_\beta Y - \tilde{F}_0 \rangle &= 2\langle F - \tilde{F}_0 + \varepsilon, F - \tilde{F}_0 + B_\beta + \mathscr{K}_\beta \varepsilon \rangle \\ &= 2\langle F - \tilde{F}_0 \rangle + 2\langle F - \tilde{F}_0, B_\beta \rangle + 2\langle F - \tilde{F}_0, \mathscr{K}_\beta \varepsilon \rangle \\ &\quad + 2\langle \varepsilon, F - \tilde{F}_0 \rangle + 2\langle \varepsilon, B_\beta \rangle + 2\langle \varepsilon, \mathscr{K}_\beta \varepsilon \rangle \end{aligned}$$

and

$$\begin{aligned} \langle \mathscr{K}_\beta Y - \tilde{F}_0 \rangle &= \langle F - \tilde{F}_0 + B_\beta + K_\beta \varepsilon \rangle \\ &= \langle F - \tilde{F}_0 \rangle + \langle B_\beta \rangle + \langle \mathscr{K}_\beta \varepsilon \rangle + 2\langle F - \tilde{F}_0, B_\beta \rangle \\ &\quad + 2\langle F - \tilde{F}_0, \mathscr{K}_\beta \varepsilon \rangle + 2\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle. \end{aligned}$$

Substituting this in the definition of T_β, we obtain the assertion of the lemma. $\square$

The next step is to show that the expansion (26) for the statistic T_β can be simplified by discarding lower order terms. Indeed we shall see below that the last three terms are relatively small and can be omitted. The terms $\tilde{E}_\beta$ and $\tilde{V}_\beta$ can be replaced by similar expressions E_β and V_β which use "true" values σ_i and κ_i instead of estimated values $\tilde{\sigma}_i$ and $\tilde{\kappa}_i$ and finally, the parametric estimator $\tilde{\theta}_n$ can be replaced by θ_n defined by

$$\theta_n = \operatorname*{arginf}_{\theta \in \Theta_n} \langle F - F_\theta \rangle \tag{27}$$

where F is a "true" regression function from (3). Suppose that all these replacements can be done. Define now

$$T'_\beta = \frac{n\sqrt{h}}{V_\beta}\big[\langle F - F_{\theta_n}\rangle - \langle B_\beta \rangle + 2\langle \mathscr{K}_\beta \varepsilon, \varepsilon\rangle - \langle \mathscr{K}_\beta \varepsilon \rangle + E_\beta\big]$$

and

$$E_\beta = \frac{1}{n}\sum_i \sum_{j \neq i} \sigma_j^2 \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j),$$

$$V_\beta^2 = h\sum_i \sum_{j\neq i} \sigma_i^2 \sigma_j^2 \Pi_\beta^2(X_i)\big|2K_{h,\beta}(X_i - X_j) - K_{h,\beta}^{(2)}(X_i, X_j)\big|^2$$
$$+ h\sum_i \kappa_i^4 \bigg|\sum_{j\neq i} \Pi_\beta^2(X_j) K_{h,\beta}^2(X_k - X_j)\bigg|^2.$$

Below we show that the tests φ_n^{**} based on the statistics T_n^{**} with

$$T_n^{**} = \sup_{\beta \in S_{n,d}} T'_\beta \tag{28}$$

have the same asymptotic behavior as φ_n^*. For the moment we only consider the tests φ_n^{**}. Note that they are not tests in the usual sense since they use the nonobservable values E_β, V_β, θ_n. Central to our proof is the analysis of the asymptotic behavior of the random variables

$$\xi_\beta = n\sqrt{h}\big[2\langle \mathscr{K}_\beta \varepsilon, \varepsilon\rangle - \langle \mathscr{K}_\beta \varepsilon\rangle + E_\beta\big]. \tag{29}$$

LEMMA 4.2. *The following assertions hold*

$$\mathbf{E}\xi_\beta = 0, \tag{30}$$

$$\mathbf{E}\xi_\beta^2 = V_\beta^2, \tag{31}$$

and uniformly in $F \in \mathscr{F}_0 \cup \mathscr{F}_1$, $\beta \in S_{n,d}$ *and* $t \in [-\log n, \log n]$,

$$\frac{\mathbf{P}\big((\xi_\beta/V_\beta) > t\big)}{1 - \Phi(t)} \to 1, \qquad n \to \infty, \tag{32}$$

$\Phi(\cdot)$ *being the standard normal distribution.*

PROOF. The first two statements are derived by direct calculation. In fact, by definition and (22),

$$\begin{aligned}
\xi_\beta &= 2\sqrt{h} \sum_i \varepsilon_i \Pi_\beta(X_i) \sum_{j \neq i} \varepsilon_j K_{h,\beta}(X_i - X_j) \\
&\quad - \sqrt{h} \sum_i \Pi_\beta^2(X_i) \Big| \sum_{j \neq i} \varepsilon_j K_{h,\beta}(X_i - X_j) \Big|^2 \\
&\quad + \sqrt{h} \sum_i \sum_{j \neq i} \sigma_j^2 \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j) \\
&= \sqrt{h} \sum_i \sum_{j \neq i} \varepsilon_i \varepsilon_j \Pi_\beta(X_i) \big[2K_{h,\beta}(X_i - X_j) - K_{h,\beta}^{(2)}(X_i, X_j)\big] \\
&\quad + \sqrt{h} \sum_i \sum_{j \neq i} \big(\sigma_j^2 - \varepsilon_j^2\big) \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j).
\end{aligned}$$

Since the errors ε_i are independent and $\mathbf{E}\varepsilon_i = 0$, $\mathbf{E}\varepsilon_i^2 = \sigma_i^2$, we immediately obtain (30) and (31). The last statement (32) is a particular case of the general central limit theorem for quadratic forms of independent random variables (U-statistics) and can be obtained in a standard way by calculation of the corresponding cumulants. We omit the details; see, for example, Härdle and Mammen (1993). □

The assertion (32) of Lemma 4.2 straightforwardly implies the following corollary.

LEMMA 4.3. *Uniformly in* $F \in \mathscr{F}_0 \cup \mathscr{F}_1$ *one has*

$$\mathbf{P}\left(\sup_{\beta \in S_{n,d}} \frac{\xi_\beta}{V_\beta} > \sqrt{4 \log N} \right) \to 0, \qquad n \to \infty. \tag{33}$$

PROOF. For any t one gets

$$\mathbf{P}\left(\sup_{\beta \in S_{n,d}} \frac{\xi_\beta}{V_\beta} > t \right) \le \sum_{\beta \in S_{n,d}} \mathbf{P}\left(\frac{\xi_\beta}{V_\beta} > t \right) \le N \sup_{\beta \in S_{n,d}} \mathbf{P}\left(\frac{\xi_\beta}{V_\beta} > t \right). \tag{34}$$

However, by (32) for n large enough,

$$\mathbf{P}\left(\frac{\xi_\beta}{V_\beta} > \sqrt{4\log N}\right) \le 2\left(1-\Phi\left(\sqrt{4\log N}\right)\right)$$
$$\le \exp\left\{-\frac{1}{2}\left|\sqrt{4\log N}\right|^2\right\} = N^{-2}$$

that implies (33) through (34). □

Now we come to the calculation of the error probabilities for the tests φ_n^{**} based on T_n^{**}. Under the hypothesis H_0 one has $F = F_\theta$, $\theta \in \Theta$. This does not automatically yield $\langle F - F_{\theta_n}\rangle = 0$ since $\theta_n \in \Theta_n$ [see (27)] and θ can be outside Θ_n, but the assumptions (H0) on the parametric family guarantee that this value is small enough.

