The Annals of Statistics
2001, Vol. 29, No. 3, 595–623

DIRECT ESTIMATION OF THE INDEX COEFFICIENT IN A SINGLE-INDEX MODEL

BY MARIAN HRISTACHE, ANATOLI JUDITSKY AND VLADIMIR SPOKOINY

ENSAI and CREST, Université Joseph Fourier and Weierstrass Institute

Single-index modeling is widely applied in, for example, econometric studies as a compromise between too restrictive parametric models and flexible but hardly estimable purely nonparametric models. By such modeling the statistical analysis usually focuses on estimating the index coefficients. The average derivative estimator (ADE) of the index vector is based on the fact that the average gradient of a single index function $f(x^\top \beta)$ is proportional to the index vector β. Unfortunately, a straightforward application of this idea meets the so-called "curse of dimensionality" problem if the dimensionality d of the model is larger than 2. However, prior information about the vector β can be used for improving the quality of gradient estimation by extending the weighting kernel in a direction of small directional derivative. The method proposed in this paper consists of such iterative improvements of the original ADE. The whole procedure requires at most $2\log n$ iterations and the resulting estimator is $\sqrt{n}$-consistent under relatively mild assumptions on the model independently of the dimensionality d.

1. Introduction. Suppose that the observations $(Y_i, X_i), i = 1, \dots, n$, are generated by the regression model

$$Y_i = f(X_i) + \varepsilon_i, \tag{1.1}$$

where Y_i are scalar response variables, $X_i \in [0,1]^d$ are d-dimensional explanatory variables, ε_i are random errors and $f(\cdot)$ is an unknown d-dimensional function f: $\mathbb{R}^d \to \mathbb{R}$. We assume that $f(x)$ has the following structure:

$$f(x) = g_0(x^\top \theta^*). \tag{1.2}$$

Here $g_0(\cdot)$ is an unknown univariate *link* function, that is, $g_0(\cdot)$: $\mathbb{R} \to \mathbb{R}$ and θ^* is an unknown *index* vector. In the statistical literature the relations as in (1.1) and (1.2) are referred to as *single-index regression* models. These models are often used in econometrics as a reasonable compromise between fully parametric and fully nonparametric modeling. See, for example, McCullagh and Nelder (1989). They are also extensively used in projection pursuit regression; see Friedman and Stuetzle (1981) and Hall (1989).

Two estimation problems in single-index models are intensively discussed in the literature. The first consists of estimation of the unknown function $f(x)$. The objective of the second is to recover the index-vector θ^*. In this paper we

Received June 1998; revised September 2000.
AMS 2000 *subject classifications*. Primary 62G05; Secondary 62H40, 62G20.
Key words and phrases. Single-index model, index coefficients, direct estimation, iteration.

focus on the second one. Note first that the vector θ^* in the representation (1.2) is not uniquely defined. Indeed, the use of the vector $c\theta^*$ and of the rescaled link function $g_c(u) = g_0(u/c)$ with some $c > 0$ leads to the same regression function f. To ensure the uniqueness of the vector θ^*, one should impose some identifiability condition on θ^*. Usually it is supposed that the Euclidean norm of θ^* is equal to 1; that is, θ^* is a unit vector in $\mathbb{R}^d$.

Several methods for estimating θ^* has been developed in the theory of semiparametric estimation. For instance, in the M-estimation approach the unknown link function g is considered as an infinite-dimensional nuisance parameter. Then the estimator $\hat{\theta}$ of θ^* is constructed by minimization of an M-functional with respect to θ, when replacing g by its nonparametric estimator,

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \psi\Big(Y_i, \hat{g}_{\theta,h}(X_i^\top \theta)\Big),$$

where $\hat{g}_{\theta,h}(\cdot)$ is a nonparametric estimator (with smoothing parameter h) of $g_\theta(\cdot) = \mathbf{E}(Y_i | X_i^\top \theta = \cdot)$ and ψ is a contrast function. Typical examples are the semiparametric maximum likelihood estimator (SMLE), with $-\psi$ being the log-likelihood of the errors ε_i, and the semiparametric least squares estimator (SLSE), with $\psi(y, r) = |y - r|^2$ being the Euclidean norm of $y - r$ squared. Klein and Spady (1993) have shown that the SMLE is asymptotically efficient in the so-called binary response model. Ichimura (1993) studied the properties of SLSE in a general single-index model. The problem of the choice of bandwidth for the nonparametric estimation of the link function has been considered in Härdle, Hall and Ichimura (1993) and Delecroix, Hristache and Patilea (1999). Delecroix and Hristache (1999) studied a rather general type of M-estimator, and the asymptotic efficiency of the general semiparametric maximum-likelihood estimator has been proved in Bonneu, Delecroix and Hristache (1997) and Delecroix, Härdle and Hristache (1997) for particular classes of single-index models.

In spite of their nice theoretical properties, M-estimators are rarely implemented in practice. The main reason for this is that the computation of these estimators leads to a hard optimization problem in a high-dimensional space.

As an alternative to M-estimators, the so-called average derivative method (ADE) has been introduced in Stoker (1986) and Powell, Stock and Stoker (1989). The idea of this method is to estimate the expected value of the (weighted) gradient $(\partial/\partial x) f(x) = \theta^* g'(x^\top \theta^*)$ of the regression function f, which is obviously proportional to θ^*. This method leads to a $\sqrt{n}$-consistent estimator of the index vector; see also Härdle and Tsybakov (1993). An advantage of this approach is that it allows estimation of the vector θ^* "directly" and does not require solving a hard optimization problem. A generalization of the average derivative estimate for the case where the components of X_i are continuous and/or discrete has been provided in Horowitz and Härdle (1996). However, the conditions required for the average derivative estimator to work are rather restrictive. In particular, the rate $n^{-1/2}$ can be attained only for the random design with a very smooth design density and only if a high-order kernel is used for its pilot estimation, as in Härdle and Tsybakov (1993) or Samarov (1993). In practical applications of the estimator for data samples of a reasonable size, a kernel estimator of the design density using a high-order kernel has very poor performance because of data sparseness (the so-called "curse of dimensionality" problem).

Another direct method of index coefficient estimation, called "sliced inverse regression" has been proposed in Li and Duan (1989), Duan and Li (1991) and Li (1991). The inverse regression method, however, requires the distribution of covariates to be elliptically symmetric.

In the present paper we introduce a new type of direct estimator of the index coefficient θ^*. It can be regarded as an iterative improvement of the average derivative estimator. We show that the proposed estimator is $\sqrt{n}$-consistent. The results are valid under rather mild conditions on the design $X_i, i = 1, \dots, n$. Another important feature of this procedure is that it is fully adaptive with respect to unknown smoothness properties of the link function. Though we do not address the problem of its asymptotic efficiency, we note that a $\sqrt{n}$-estimator can be used as a departure point for a so-called "one-step efficient estimator" as discussed, for example, in Delecroix, Härdle and Hristache (1997).

The paper is organized as follows. In the following section we describe the estimation algorithm. The properties of the proposed algorithm are studied in Section 3. In Section 4 we consider details of implementation of the proposed estimator and present some simulation results. The proofs are collected in Section 5.

2. Estimation procedure.

We start with the informal description of the proposed estimator. Our approach (as in the case of the average derivative estimator) is based on the obvious fact that under the model (1.2), the gradient $\nabla f(x) = \partial f(x)/\partial x = \theta^* g_0'(X_i^\top \theta^*)$ of the regression function f at every point x is proportional to θ^*. This leads to the idea of estimating the average gradient,

$$\beta^* := \frac{1}{n} \sum_{i=1}^{n} \nabla f(X_i) = \theta^* \frac{1}{n} \sum_{i=1}^{n} g_0'(X_i^\top \theta^*). \tag{2.1}$$