LEMMA 4.4. *Let $F = F_\theta$, $\theta \in \Theta$. Then*

$$\langle F_\theta - F_{\theta_n}\rangle \le C_\theta^2 \frac{\ln^2 n}{n}.$$

PROOF. Let

$$\theta_n' = \underset{\theta' \in \Theta_n}{\operatorname{arginf}} |\theta - \theta'|.$$

The definition of the grid Θ_n provides $|\theta - \theta_n'|^2 \le (\ln^2 n)/n$. Now from the definition of θ_n and the assumptions (H0) on the parametric family we obtain

$$\langle F_\theta - F_{\theta_n}\rangle \le \langle F_\theta - F_{\theta_n'}\rangle = \frac{1}{n}\sum_i \left|F_\theta(X_i) - F_{\theta_n'}(X_i)\right|^2 \le C_\Theta^2 |\theta - \theta'|^2$$
$$\le C_\Theta^2 \frac{\ln^2 n}{n}.$$

Using this result, we have for $F = F_\theta$ by Lemma 4.3,

$$\mathbf{P}\left(T_n^{**} > \sqrt{4\log N}\right) \le \mathbf{P}\left(\sup_{\beta \in S_{n,d}} \frac{\xi_\beta}{V_\beta} > \sqrt{4\log N} - C_\theta^2 \frac{\ln^2 n}{n} n\sqrt{h}\right) \to 0,$$
$$n \to \infty,$$

that is,

$$\alpha_0(\varphi_n^{**}) = \sup_{F \in \mathscr{F}_0} \mathbf{P}_F(\varphi_n^{**} = 1) \to 0, \qquad n \to \infty. \qquad \square$$

Next we evaluate the error probability of the second type.

LEMMA 4.5. *Let $F \in \mathscr{F}_1$. Then for n large enough*

$$\langle F - F_{\theta_n}\rangle \ge c_n/2.$$

PROOF. Let $F \in \mathscr{F}_1$ be fixed and

$$\theta_F = \operatorname*{arginf}_{\theta \in \Theta} \|F - F_\theta\|.$$

By the triangle inequality and Lemma 4.4 one has

$$\langle F - F_{\theta_F} \rangle \leq \langle F - F_{\theta_n} \rangle + \langle F_{\theta_n} - F_{\theta_F} \rangle \leq \langle F - F_{\theta_n} \rangle + C_\theta^2 \frac{\ln^2 n}{n}.$$

It remains to check that the inequality $\|F - F_{\theta_F}\| \geq c_n$ implies $\langle F - F_{\theta_F} \rangle \geq c_n/2$. For n large enough that is obviously the case. □

The following lemma is a direct consequence of assumptions (E3) and (D).

LEMMA 4.6. *There exist constants* C_π, σ^* *and* V^* *such that*

$$\left|\Pi_\beta(X_i) K_{h,\beta}(X_i - X_j)\right| \leq C_\pi \left|\Pi_\beta(X_j) K_{h,\beta}(X_i - X_j)\right| \qquad \forall\, \beta, X_i, X_j, \tag{35}$$

$$\sup_i \sigma_i \leq \sigma^* \tag{36}$$

and

$$\sup_\beta V_\beta \leq V^*. \tag{37}$$

Recall now that each function $F(\cdot)$ from $\mathscr{F}_1$ is of the form $F(x) = f(x^\top \beta_0)$ with some $\beta_0 \in S_d$. As a consequence $F(\cdot)$ should be well approximated by the smoothing operator $\mathscr{K}_\beta$ with β coinciding or close to β_0. More precisely, the following can be stated.

LEMMA 4.7. *There is a positive constant* C_b *such that for each* $F(\cdot) \in \mathscr{F}_1$, $F(x) = f(x^\top \beta_0)$,

$$\langle B_{\beta_n} \rangle \leq C_b h^{2s} \tag{38}$$

with

$$\beta_n = \operatorname*{arginf}_{\beta \in S_{n,d}} |\beta - \beta_0|. \tag{39}$$

PROOF. The definition of the grid $S_{n,d}$ provides $|\beta_n - \beta_0| \leq h^{2s+2}$. Then, it is well known, for example, from Ibragimov and Khasminskii (1981), that for $F(x) = f(x^\top \beta_0)$ with $f \in \Sigma(s, L)$ one has

$$\langle B_{\beta_0} \rangle = \langle \mathscr{K}_{\beta_0} F - F \rangle \leq L' h^{2s+1} \tag{40}$$

with $L' = L\|K\|/(s-1)!$, but

$$\begin{aligned}\left|\langle B_{\beta_n}\rangle - \langle B_{\beta_0}\rangle\right| &\le \langle B_{\beta_n} - B_{\beta_0}\rangle\\ &\le \langle \mathscr{K}_{\beta_0} F - \mathscr{K}_{\beta_n} F\rangle\\ &\le \frac{1}{n}\sum_i \Bigg|\Pi_{\beta_n}(X_i)\sum_{j\neq i} F(X_j)K_{h,\beta_n}(X_i - X_j)\\ &\qquad - \Pi_{\beta_0}(X_i)\sum_{j\neq i} F(X_j)K_{h,\beta_0}(X_i - X_j)\Bigg|.\end{aligned}$$

Now using assumptions (D) and (K1)–(K5) we obtain

$$\begin{aligned}\left|\Pi_{\beta_n}^{-1}(X_i) - \Pi_{\beta_0}^{-1}(X_i)\right| &\le \sum_{j\neq i}\left|K_{h,\beta_n}(X_i - X_j) - K_{h,\beta_0}(X_i - X_j)\right|\\ &\le C\Pi_{\beta_0}(X_i)\frac{|\beta_n - \beta_0|}{h}\end{aligned}\tag{41}$$

and similarly

$$\begin{aligned}&\sum_{j\neq i}\left|F(X_j)K_{h,\beta_n}(X_i - X_j) - F(X_j)K_{h,\beta_0}(X_i - X_j)\right|\\ &\qquad \le C\Pi_{\beta_0}(X_i)\frac{|\beta_n - \beta_0|}{h}.\end{aligned}\tag{42}$$

Putting together (41) and (42) we conclude that

$$\left|\langle B_{\beta_n}\rangle - \langle B_{\beta_0}\rangle\right| \le C\frac{|\beta_n - \beta_0|}{h} \le Ch^{2s+1}$$

and the lemma follows from $C_b = L' + 1$.

To complete the proof for the tests φ_n^{**} it remains to note that for each $F \in \mathscr{F}_1$,

$$T_n^{**} \ge \frac{n\sqrt{h}}{V_{\beta n}}\left|\langle F - F_{\theta_n}\rangle - \langle B_{\beta_n}\rangle\right| + \frac{\xi_{\beta_n}}{V_{\beta_n}}$$

and that if

$$\langle F - F_{\theta_n}\rangle \ge C_b h^{2s} + \frac{2V^*}{n\sqrt{h}}\sqrt{4\log N},\tag{43}$$

with V^* from Lemma 4.6, then by Lemma 4.3 we obtain

$$\begin{aligned}\mathbf{P}\left(T_n^{**} < \sqrt{4\log N}\right) &\le \mathbf{P}\left(\frac{n\sqrt{h}}{V_{\beta_n}}\,\frac{2V^*}{n\sqrt{h}}\sqrt{4\log N} + \frac{\xi_{\beta_n}}{V_{\beta_n}} < \sqrt{4\log N}\right)\\ &\le \mathbf{P}\left(\left|\frac{\xi_{\beta_n}}{V_{\beta_n}}\right| > \sqrt{4\log N}\right) \to 0, \qquad n\to\infty.\end{aligned}$$

Finally we remark that $\log N \le C \log n$ and the choice of h by (8) yields

$$C_b h^{2s} + \frac{2V^*}{n\sqrt{h}}\sqrt{4\log N} \le C'\left(\frac{\sqrt{\log n}}{n}\right)^{4s/(4s+1)} = C'h^{2s},$$

that is, (43) holds true if c_n in the definition of the alternative H_1 is taken with $c_n^2 \ge 2C'h^{2s}$. This completes the proof for the tests φ_n^{**}. □

Now we explain why the statistics T_n^{**} can be considered in place of T_n^*. The idea is to show that the difference $T_n^{**} - T_n^*$ is relatively small (being compared with the test level $\sqrt{4\log N}$ or deviation $\langle F - F_{\theta_n}\rangle$). First we treat the preliminary parametric estimator $\tilde{\theta}_n$. Denote for given $F \in \mathscr{F}_0 \cup \mathscr{F}_1$,

$$d_n(F) = \langle F - F_{\theta_n}\rangle + \frac{\ln^2 n}{n},$$

θ_n being from (27).