Clearly β^* is a linear functional of the unknown regression function f, so that one can apply here the well-developed theory of estimation of linear functionals [see, e.g., Ibragimov and Khasminski (1987) and references therein]. The main problem which arises when implementing this approach is that the gradient function is not smooth and some rather restrictive assumptions on the design and on the link function g_0 must hold to ensure the desirable $\sqrt{n}$-consistency of the corresponding estimator [Samarov (1991, 1993), Härdle and Tsybakov (1993)]. The vectors β^* and θ^* can be estimated naturally using the expression

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} \widehat{\nabla f}(X_i) \quad \text{and} \quad \hat{\theta}_1 = \frac{\hat{\beta}_1}{|\hat{\beta}_1|}, \tag{2.2}$$

where $\widehat{\nabla f}(X_i)$ is the pilot estimator of the gradient $\nabla f(X_i)$ of f w.r.t. x at the point X_i. A standard way to estimate both the value $f(X_i)$ and the gradient vector $\nabla f(X_i)$ [cf. Fan and Gijbels (1996)] is to use the local least squares algorithm,

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \operatorname*{arginf}_{c\in\mathbb{R},\, \beta\in\mathbb{R}^d} \sum_{j=1}^{n} \Big[Y_j - c - \beta^\top (X_j - X_i) \Big]^2 K\bigg(\frac{|X_j - X_i|^2}{h^2} \bigg),$$

where *a kernel* $K(\cdot)$ is positive and supported on $[-1, 1]$, so that the weights of all points X_j outside a neighborhood $U_h(X_i)$ of diameter h around X_i vanish. The solution to this quadratic optimization problem can be represented as

$$\begin{aligned} \begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} &= \bigg\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top K\bigg(\frac{|X_{ij}|^2}{h^2} \bigg) \bigg\}^{-1} \\ &\quad \times \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\bigg(\frac{|X_{ij}|^2}{h^2} \bigg), \end{aligned} \tag{2.3}$$

where $X_{ij} = X_j - X_i$. One can show (see Proposition 1 below) that the loss $|\hat{\beta} - \beta^*|$ of this estimator (which is the Euclidean norm of the vector $\hat{\beta} - \beta^*$) can be bounded as follows:

$$|\hat{\beta}_1 - \beta^*| \le C_1 h + \frac{|\xi|}{h\sqrt{n}}. \tag{2.4}$$

Here ξ is a Gaussian random vector in $\mathbb{R}^d$ with $\mathbf{E}\xi = 0$ and $\mathbf{E}|\xi|^2 \le C_2$, and C_1, C_2 are some fixed constants. The right-hand side of (2.4) consists of two terms. The first term bounds the deterministic error (the bias), which is due to the error of local approximation of f by a linear function. This error is proportional to h. The second term is the stochastic error $|\xi|/(\sqrt{n}h)$ which is independent of f; this term is typically of order $(\sqrt{n}h)^{-1}$. The balance of these two terms leads to the choice of h of order $n^{-1/4}$ and hence to the error

$$|\hat{\beta}_1 - \beta^*| = O(n^{-1/4})$$

and similarly for $|\hat{\theta} - \theta^*|$ [provided that the quantity $|\beta^*| = |(1/n)\sum_{i=1}^{n} g'(X_i^\top \theta^*)|$ is separated away from zero]. The situation becomes even worse if the dimension $d > 4$. The reason for this lies in the data sparseness. Indeed, in order to provide $d + 1$ design points which are necessary to compute the local linear approximation (2.3) in a ball of radius h, one should take h of order $n^{-1/d}$. This leads to the bias $O(n^{-1/d})$ in (2.4). Therefore, for $d > 4$ the accuracy of such an estimator would be order $n^{-1/d}$. This rate of convergence $(n^{-1/(4\vee d)})$ is, of course, much worse than $n^{-1/2}$ that can be attained for this problem. Fortunately, the simple estimator $\hat{\theta}_1$ can be significantly refined. The idea is to employ the structural assumption (1.2) for improving the quality of the gradient estimation.

We use the following observation. To recover the vector θ^* we do not need to know the vector β^* entirely. Only the direction of β^* is important and therefore, if $\hat{\beta}$ estimators β^*, it suffices to ensure that the difference between $\hat{\beta}$ and its projection $(\hat{\beta}^\top\theta^*)\theta^*$ on the vector θ^* is small. If, in addition, $|\beta^*|$ is separated away from zero, then $\hat{\beta}/|\hat{\beta}|$ estimates reasonably the unit vector θ^*. Therefore, our intention is to modify the simple estimators (2.3) in a way that the quality of estimation improves in the direction orthogonal to θ^*.

Suppose for a moment that we know θ^* and estimate $\nabla f(X_i) = \theta^* g_0'(X_i^\top\theta^*)$. Note that the regression function $f(x)$ and, hence, the gradient function $\nabla f(x)$ do not vary within the subspace which is orthogonal to θ^*. This implies that within the strip $S = \{x\colon |(x - X_i)^\top\theta^*| \le \rho\}$, where ρ is small, the function f can be nicely approximated with a linear function. Let now $\widehat{\nabla f}(X_i)$ be the estimator of $\nabla f(X_i)$ based [in the same way as in (2.3)] on the local linear approximation of f over the strip $S(X_i)$; that is,

$$\begin{aligned}\begin{pmatrix}\hat{f}(X_i)\\ \widehat{\nabla f}(X_i)\end{pmatrix} &= \operatorname*{arginf}_{c\in\mathbb{R},\, \beta\in\mathbb{R}^d} \sum_{j=1}^n \Big[Y_j - c - \beta^\top(X_j - X_i)\Big]^2 \\ &\qquad\qquad \times K\left(\frac{|(X_j - X_i)^\top\theta^*|^2}{\rho^2}\right).\end{aligned} \tag{2.5}$$

When applying the averaging we obtain the correspondent estimators $\hat{\beta}$ and $\tilde{\theta}$,

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^n \widehat{\nabla f}(X_i) \quad\text{and}\quad \tilde{\theta} = \hat{\beta}/|\hat{\beta}|.$$

The parameter ρ in (2.5) can be selected small, for example, $n^{-1/3}$, independently of the dimensionality d and one still has enough design points in (almost) every strip $S(X_i)$. One can also show that the bias of the so-defined gradient estimators $\widehat{\nabla f}(X_i)$ in the direction orthogonal to θ^* is of order ρ^2. Moreover, for a properly selected ρ and under some regularity conditions, the estimator $\hat{\theta}$ is asymptotically normal with the rate of convergence $O(n^{-1/2})$. Unfortunately, this estimator cannot be implemented since it involves explicitly the target vector θ^*. A natural idea here is to replace θ^* in this construction by its pilot estimator $\hat{\theta}_1$ as in (2.2). This leads to the following iterative procedure; compare Carroll, Fan, Gijbels and Wand (1997). We start with the usual estimator $\hat{\beta}_1$ from (2.2) and (2.3) with some $h = h_1$. Although this estimator is very rough, it delivers some useful information about θ^*. At the next step we update the gradient estimators $\widehat{\nabla f}_2(X_i)$, using the elliptic windows $\{x\colon |S_2(x - X_i)| \le h_2\}$, with $S_2 = (I + \rho_2^{-2}\hat{\beta}_1\hat{\beta}_1^\top)^{-1/2}$ for some $\rho_2 < \rho_1 = 1$ and $h_2 > h_1$ instead of the spherical windows $\{x\colon |x - X_i| \le h_1\}$. In other words, we shrink the original windows in the direction $\hat{\beta}_1$ (since $\rho_2 < 1$) and stretch them in all the orthogonal directions (since $h_2 > h_1$):

$$\begin{pmatrix} \hat{f}_2(X_i) \\ \widehat{\nabla f}_2(X_i) \end{pmatrix} = \operatorname*{arginf}_{c\in\mathbb{R},\, \beta\in\mathbb{R}^d} \sum_{j=1}^{n} \bigl[Y_j - c - \beta^\top (X_j - X_i)\bigr]^2 K\biggl(\frac{|S_2(X_j - X_i)|^2}{h_2^2}\biggr)$$

$$= \biggl\{\sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K\biggl(\frac{|S_2 X_{ij}|^2}{h_2^2}\biggr)\biggr\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\biggl(\frac{|S_2 X_{ij}|^2}{h_2^2}\biggr).$$