LEMMA 4.8. *Uniformly in $F \in \mathscr{F}_0 \cup \mathscr{F}_1$ we have for each $\delta > 0$,*

$$\mathbf{P}\left(\frac{1}{d_n(F)}\left|\langle F - \tilde{F}_0\rangle - \langle F - F_{\theta_n}\rangle\right| > \delta\right) \to 0,$$
$$\mathbf{P}\left(\frac{1}{d_n(F)}\left|\langle F - \tilde{F}_0, \varepsilon\rangle\right| > \delta\right) \to 0. \tag{44}$$

PROOF. Let us fix some $\delta > 0$ and some $\theta \in \Theta_n$. First we show that the probability of the event

$$\left\{|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right\}$$

is asymptotically small. More precisely, we state the following assertion:

$$\sum_{\theta\in\Theta_n}\mathbf{P}\left(|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right) \to 0, \qquad n \to \infty. \tag{45}$$

In fact, if we put $d_\theta^2 = \mathbf{E}|\langle F - F_\theta, \varepsilon\rangle|^2$, then we have

$$\begin{aligned} d_\theta^2 &= \mathbf{E}\left|\frac{1}{n}\sum_i \varepsilon_i[F(X_i) - F_\theta(X_i)]\right|^2 \\ &= \frac{1}{n^2}\sum_i \sigma_i^2|F(X_i) - F_\theta(X_i)|^2. \end{aligned}$$

Using Lemma 4.6 we have

$$d_\theta^2 \le \frac{\sigma^{*2}}{n^2}\sum_i |F(X_i) - F_\theta(X_i)|^2 = \frac{\sigma^{*2}}{n}\langle F - F_\theta\rangle.$$

Further,

$$\frac{1}{2}\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right) \geq \sqrt{\langle F - F_\theta\rangle \frac{\ln^2 n}{n}}$$

and

$$\begin{aligned}
&\mathbf{P}\left(|\langle F - F_\theta, \varepsilon\rangle| > \delta\left(\langle F - F_\theta\rangle + \frac{\ln^2 n}{n}\right)\right)\\
&\quad \leq \mathbf{P}\left(\frac{1}{d_\theta}|\langle F - F_\theta, \varepsilon\rangle| > \frac{\delta}{d_\theta}\log n \sqrt{\frac{\langle F - F_\theta\rangle}{n}}\right)\\
&\quad \leq \mathbf{P}\left(\frac{1}{d\theta}|\langle F - F_\theta, \varepsilon\rangle| > \frac{\delta}{\sigma^*}\log n\right).
\end{aligned}$$

Now we use an estimate of the large deviation probability for the centered and normalized random variables $(1/d_\theta)\langle F - F_\theta, \varepsilon\rangle$; see Lemma 4.11. Indeed, for n large enough,

$$\begin{aligned}
\sum_{\theta\in\Theta_n} \mathbf{P}\left(\frac{1}{d_\theta}\langle F - F_\theta, \varepsilon\rangle > \frac{\delta}{\sigma^*}\log n\right) &\leq \sum_{\theta\in\Theta_n} \exp\left\{-\frac{\delta^2}{2\sigma^{*2}}\ln^2 n\right\}\\
&\leq n^d \exp\{-(d+1)\log n\} \leq n^{-1}
\end{aligned}$$

which implies (45). Here we used that the cardinality of Θ_n is of order n^d. Let $\theta \in \Theta_n$ be such that

$$\langle F - F_\theta\rangle - \langle F - F_{\theta_n}\rangle > 2\delta d_n(F). \tag{46}$$

For δ small enough this yields

$$\langle F - F_\theta\rangle - \langle F - F_{\theta_n}\rangle > \delta\big(\langle F - F_\theta\rangle + \langle F - F_{\theta_n}\rangle\big). \tag{47}$$

Now by definition of $\tilde{\theta}_n$, we obtain by (46) and (47)

$$\begin{aligned}
\big\{\tilde{\theta}_n = \theta\big\} &\subseteq \big\{\langle Y - F_\theta\rangle \leq \langle Y - F_{\theta_n}\rangle\big\}\\
&= \big\{\langle F - F_\theta + \varepsilon\rangle \leq \langle F - F_{\theta_n} + \varepsilon\rangle\big\}\\
&= \big\{\langle F - F_\theta\rangle - \langle F - F_{\theta_n}\rangle \leq 2\langle F - F_\theta, \varepsilon\rangle + 2\langle F - F_{\theta_n}, \varepsilon\rangle\big\}\\
&\subseteq \left\{\langle F - F_\theta, \varepsilon\rangle > \frac{\delta}{2}\langle F - F_\theta\rangle\right\} \cup \left\{\langle F - F_{\theta_n}, \varepsilon\rangle > \frac{\delta}{2}\langle F - F_{\theta_n}\rangle\right\}.
\end{aligned}$$

Using this relation and (45) we deduce

$$\begin{aligned}
&\mathbf{P}\big(\big|\langle F - F_{\tilde{\theta}_n}\rangle - \langle F - F_{\theta_n}\rangle\big| > 2\delta d_n(F)\big)\\
&\quad \leq \sum_{\theta\in\Theta_n} \mathbf{1}\big(\big|\langle F - F_\theta\rangle - \langle F - F_{\theta_n}\rangle\big| > 2\delta d_n(F)\big)\mathbf{P}\big(\tilde{\theta}_n = \theta\big)\\
&\quad \leq \sum_{\theta\in\Theta_n} \mathbf{P}\left(\langle F - F_\theta, \varepsilon\rangle > \frac{\delta}{2}\langle F - F_\theta\rangle\right) \to 0, \qquad n \to \infty;
\end{aligned}$$

that proves (44). The second statement of the lemma follows directly from (45). □

The next step is to show that the last two terms in the expansion (29) are vanishing.

LEMMA 4.9. *Given F let*

$$b_\beta = \langle B_\beta \rangle + \frac{\ln^2 n}{n}.$$

Then uniformly in $F \in \mathscr{F}_0 \cup \mathscr{F}_1$ for each $\delta > 0$, the following assertions hold:

$$\sum_{\beta \in S_{n,d}} \mathbf{P}\big(\langle B_\beta, \varepsilon \rangle > \delta b_\beta\big) \to 0,$$

$$\sum_{\beta \in S_{n,d}} \mathbf{P}\big(\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle > \delta b_\beta\big) \to 0.$$

REMARK 4.1. The statements of this lemma yield immediately that

$$\mathbf{P}\big(\langle B_\beta, \varepsilon \rangle \le \delta b_\beta \qquad \forall\, \beta \in S_{n,d}\big) \to 1$$

and similarly for $\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle$.

PROOF. The statements of the lemma are proved in the same manner as in the last part of the proof of Lemma 4.8. For the second statement, we use in addition the fact that

$$\operatorname{Var}\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle \le \frac{C}{n} \langle B_\beta \rangle. \tag{48}$$

Indeed, using assumptions (E1)–(E3) and (K1)–(K5), Lemma 4.6 and Jensen's inequality, we have

$$\begin{aligned}
\mathbf{E}\big|\langle B_\beta, \mathscr{K}_\beta \varepsilon \rangle\big|^2 &= \frac{1}{n^2} \mathbf{E}\Big|\sum_i B_\beta(X_i) \Pi_\beta(X_i) \sum_{j \ne i} \varepsilon_j K_{h,\beta}(X_i - X_j)\Big|^2 \\
&= \frac{1}{n^2} \mathbf{E}\Big|\sum_j \varepsilon_j \sum_{i \ne j} B_\beta(X_i) \Pi_\beta(X_i) K_{h,\beta}(X_i - X_j)\Big|^2 \\
&= \frac{1}{n^2} \sum_j \sigma_j^2 \Big|\sum_{i \ne j} B_\beta(X_i) \Pi_\beta(X_i) K_{h,\beta}(X_i - X_j)\Big|^2 \\
&\le \frac{1}{n^2} \sigma^{*2} C_\pi^2 \sum_j \Pi_\beta^2(X_j) \Big|\sum_{i \ne j} B_\beta(X_i) K_{h,\beta}(X_i - X_j)\Big|^2 \\
&\le \frac{1}{n^2} \sigma^{*2} C_\pi^2 \sum_j \left|\frac{\sum_{i \ne j} B_\beta(X_i) K_{h,\beta}(X_i - X_j)}{\sum_{i \ne j} K_{h,\beta}(X_i - X_j)}\right|^2 \\
&\le \frac{1}{n^2} \sigma^{*2} C_\pi^2 C \langle B_\beta \rangle.
\end{aligned}$$

□

Next we show that the quantities $\tilde{E}_\beta$ and $\tilde{V}_\beta$ estimate E_β and V_β well enough.