This leads to the estimator $\hat{\beta}_2 = (1/n)\sum_{i=1}^{n} \widehat{\nabla f}_2(X_i)$ of β^*. We continue this way each time compressing the averaging windows in the direction of the current estimator $\hat{\beta}_k$ and expanding them in the hyperplane orthogonal to $\hat{\beta}_k$, so that the final windows look very much like flat layers orthogonal to β^*. After $k = k(n)$ iterations, the algorithm delivers the estimator $\hat{\theta} = \hat{\beta}_{k(n)}/|\hat{\beta}_{k(n)}|$.

Now we present the formal description of the estimator.

2.1. *Iterative procedure.* The procedure involves input parameters $\rho_{\min} < \rho_1$ and $h_1 < h_{\max}$, so that ρ decreases geometrically from ρ_1 to $\rho_{\min}$ by the factor a_ρ and h increases geometrically from h_1 to $h_{\max}$ by the factor a_h during iterations. The choice of these parameters will be discussed in the next section. The algorithm reads as follows:

1. Initialization: specify parameters $\rho_1, \rho_{\min}, a_\rho, h_1, h_{\max}, a_h, k = 1, \hat{\beta}_0 = 0$.
2. Compute $S_k = (I + \rho_k^{-2}\hat{\beta}_{k-1}\hat{\beta}_{k-1}^\top)^{1/2}$.
3. For every $i = 1, \dots, n$, compute $\widehat{\nabla f}_k(X_i)$ from the expression

$$\begin{pmatrix} \hat{f}_k(X_i) \\ \widehat{\nabla f}_k(X_i) \end{pmatrix} = \biggl\{\sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K\biggl(\frac{|S_k X_{ij}|^2}{h_k^2}\biggr)\biggr\}^{-1} \times \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\biggl(\frac{|S_k X_{ij}|^2}{h_k^2}\biggr),$$

   where $X_{ij} = X_j - X_i$.
4. Compute the vector $\hat{\beta}_k = (1/n)\sum_{i=1}^{n} \widehat{\nabla f}_k(X_i)$.
5. Set $h_{k+1} = a_h h_k$, $\rho_{k+1} = a_\rho \rho_k$. If $\rho_{k+1} > \rho_{\min}$, then set $k = k + 1$ and continue with step 2; otherwise terminate.

By $k(n)$ we denote the total number of iterations. The last iteration estimator $\hat{\beta} = \hat{\beta}_{k(n)}$ will be used for constructing the estimator of θ^*: $\hat{\theta} = \hat{\beta}/|\hat{\beta}|$.

It is worth mentioning that the number of iterations in the proposed algorithm is always finite (logarithmic in n) and hence one cannot speak of convergence of the estimator during iterations.

2.2. *Choice of parameters of the algorithm.* It is obvious that the quality of estimation by the proposed method strongly depends on the rule for changing the parameters h and ρ, and, in particular, on their values at the initial and final iteration. The values h_k increase during iteration from h_1 to $h_{\max}$ while