LEMMA 4.10. *For each* $\delta > 0$ *and uniformly in* $F \in \mathscr{F}_0 \cup \mathscr{F}_1$,

$$\mathbf{P}\left(\sup_{\beta \in S_{n,d}} |\tilde{E}_\beta - E_\beta| > \frac{1}{n\sqrt{h}\,\log n} \right) \to 0,$$

$$\mathbf{P}\left(\sup_{\beta \in S_{n,d}} \left| \frac{\tilde{V}_\beta}{V_\beta} - 1 \right| > \delta \right) \to 0.$$

PROOF. The assumption (E3) implies for each $j = 1, \dots, n$,

$$|\sigma_j^2 - \tilde{\sigma}_j^2| \le C_\sigma \left| \tilde{F}_1(X_j) - F(X_j) \right|$$

and hence

$$|\tilde{E}_\beta - E_\beta| \le \frac{1}{n} \sum_i \sum_{j \ne i} |\sigma_j^2 - \tilde{\sigma}_j^2| \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j).$$

Now by the design and kernel properties we derive for each $j = 1, \dots, n$,

$$\sum_{j \ne i} \Pi_\beta^2(X_i) K_{h,\beta}^2(X_i - X_j) \le \frac{C}{nH}$$

and using the Cauchy–Schwarz inequality we obtain

$$|\tilde{E}_\beta - E_\beta| \le \frac{C}{n^2 h} \sum_i \left| \tilde{F}_1(X_j) - F(X_j) \right| \le \frac{C}{n^2 h} \left[\frac{1}{n} \sum_i \left| \tilde{F}_1(X_j) - F(X_j) \right|^2 \right]^{1/2}.$$

The pilot estimator $\tilde{F}_1$ fulfills with high probability,

$$\langle \tilde{F}_1 - F \rangle \le C n^{-2s/(2s+d)}.$$

Hence using the inequality $2s/(2s + d) > 1/(4s + 1)$ and the definition of h we arrive at the conclusion that

$$n\sqrt{h}\,|\tilde{E}_\beta - E_\beta| \le \frac{C}{\sqrt{h}} n^{-2/(2s+d)} = o\left(\frac{1}{\log n} \right). \qquad \square$$

Lemmas 4.8–4.10 together imply the asymptotic equivalence of the tests based on T_β and T'_β. We finish the proof of the theorem with a result on probabilities of deviations of centered and normalized sums of independent errors ε_i over the logarithmic level. The following lemma was already used in the proof of Lemma 4.8.

LEMMA 4.11. *For each pair of positive constants* r, a, *the following relation holds uniformly in functions* F *from the Lipschitz class* $\Sigma_d(1, L)$ *of functions in* $\mathbb{R}^d$:

$$n^r \mathbf{P}(\xi(F) > a \log n) \to 0, \qquad n \to \infty,$$

where

$$\xi(F) = \frac{\langle F, \varepsilon \rangle}{\sqrt{\mathbf{E}\langle F, \varepsilon \rangle^2}}.$$

For the proof, we proceed in a standard way using (E1) through (E3) and the Chebyshev exponential inequality: for any $\lambda, z > 0$, one has $P(\xi > z) \le e^{-\lambda z} E e^{\lambda \xi}$. The details are omitted.

5. A simulation and an application. The purpose of our simulation experiments was to study the quantiles of the test statistic T_n^* and the power of the test in finite samples. All calculations have been performed in the languages GAUSS and XploRe [Härdle, Klinke, and Turlach (1995)]. The observations were generated according to a binary response model. The explanatory variables were identical independent uniformly distributed on $[-1, 1]$. We took the parameter $\theta = \binom{1}{1} 1/\sqrt{2}$ and considered the functions

$$f_0(u) = \frac{1}{1 + \exp^{-u}} \tag{49}$$

$$f_1(u) = f_0(u) + \eta \varphi'(u) \tag{50}$$

$$f_2(u) = 1 - \exp(-\exp(u)) \tag{51}$$

for different $0 < \eta \le 1$, where φ is the density function of the standard normal distribution. While f_0 is a logit function, f_1 consists of a logit disturbed by a bump (Figure 3). The response Y under H_0 was generated such that $P(Y = 1 | x^T \theta_0 = u) = f_0(u)$. We are thus interested in the hypothesis H_0,

$$H_0\colon F_\theta(x) = \mathbf{E}[Y | u(x, \theta) = u] = f_0(u), \qquad \theta \in S_2.$$

In a first step we calculated empirically the 90 and 95% quantiles of T_n^* for $n = 100$ and 200 observations generated by f_0. They were used then as rejection boundaries, defined as $\sqrt{4 \log N}$; see (24). We calculated T_n^* by optimizing T_β over a grid; see (23), with $N = 50$ gridpoints. As kernel function K we used always the quartic kernel,

$$K(u) = \tfrac{15}{16}(1 - u^2)^2 \mathbf{1}_{\{|u| < 1\}}.$$

In the second step we analyzed the effect of increasing sample size on the power. In Table 1 we show the power of the test when the data were generated with functions f_{1a}, that is, f_1 for $\eta = 0.2$, f_{1c}, where $\eta = 0.6$ and f_2. In order not to oversmooth, we used the bandwidth $h_1 = h = 0.5$ for $n = 100$, 200 and $h_1 = h = 0.25$ for $n = 350$, 500. Although we substituted, for speed reasons in the cases $n = 350$ and 500, $\tilde{V}_\beta$ by $\tilde{V}_\theta$ for all β, the power increases very fast with n. Therefore, it could be of interest to compare the power with regard to the bump η in the logit model. In Table 2 we show for $n = 200$ and 350 the power of the test as a function of η. We see that for $\eta > 0.4$ this test procedure works very well.

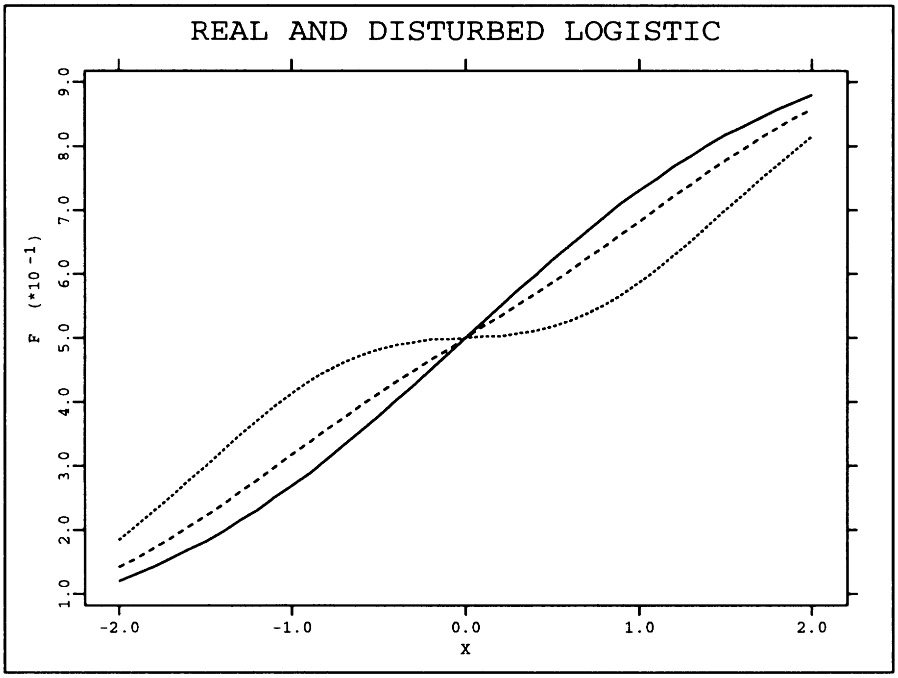

FIG. 3. *Solid line*: f_0; *dashed line*: f_1 *with* $\eta = 0.2$; *pointed line*: f_1 *with* $\eta = 0.6$.

TABLE 1
Power and rejection boundaries for different alternatives

$\boldsymbol{n, h} =$	**100, 0.5**		**200, 0.5**		**350, 0.25**		**500, 0.25**	
level	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
rejection boundary	**4.00**	**3.35**	**3.30**	**3.25**	**3.75**	**2.90**	**3.20**	**2.76**
f_{1a}	0.056	0.096	0.112	0.215	0.133	0.207	0.150	0.200
f_{1c}	0.224	0.294	0.530	0.690	0.798	0.856	0.900	0.960
f_2	0.316	0.376	0.946	0.991	0.995	1.000	0.995	1.000

TABLE 2
Power for different bumps η

$\boldsymbol{\eta} =$	**0.2**		**0.4**		**0.6**		**1.0**	
level	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
n, h 200, 0.50	0.112	0.215	0.227	0.419	0.530	0.690	0.687	0.801
350, 0.25	0.133	0.207	0.321	0.478	0.798	0.856	0.889	0.926

The last step of the simulation experiment was the study of bandwidth choice. For the sake of simplicity, we set $h_1 = h$ as above. First, we always have had to determine numerically the rejection boundaries for the special bandwidth h. Here we observed shrinking boundaries, when h grew from 0.25 u to 2.25. In Figure 4 we plot the bandwidth versus the power of the test with observations generated by f_{1c}. Obviously for this kind of alternative we get better power for larger bandwidths.

In the introductory example we dealt with youth unemployment. The question is, can we explain the youth unemployment with the aforementioned predictor variables X in a single index model with logit link? In the application to this dataset, we used a slightly modified numerical procedure as described in Proenca and Ritter (1995). Further, we rescaled the explanatory variables of each dimension to $[-1, 1]$. Since there are three dimensions ($d = 3$) for a sample size of $n = 462$, we choose the bandwidth h_1 large, definitively 1.5, whereas $h = 0.3$. By Monte Carlo studies described above we determined the 90 and 95% one-sided quantiles of T^*_{462} and got 1.74, respectively 2.38. Then we ran the test for our data and got the statistic value $T^*_{462} = 3.076$ for $\beta = (-0.18010, -0.10725, 0.97778)$. For the purpose of comparison, in Table 3 we switch the norm of β and set this first component equal to the corresponding one of θ, the parameter of the logit fit in Figure 1.

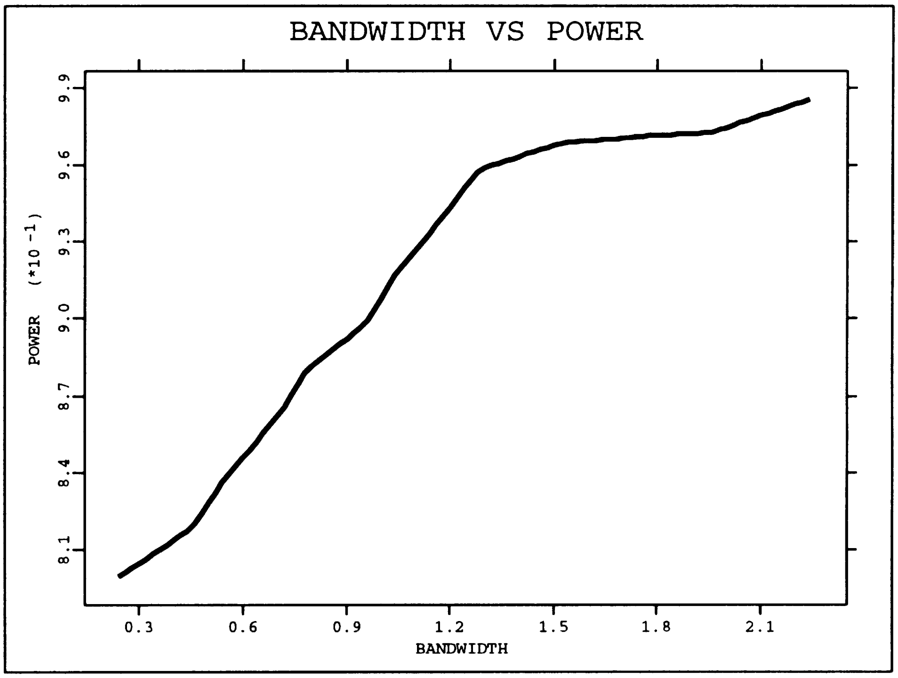

FIG. 4. *Power function of the test with respect to the bandwidth for function* f_{1c}.

TABLE 3
Comparison of θ and β

Explanatory variables	**Intercept**	**Earnings as an apprentice**	**Percentage of apprentices divided by employees**	**Unemployed rate**
θ	-2.40996	-0.07999	-0.17989	0.95113
β	—	-0.07999	-0.04763	0.43422

APPENDIX

PROOF OF THEOREM 2.1. To simplify our exposition and to emphasize the main idea, we consider the case when the parametric family consists of one point, namely, a zero regression function, and errors ε_i are independent and standard Gaussian. Moreover, we assume random design with a design density $\pi(x)$ in $\mathbb{R}^d$ of the form $\pi(x) = \pi_1(|x|)$, where a univariate function $\pi_1(\cdot)$ is compactly supported on $[-1, 1]$, symmetric, twice continuously differentiable and satisfies $\pi_1(t) = 3/4$ for $|t| \le 1/2$.

The method of the proof is standard; see, for example, Ingster (1993). We replace the minimax problem by a Bayes one, where we consider, instead of the set $\mathscr{F}_1$ of alternatives, one Bayes alternative corresponding to a prior ν concentrated on $\mathscr{F}_1$. We try to choose this prior ν in such a way that the likelihood $Z_\nu = d\mathbf{P}_\nu/d\mathbf{P}_0$ is close to 1 where the measure $\mathbf{P}_\nu$ is the Bayes measure for the prior ν and $\mathbf{P}_0$ corresponds to the case of zero regression function. The Neyman–Pearson lemma yields that the hypothesis H_0: $\mathbf{P} = \mathbf{P}_0$ cannot be consistently distinguished from the Bayes alternative H_ν: $\mathbf{P} = P_\nu$ and hence from the composite alternative H_1: $\mathbf{P} \in \mathscr{F}_1$. In the proof we use a hypercube argument as in Bickel and Ritov (1988).

Now we describe the structure of the prior ν. Let $g(\cdot)$ be some function from the Hölder class $\Sigma(s, L)$, supported on $[-1, 1]$ and satisfying the conditions

$$\int g(t)\,dt = 0, \qquad \|g\|^2 = \int g^2(t)\,dt > 0, \qquad \|g\|_\infty = \sup_t |g(t)| \le 1. \tag{52}$$

Set

$$h = \left(\frac{a\sqrt{\log n}}{n} \right)^{2/(4s+1)}, \tag{53}$$

where a constant a will be chosen later. Denote by $\mathscr{I}_n$ the partition of the interval $[-\frac{1}{2}, \frac{1}{2}]$ into intervals of length h. Without loss of generality, we assume that the cardinality of the set $\mathscr{I}_n$ coincides with $1/h$,

$$\#\mathscr{I}_n = \frac{1}{h}. \tag{54}$$

For each interval $i \in \mathcal{I}_n$ introduce a function $g_I(t)$ of the form

$$g_I(t) = h^s g\left(\frac{t - t_I}{h}\right), \tag{55}$$

t_I being the center of I. Evidently $g_I(\cdot)$ is supported on I, $g_I \in \Sigma(s, L)$ and the following hold for h small enough:

$$\int g_I(t)\, dt = 0, \qquad \int g_I^2(t)\, dt = h^{2s+1}\|g\|^2. \tag{56}$$

Let now μ be a set of binary values $\{\mu_I,\ I \in \mathcal{I}_n\}$, that is, $\mu_I = \pm 1$. Define a function $G_\mu(t)$ with

$$G_\mu(t) = \sum_{I \in \mathcal{I}_n} \mu_I g_I(t). \tag{57}$$

This function $G_\mu \in \Sigma(s, L)$ vanishes outside $[-\frac{1}{2}, \frac{1}{2}]$ and by (56),

$$\int G_I^2(t)\, dt = \sum_{i \in \mathcal{I}_n} \int g_I^2(t)\, dt = \frac{1}{h} h^{2s+1} \int g^2(t)\, dt = h^{2s}\|g\|^2. \tag{58}$$

Taking into account (53) we see that the distance between zero function and each G_μ is just of the rate c_n^2 from Theorems 2.1 and 2.2.