ρ_k decrease from $\rho_1 = 1$ to $\rho_{\min}$. The value h_1 is to be selected in such a way that for every (or almost every) point X_i, the estimator $\widehat{\nabla f}(X_i)$ is well defined. A necessary (and usually sufficient) condition is that every ball $\{x\colon |x - X_i| \le h_1\}$ contains at least $d+1$ design points (see the modified procedure in the next section for more discussion). The values of h and ρ at the last iteration $k(n)$ can be obtained by minimizing the risk of the estimator $\hat{\theta}$ (see Corollary 1 in Section 3.3). It leads to the following recommendation: the value $h_{k(n)}$ at the last iteration should be as large as possible, that is, about 1; in opposition, the value $\rho_{\min}$ should be selected as small as possible, but still providing enough design points in every (or almost every) local ellipsoidal neighborhood $E_k(X_i) = \{x\colon |S_k(x - X_i)| \le h_k\}$. We propose the following empirical rule:

$$\begin{array}{lll} \rho_1 = 1, & \rho_{\min} = n^{-1/3}, & a_\rho = e^{-1/6}, \\ h_1 = C_0 n^{-1/4\vee d}, & h_{\max} = C_0, & a_h = e^{1/2(4\vee d)}, \end{array} \tag{2.6}$$

where $C_0 \ge 1$ is to be defined depending on the design; see the modified procedure for a proposal. The rule (2.6) obviously leads to the number of iterations $k(n) \approx \log_{a_\rho}(\rho_1/\rho_{\min}) = 2\log n$ providing $h_{k(n)} \approx h_{\max}$.

Note also that every neighborhood $E_k(X_i)$ is stretched at each iteration step by factor a_h in all directions and is shrunk by factor a_ρ in direction of the estimator $\hat{\theta}_k$. Therefore, the Lebesgue measure of every such neighborhood is changed each time by the factor $e^{d/(2(4\vee d)-1/6)}$ which is larger or equal to 1 for all $d \ge 2$. Under the assumption of a random design with a positive density, this would lead to an increase of the mean number of design points inside each $E_k(X_i)$.

3. Theoretical properties of the index estimator.

In this section we present some results describing the properties of the estimator $\hat{\theta}$. First we introduce another model representation which seems to be more convenient for our purposes.

Let the average gradient vector β^* be defined by (2.1). Clearly β^* is proportional to θ^*. Therefore, if $|\beta^*| > 0$, we may rewrite the model assumption (1.2) in the form

$$f(x) = g(x^\top \beta^*), \tag{3.1}$$

where the new link function g is defined by $g(u) = g_0(u/|\beta^*|)$.

3.1. *Assumptions.*

We consider the following assumptions.

ASSUMPTION 1 (Kernel). The kernel $K(\cdot)$ is a continuously differentiable decreasing function on $\mathbb{R}_+$ with $K(0) = 1$ and $K(x) = 0$ for all $|x| \ge 1$.

ASSUMPTION 2 (Errors). The random variables ε_i in (1.1) are independent and normally distributed with zero mean and variance σ^2.

ASSUMPTION 3 (Link function). The function g from (3.1) is twice differentiable with a bounded second derivative, so that, for some constant C_g and for all $u, v \in \mathbb{R}$, it holds

$$|g(v) - g(u) - (v-u)g'(u)| \le C_g |u - v|^2.$$

Our last assumption concerns the design properties. In what follows we assume that the design is deterministic. That is, $X_1, \dots, X_n$ are nonrandom points in $\mathbb{R}^d$. Note, however, that the case of a random design can be considered as well, supposing $X_1, \dots, X_n$ independent and identically distributed random points in $\mathbb{R}^d$ with a design density $p(x)$. Then all the results should be understood conditionally on the design.

In order for Algorithm 1 to work, we have to suppose that the design points (X_i) are "well diffused" and, as a consequence, all the matrices

$$V_k(X_i) = \sum_{j=1}^{n} \binom{1}{X_{ij}} \binom{1}{X_{ij}}^{\top} K\left(\frac{|S_k X_{ij}|^2}{h_k^2}\right),$$

$$i = 1, \dots, n, \qquad k = 1, \dots, k(n)$$

are nondegenerated. Here $S_k = (I + \rho_k^{-2} \hat{\beta}_{k-1} \hat{\beta}_{k-1}^{\top})^{1/2}$ with $\hat{\beta}_{k-1}$ being the estimator of the vector β^* constructed at the preceding iteration step. We also introduce an "ideal" matrix $S_k^* = (I + \rho_k^{-2} \beta^* (\beta^*)^{\top})^{1/2}$ and define the matrix