Denote by $\mathcal{M}_n$ the set of all possible collections $\{\mu_I,\ I \in \mathcal{I}_n\}$ with binaries $\mu_I = \pm 1$, and let $m(d\mu)$ be the uniform measure on $\mathcal{M}_n$. This measure can be represented as the direct product of binary measures $m_I(d\mu_I)$ with $m_I(\mu_I \pm 1) = 1/2$.

Now we pass to the semiparametric model. Let S_n be a grid on the unit sphere S_d with the step b_n,

$$b_n = h^{1/6}, \tag{59}$$

h being from (53). This means that $|\beta - \beta'| \ge b_n = h^{1/6}$ for each $\beta, \beta' \in S_n$, $\beta \neq \beta'$. Below we will use that for some $\alpha > 0$,

$$N = \# S_n \asymp n^\alpha \tag{60}$$

and

$$\frac{h^{1/2}\log n}{|\beta - \beta'|^2} \le \frac{h^{1/2}\log n}{b_n^2} \to 0, \qquad n \to \infty\ \forall\, \beta, \beta' \in S_n,\ \beta \neq \beta'. \tag{61}$$

For each $\beta \in S_n$ and each $\mu \in \mathcal{M}_n$, define a multivariate function $G_{\beta,\mu}(x)$ on $\mathbb{R}^d$ with

$$G_{\beta,\mu}(x) = G_\mu(x^\top \beta).$$

It is clear that the function $G_{\beta,\mu}(x)$ is Hölder, $G_{\beta,\mu}(x) \in \Sigma_d(s, L)$, and by (58) we get

$$\begin{aligned} \int G_{\beta,\mu}^2(x)\pi(x)\, dx &= \int G_\mu^2(x^\top \beta)\pi_1(|x|)\, dx \\ &= \int G_\mu^2(t)\pi_2(t)\, dt = C_0 h^{2s} \end{aligned} \tag{62}$$

with $\pi_2(t) = (d/dt)\mathbf{1}(x^\top \beta \le t)\pi_1(|x|)\,dx$ and $C_0 \in [\frac{1}{2}\|g\|^2, \|g\|^2]$. Finally we take the prior ν as the uniform measure on the set of functions $\{G_{\beta,\mu},\ \beta \in S_n,\ \mu \in \mathscr{M}_n\}$, and

$$\mathbf{P}_\nu = \frac{1}{N}\sum_{\beta\in S_n}\frac{1}{M}\sum_{\mu\in\mathscr{M}_n}\mathbf{P}_{G_{\beta,\mu}}. \tag{63}$$

Here $M = \#\mathscr{M}_n = 2^{1/h}$, N being from (60). Denote also $Z_\nu = d\mathbf{P}_\nu/d\mathbf{P}_0$ and notice that this likelihood can be represented in the form $Z_\nu = (1/N)\sum_{\beta\in S_n} Z_\beta$ with

$$Z_\beta = \frac{1}{M}\sum_{\mu\in\mathscr{M}_n} Z_{\beta,\mu} = \frac{1}{M}\sum_{\mu\in\mathscr{M}_n}\frac{d\mathbf{P}_{G_{\beta,\mu}}}{d\mathbf{P}_0}. \tag{64}$$

Our goal is to prove that for a small enough in (53) one has

$$Z_\nu \to 1 \tag{65}$$

under the measure $\mathbf{P}_0$.

We start from a decomposition and an asymptotic expansion for each Z_β from (64). For that we need some more notation. Fix some $\beta \in S_n$ and put

$$\sigma^2_{\beta,I} = \sum_i g_I^2(X_i^\top \beta), \qquad I \in \mathscr{I}_n, \tag{66}$$

$$\xi_{\beta,I} = \frac{1}{\sigma_{\beta,I}}\sum_i g_I(X_i^\top \beta)\varepsilon_i, \qquad I \in \mathscr{I}_n. \tag{67}$$

We see that $\xi_{\beta,I}$ are standard normal and independent for different $I \in \mathscr{I}_n$, and

$$\sum_i G^2_{\beta,\mu}(X_i) = \sum_i G^2_\mu(X_i^\top \beta) = \sum_{I\in\mathscr{I}_n}\sigma^2_{\beta,I}.$$

Recall that we assume random design and

$$\mathbf{E}\sum_i G^2_{\beta,\mu}(X_i) = n\int G^2_{\beta,\mu}(x)\pi(x)\,dx = nC_0h^{2s}. \tag{68}$$

Similarly for each $\sigma^2_{\beta,I}$,

$$\mathbf{E}\sigma^2_{\beta,I} = n\int g_I^2(x^\top \beta)\pi(x)\,dx = n\int g_I^2(x^\top \beta)\pi_1(|x|)\,dx = nC_Ih^{2s+1}, \tag{69}$$

where C_I does not depend on β and $C_I \in [C_0/\sqrt{2}, \sqrt{2}\,C_0]$.

LEMMA A.1.

$$Z_\beta = \prod_{i\in\mathscr{I}_n} ch(\sigma_{\beta,I}\xi_{\beta,I})\exp\left(-\tfrac{1}{2}\sigma^2_{\beta,I}\right),$$

where $ch(z) = \frac{1}{2}(e^z + e^{-z})$.

PROOF. By Girsanov's formulas and (66) and (67),

$$\begin{aligned} Z_{\beta,\mu} &= \exp\Big\{ \sum_i G_{\beta,\mu}(X_i)\,\varepsilon_i - \tfrac{1}{2} G^2_{\beta,\mu}(X_i) \Big\} \\ &= \exp\Big\{ \sum_{I\in\mathscr{I}_n} \mu_I \sigma_{\beta,I} \xi_{\beta,I} - \tfrac{1}{2} \sum_{I\in\mathscr{I}_n} \sigma^2_{\beta,I} \Big\} \\ &= \prod_{I\in\mathscr{I}_n} \exp\big\{ \mu_I \sigma_{\beta,I} \xi_{\beta,I} - \tfrac{1}{2}\sigma^2_{\beta,I} \big\}. \end{aligned}$$

Now the lemma's assertion follows from the direct product structure of the measure $m(d\mu)$. $\square$

Denote also

$$v_\beta^2 = \tfrac{1}{2} \sum_{I\in\mathscr{I}_n} \sigma^4_{\beta,I}, \tag{70}$$

$$\zeta_\beta = \frac{1}{v_\beta} \sum_{I\in\mathscr{I}_n} \sigma^2_{\beta,I} \big(\xi^2_{\beta,I} - 1 \big). \tag{71}$$

LEMMA A.2. *The following statements hold*:

(i) $\mathbf{E}_0 \zeta_\beta = 0$.
(ii) $\mathbf{E}_0 \zeta_\beta^2 = 1$.
(iii) $\mathbf{E} v_\beta^2 \le C_1 n^2 h^{4s+1} = C_1 \log n$ *with* $C_1 \le a$; *see* (53).
Var $v_\beta^2 \le C_2 n^{-1} \log n$ *with some positive* C_2.
(iv) *The random variables* ζ_β *are asymptotically normal and there exist standard normal random variables* $\tilde{\zeta}_\beta$ *such that*

$$\log n \sup_{\beta\in S_n} \mathbf{E}_0 \big(\tilde{\zeta}_\beta - \zeta_\beta \big)^2 \to 0,$$

$$\log n \sup_{\beta,\beta'\in S_n} |\mathbf{E}_0 \tilde{\zeta}_\beta \tilde{\zeta}_{\beta'} - \mathbf{E}_0 \zeta_\beta \zeta_{\beta'}| \to 0.$$

For the proof, the first two statements are obvious. Statement (iii) follows from (69). Finally, (iv) is the application of the Strassen type invariance principle [see, e.g., Csörgő and Révész (1981)].

The next step is the asymptotic expansion for each Z_β.

LEMMA A.3. *The following statements are satisfied uniformly in* $\beta \in S_n$, *for each* $\delta > 0$:

(i) $$\mathbf{P}_0\Big(\Big| Z_\beta - \exp\big\{ v_\beta \zeta_\beta - \tfrac{1}{2} v_\beta^2 \big\} \Big| > \delta \Big) \to 0,$$

(ii) $$\mathbf{P}_0\Big(\Big| Z_\beta - \exp\big\{ v_\beta \tilde{\zeta}_\beta - \tfrac{1}{2} v_\beta^2 \big\} \Big| > \delta \Big) \to 0.$$

PROOF. The first statement is equivalent to the following one:

$$\mathbf{P}_0\Big(\big|\log Z_\beta - v_\beta \zeta_\beta + \tfrac{1}{2} v_\beta^2\big| > \delta\Big) \to 0,$$

but the latter can be obtained using the Taylor expansion for $\log Z_\beta$:

$$\begin{aligned}\log Z_\beta &= \sum_{I \in \mathscr{I}_n} \log ch(\sigma_{\beta, I}\xi_{\beta, I}) - \tfrac{1}{2}\sigma_{\beta, I}^2 \\ &= \sum_{I \in \mathscr{I}_n} \Big[\tfrac{1}{2}\sigma_{\beta, I}^2(\xi_{\beta, I}^2 - 1) - \tfrac{1}{12}\sigma_{\beta, I}^4 \xi_{\beta, I}^4 + O(\sigma_{\beta, I}^6 \xi_{\beta, I}^6)\Big]\end{aligned}$$

and the following asymptotic relations which hold uniformly in β:

$$\mathbf{P}_0\Bigg(\bigg|\sum_{I \in \mathscr{I}_n} \sigma_{\beta, I}^4(\xi_{\beta, I}^4 - 3)\bigg| > \delta\Bigg) \to 0,$$

$$\mathbf{P}_0\Bigg(\bigg|\sum_{I \in \mathscr{I}_n} \sigma_{\beta, I}^6 \xi_{\beta, I}^6\bigg| > \delta\Bigg) \to 0;$$

for details we refer to Ingster (1993).

The second statement of the lemma follows directly from Lemma A.2(iii). □

Now we arrive at the central point of the proof. Actually we prove that "submodels" corresponding to different β are in some sense asymptotically independent. That is why we have to pay with the extra log term for the choice of "direction" β.

LEMMA A.4. *There exists a constant C such that for any $\beta, \beta' \in S_n$, $\beta \neq \beta'$,*

$$|\mathbf{E}\zeta_\beta \zeta_{\beta'}| \le \frac{Ch^{1/2}}{|\beta - \beta'|^2}. \tag{72}$$

PROOF. Let us fix some β, β' for S_n. Denote by ρ their scalar product,

$$\rho = (\beta, \beta').$$

We will also denote by $\mathbf{E}^\pi$ the condition expectation given the design points X_i, $i = 1, \dots, n$.

Set for I, I' from $\mathscr{I}_n$,

$$r = r_{I, I'} = \mathbf{E}^\pi \xi_{\beta, I}\xi_{\beta', I'} = \frac{1}{\sigma_{\beta, I}\sigma_{\beta', I'}} \sum_i g_I(X_i^\top \beta) g_{I'}(X_i^\top \beta'). \tag{73}$$

Obviously, $|r_{I, I'}| \le 1$.

Using the normality of $\xi_{\beta, I}$ and $\xi_{\beta', I'}$ we calculate easily

$$\mathbf{E}^\pi(\xi_{\beta, I}^2 - 1)(\xi_{\xi', I'}^2 - 1) = 4r_{I, I'}^2 - 2r_{I, I'}. \tag{74}$$

One has

$$\begin{aligned}\mathbf{E}_0\,\zeta_\beta\zeta_{\beta'} &= \mathbf{E}_0\frac{1}{v_\beta}\sum_{I\in\mathscr{I}_n}\sigma_{\beta,I}^2\big(\xi b_{\beta,I}^2-1\big)\frac{1}{v_{\beta'}}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta',I'}^2\big(\xi_{\beta',I'}^2-1\big)\\ &= \mathbf{E}_0\frac{1}{v_\beta}\frac{1}{v_{\beta'}}\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta,I}^2\sigma_{\beta',I'}^2\big[4r_{I,I'}^2-2r_{I,I'}\big]\end{aligned}$$

and using $|r_{I,I'}|\le 1$ and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}|\mathbf{E}_0\,\zeta_\beta\zeta_{\beta'}| &\le 6\mathbf{E}_0\frac{1}{v_\beta}\frac{1}{v_{\beta'}}\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta,I}^2\sigma_{\beta',I'}^2|r_{I,I'}|\\ &\le 6\left[\mathbf{E}_0\frac{1}{v_\beta^2}\frac{1}{v_{\beta'}^2}\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta,I}^2\sigma_{\beta',I'}^2\right]^{1/2}\left[\mathbf{E}_0\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta,I}^2\sigma_{\beta',I'}^2r_{I,I'}^2\right]^{1/2}.\end{aligned}$$

Next, using the Hölder inequality,

$$\begin{aligned}\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\sigma_{\beta,I}^2\sigma_{\beta',I'}^2 &= \Big(N\sum_{I\in\mathscr{I}_n}\sigma_{\beta,I}^4\Big)^{1/2}\Big(N\sum_{I'\in\mathscr{I}_n}\sigma_{\beta',I'}^4\Big)^{1/2}\\ &= Nv_\beta v_{\beta'}\Big(\sum_{I\in\mathscr{I}_n}\sigma_{\beta,I}^2\Big)\Big(\sum_{I'\in\mathscr{I}_n}\sigma_{\beta',I'}^2\Big)\le Nv_\beta v_{\beta'},\end{aligned}$$

where, recall, $N=\#\mathscr{I}_n=h^{-1}$. This inequality and Lemma A.2(iii) imply

$$\begin{aligned}|\mathbf{E}_0\,\zeta_\beta\zeta_{\beta'}| &\le C(\log n)^{-1}N\left[\sum_{I\in\mathscr{I}_n}\sum_{I'\in\mathscr{I}_n}\mathbf{E}\,\sigma_{\beta,I}^2\sigma_{\beta',I'}^2r_{I,I'}^2\right]^{1/2}\\ &\le C(\log n)^{-1}N^2\sup_{I,I'\in\mathscr{I}_n}\big[\mathbf{E}\,\sigma_{\beta,I}^2\sigma_{\beta',I'}^2r_{I,I'}^2\big]^{1/2}\end{aligned}\tag{75}$$

with some $C>0$.

Let us fix for a moment some $I, I'\in\mathscr{I}_n$ and denote

$$w_{I,I'}^2=\mathbf{E}\,\sigma_{\beta,I}^2\sigma_{\beta',I'}^2r_{I,I'}^2.$$

Below we will prove the following estimate:

$$w_{I,I'}\le\frac{Cnh^{2s+3}}{1-\rho^2}\tag{76}$$

with some constant C depending only on the design density π. Substituting the estimates in (75) and using (53), we get that $|\mathbf{E}_0\,\zeta_\beta\zeta_{\beta'}|\le Ch^{1/2}(1-\rho^2)^{-1}$. Now the assertion follows from the simple fact that $(\beta-\beta')^2\ge 1-\rho^2$.