$$U_k = (S_k^*)^{-1} S_k^2 (S_k^*)^{-1}.$$

This matrix U_k characterizes the accuracy of estimating the vector β^* by $\hat{\beta}_{k-1}$. If $\hat{\beta}_{k-1} = \beta^*$, then $U_k = I$. We shall see that these matrices U_k are typically close to I. Given a matrix U and $k \le k(n)$ we define

$$N_{i,k}(U) = \sum_{j=1}^{n} K(Z_{ij,k}^{\top} U Z_{ij,k}), \qquad i = 1, \dots, n,$$

$$\mathscr{V}_{i,k}(U) = \sum_{j=1}^{n} \binom{1}{Z_{ij}} \binom{1}{Z_{ij}}^{\top} K(Z_{ij,k}^{\top} U Z_{ij,k}), \qquad i = 1, \dots, n,$$

where $Z_{ij,k} = h_k^{-1} S_k^* (X_j - X_i)$. Our design assumption means in particular that the $(d+1) \times (d+1)$-matrices $\mathscr{V}_{i,k}(U)$ are well defined for all U close to I and for all $i \le n$.

In what follows $\|A\|$ stands for the matrix norm associated with the Euclidean vector norm, $\|A\| = \sup_{\lambda} |A\lambda| / |\lambda|$.

ASSUMPTION 4 (Design). There exist constants $C_V, C_K, C_{K'}$ and some $\alpha > 0$, such that for all matrices U satisfying $\|U - I\| \le \alpha$ and for all $k \le k(n)$ the following conditions hold:

(i) The inverse matrices $\mathscr{V}_{i,k}(U)^{-1}$ are well defined and

$$N_{i,k}(U) \|\mathscr{V}_{i,k}(U)^{-1}\| \le C_V, \qquad i = 1, \dots, n.$$

(ii) For $j = 1, \dots, n$,

$$\sum_{i=1}^{n} \frac{1}{N_{i,k}(U)} K\big(Z_{ij,k}^{\top} U Z_{ij,k}\big) \le C_K,$$

$$\sum_{i=1}^{n} \frac{1}{N_{i,k}(U)} \Big| K'\big(Z_{ij,k}^{\top} U Z_{ij,k}\big)\Big| \le C_{K'}.$$

Here K' means the derivative of the kernel K.

REMARK 1. It can be checked that in the case of random design with a continuous positive density one can fix some constants C_V, C_K and $C_{K'}$ (which depend on the dimension d and the design distribution) such that the bounds in Assumption 4 hold with probability which converges to 1 exponentially as n grows.

3.2. *Accuracy of the estimator* $\hat{\theta}$. In what follows by C, C_1, C_2, etc. we denote generic constants depending on d, C_g, C_V, C_K and σ only.

THEOREM 1. *Let Assumptions* 1 *through* 4 *hold. Under the condition*

$$|\beta^*| > 4\sqrt{2}\sigma C_V C_K z_n n^{-1/2} \tag{3.2}$$

with $z_n = (1 + 2\log n + 2\log\log n)^{1/2}$, *it holds for n sufficiently large*,

$$\mathbf{P}\left(\left|(\hat{\theta} - \theta) - \frac{\gamma^*}{\sqrt{n}}\right| > \frac{C z_n^2 n^{-2/3}}{|\beta^*|}\right) \le \frac{3k(n)}{n},$$

where γ^* is a Gaussian random vector in $\mathbb{R}^d$ with $\mathbf{E}\gamma^* = 0$ *and*

$$\mathbf{E}|\gamma^*|^2 \le 2\sigma^2 C_V^2 C_K^2 |\beta^*|^{-2}.$$

Here $k(n)$ is the total number of iterations, $k(n) \le C \log n$.