It remains to check (76). First we note that from (73),

$$
\begin{aligned}
w_{I,I'}^2 &= \mathbf{E}\Big[\sum_i g_I(X_i^\top \beta) g_{I'}(X_i^\top \beta')\Big]^2 \\
&= (n^2-n)\Big[\int g_1(x^\top \beta) g_{I'}(x^\top \beta')\pi(x)\,dx\Big]^2 \\
&\quad + n\int g_I^2(x^\top \beta) g_{I'}^2(x^\top \beta')\pi(x)\,dx.
\end{aligned}
$$

To simplify the notation, suppose without loss of generality that $\beta = (1,0,\dots,0)$, $\beta' = (\rho, \sqrt{1-\rho^2}, 0, \dots, 0)$. Introduce new variables $y_1, y_2, \dots, y_d$ by the equality $x^\top \beta = t_I + hy_1$, $x^\top \beta' = t_{I'} + hy_2$, $y_k = x_k$, $k = 3,\dots,d$ and denote by T the linear transformation in R^d such that $y = Tx$. It is easy to verify that $\partial T/\partial y_1 = (h, 0, \dots, 0)$, $\partial T/\partial y_2 = (h\varrho, h(1-\varrho^2)^{-1/2}, \dots, 0)$ and $|\det T| = h^2(1-\varrho^2)^{-1/2}$. We have

$$
\begin{aligned}
w_{I,I'}^2 &= (n^2-n)\Bigg[\frac{h^{2s+2}}{\sqrt{1-\rho^2}}\int g(y_1)g(y_2)\pi\circ T(y)\,dy\Bigg]^2 \\
&\quad + \frac{nh^{4s+2}}{\sqrt{1-\rho^2}}\int g^2(y_1)g^2(y_2)\pi\circ T(y)\,dy.
\end{aligned}
$$

Here $\pi\circ T$ means the function on R^d defined as the superposition of the linear transformation T and the design density π. Since the function g has the support in $[-1,1]$, the latter integrals can be considered only on the domain with $|y_1| \le 1$ and $|y_2| \le 1$. For any such point, one has $|\pi\circ T(y_1, y_2, y_3, \dots, y_d) - \pi\circ T(0,0,y_3,\dots,y_d)| \le \|\pi'\|(|\partial T/\partial y_1| + |\partial T/\partial y_2|)$ $\|\pi'\|(2h + h(1-\varrho^2)^{-1/2})$ where $\|\pi'\|$ means the maximum of the norm of the first derivative of π. This implies

$$
\begin{aligned}
\Big|\int g(y_1)g(y_2)\pi\circ T(y)\,dy\Big| &= \Big|\int g(y_1)g(y_2)\pi\circ T(y_1,y_2,y_3,\dots,y_d)\,dy \\
&\qquad - \int g(y_1)g(y_2)\pi\circ T(0,0,y_3,\dots,y_d)\,dy\Big| \\
&\le Ch(1-\varrho^2)^{-1/2}.
\end{aligned}
$$

Here we have used the equality $\int g(t)\,dt = 0$ and boundedness of g. Similarly, one estimates the second integral $\int g^2(y_1)g^2(y_2)\pi\circ T(y)\,dy$ and (76) follows.

Now everything is prepared to complete the proof of (65). The results of Lemmas A.2 and A.3, reduce this assertion to the following one:

$$
\frac{1}{N}\sum_{\beta\in S_n}\Big[\exp\Big\{v_\beta\tilde{\zeta}_\beta - \frac{1}{2}v_\beta^2\Big\} - 1\Big] \to 0 \tag{77}
$$

under the measure $\mathbf{P}_0$. It suffices to check that

$$\frac{1}{N^2}\mathbf{E}_0\Big|\sum_{\beta\in S_n}\Big(\tilde{Z}_\beta - 1\Big)\Big|^2 \to 0$$

with

$$\tilde{Z}_\beta = \exp\Big\{v_\beta\tilde{\zeta}_\beta - \tfrac{1}{2}v_\beta^2\Big\}.$$

Using the normality of $\tilde{\zeta}_\beta$ and Lemma A.2(iii), one derives

$$\mathbf{E}_0\Big[\exp\Big\{v_\beta\tilde{\zeta}_\beta - \tfrac{1}{2}v_\beta^2\Big\} - 1\Big]^2 = \exp\Big\{v_\beta^2\Big\} \le n^a.$$

For different, $\beta, \beta' \in S_n$, denote $\alpha = \mathbf{E}_0\,\tilde{\zeta}_\beta\tilde{\zeta}_{\beta'}$. Then by direct calculation,

$$\mathbf{E}_0\tilde{Z}_\beta\tilde{Z}_{\beta'} = \mathbf{E}_0\exp\{\alpha v_\beta v_{\beta'}\}.$$

The results of Lemmas A.4 and A.2(iv) allow us to obtain

$$\mathbf{E}_0\Big(\tilde{Z}_\beta b - 1\Big)\Big(\tilde{Z}_{\beta'} - 1\Big) = \mathbf{E}_0\tilde{Z}_\beta\tilde{Z}_\beta - 1 \le C\alpha\log n.$$

Finally, we derive by (61), Lemmas A.4 and A.2(iii), (iv),

$$\begin{aligned}
&\frac{1}{N^2}\mathbf{E}_0\Big|\sum_{\beta\in S_n}\Big(\tilde{Z}_\beta - 1\Big)\Big|^2\\
&\quad= \frac{1}{N^2}\sum_{\beta\in S_n}\mathbf{E}_0\Big(\tilde{Z}_\beta - 1\Big)^2 + \frac{1}{N^2}\sum_{\beta\in S_n}\sum_{\substack{\beta'\in S_n\\ \beta'\neq\beta}}\mathbf{E}_0\Big(\tilde{Z}_\beta - 1\Big)\Big(\tilde{Z}_{\beta'} - 1\Big)\\
&\quad\le \frac{1}{N^2}\sum_{\beta\in S_n}\exp\Big(v_\beta^2\Big) + \frac{1}{N^2}\sum_{\beta\in S_n}\sum_{\substack{\beta'\in S_n\\ \beta'\neq\beta}}\frac{Ch^{1/2}\log n}{|\beta-\beta'|^2}\\
&\quad\le \frac{1}{N^2}n^a + \frac{Ch^{1/2}\log n}{b_n^2} \to 0,
\end{aligned}$$

if a in (53) is taken small enough.

Acknowledgment. We thank O. Lepski for helpful discussion and comments.

REFERENCES

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

CSÖRGŐ, M. and RÉVÉSZ, P. (1982). *Strong Approximation in Probability and Statistics*. Akadémiai Kiadó, Budapest.

ENGLE, R. F., GRANGER, W. J., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FRIEDMAN, F. H. and STUETZLE, W. (1984). A projection pursuit regression. *J. Amer. Statist. Assoc.* **79** 599–608.

GOLUBEV, G. (1992). Asymptotic minimax regression estimation in additive model. *Problems Inform. Transmission* **28** 3–15. (In Russian.)

GREEN, P. and SILVERMAN, B. W. (1994). *The Penalized Likelihood Approach*. Chapman and Hall, London.

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

HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single index models. *Ann. Statist.* **21** 157–178.

HÄRDLE, W., KLINKE, S. and TURLACH, T. A. (1995). *XploRe: An Interactive Statistical Computing Environment*. Springer, Berlin.

HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947.

HALL, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–578.

HOROWITZ, J. (1993). Semiparametric and nonparametric estimation of quantal response models. In *Handbook of Statistics* (G. S. Maddala, C. R. Rao and H. D. Vinod, eds.) 45–72. Elsevier, New York.

HOROWITZ, J. and HÄRDLE, W. (1994). Testing a parametric model against a semiparametric alternative. *Econometric Theory* **10** 821–848.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

HUET, S., JOLIVET, E. and MÉSSEAU, A. (1993). *La regression Non-lineaire: Methodes et Applications en Biologie*. INRA, Paris.

IBRAGIMOV, I. A. and KHASMINSKI, R. Z. (1977). One problem of statistical estimation in Gaussian white noise. *Sov. Math. Dokl.* **236** 1351–1354.

IBRAGIMOV, I. A. and KHASMINSKI, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, Berlin.

INGSTER, YU. I. (1982). Minimax nonparametric detection of signals in white Gaussian noise. *Problems Inform. Transmission* **18** 130–140.

INGSTER, YU. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Math. Methods Statist.* **2** 85–114.

MADDALA, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge Univ. Press.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models* **2**. Chapman and Hall, London.

MÜLLER, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–238.

PROENCA, I. and RITTER, CHR. (1995). Negative bias in the H-H Statistik. *Comput. Statist.*

RICE, J. A. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* **4** 203–208.

SEVERINI, T. A. and STANISWALIS, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89** 501–511.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–446.

SPOKOINY, V. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498.

W. HÄRDLE
S. SPERLICH
INSTITUT FÜR STATITIK UND OKONOMETRIE
HUMBOLDT UNIVERSITÄT ZU BERLIN
SFB 373, SPANDAUER STR. 1
10178 BERLIN
GERMANY
E-MAIL: stat@wiwi.hu-berlin.de

V. SPOKOINY
WEIERSTRASS INSTITUTE FOR
APPLIED ANALYSIS AND STOCHASTICS
MOHRENSTR. 39
10117 BERLIN
GERMANY