REMARK 2. Note that one of the consequences of Theorem 1 is that the normalized error $\sqrt{n}(\hat{\theta} - \theta)$ is close in distribution to the Gaussian vector γ^*. Further, since the error $\hat{\theta} - \theta$ is bounded, it also implies that $E|\hat{\theta} - \theta| = O(n^{-1/2})$.

REMARK 3. The result of Theorem 1 is essentially nonasymptotic; that is, the estimator $\hat{\theta}$ delivers the given accuracy of estimation with a given probability close to 1. The only requirement for this is some minimal number of observations; that is, n should be larger that some value n_0 depending on the model we consider.

3.3. *Properties of the estimator $\hat{\beta}$.* The following results are used in the proof of Theorem 1. They are of certain interest on their own. We start with the description of the accuracy of the first step estimator $\hat{\beta}_1$.

PROPOSITION 1. *Under Assumptions* 1 *through* 4, *it holds*

$$\hat{\beta}_1 - \beta^* = s_1 h_1 + \frac{\eta_1}{h_1\sqrt{n}},$$

where s_1 is a deterministic vector in $\mathbb{R}^d$ satisfying $|s_1| \le \sqrt{2}C_g C_V$ and η_1 is a Gaussian random vector in $\mathbb{R}^d$ with zero mean satisfying $\mathbf{E}|\eta_1|^2 \le 2\sigma^2 C_V^2 C_K^2$. Also

$$\mathbf{P}\biggl(|\hat{\beta}_1 - \beta^*| > \sqrt{2}C_g C_V h_1 + \frac{\sqrt{2}\sigma C_V C_K z}{h_1\sqrt{n}}\biggr) \le z e^{-(z^2-1)/2} \qquad \forall\, z \ge 1.$$

Consider now the "final" estimator $\hat{\beta}$ of β^*. The losses $|\hat{\beta} - \beta^*|$ are not homogeneous w.r.t. the orientation in the space $\mathbb{R}^d$ that is induced by application of elliptic windows for estimating the gradient vectors $\nabla f(X_i)$. To emphasize this property, we introduce for every $k \le k(n)$ the $d \times d$-matrix $P^*_{\rho_k} = (I + \rho_k^{-2}\beta^*(\beta^*)^\top)^{-1/2} = (S_k^*)^{-1}$. Note that when restricted to the hyperplane orthogonal to θ^*, $P^*_{\rho_k}$ coincides with the identity mapping. However, its eigenvalue which corresponds to the eigenvector θ^* is of order ρ_k.

THEOREM 2. *Let Assumptions* 1 *through* 4 *hold. There exists a Gaussian zero mean random vector $\xi^* \in \mathbb{R}^d$ such that, with $\rho = \rho_{k(n)}$ and n large enough,*

$$\mathbf{P}\biggl(\biggl|P^*_\rho(\hat{\beta} - \beta^*) - \frac{\xi^*}{\sqrt{n}}\biggr| > C_1 z_n^2 n^{-2/3}\biggr) \le \frac{3k(n) - 1}{n}$$

and $\mathbf{E}|\xi^*|^2 \le 2\sigma^2 C_V^2 C_K^2$.

COROLLARY 1. *Under the conditions of Theorem* 2, *for every $z \ge 1$,*

$$\mathbf{P}\biggl(\bigl|P^*_\rho(\hat{\beta} - \beta^*)\bigr| > \frac{\sqrt{2}\sigma C_V C_K z}{\sqrt{n}} + C_1 z_n^2 n^{-2/3}\biggr) \le z e^{-(z^2-1)/2} + \frac{3k(n) - 1}{n}.$$

3.4. *Comments.* By inspecting the proof of Theorems 1 and 2 one may conclude that all the results hold in the case of heteroskedastic Gaussian errors ε_i, however, σ^2 is to be understood as $\sup_{1\le i\le n} \mathbf{E}\varepsilon_i^2$. Similarly, the results can be extended to the case of non-Gaussian errors under the condition $\sup_{1\le i\le n} \mathbf{E}\exp(\lambda\varepsilon_i) \le \varkappa_\lambda$ for some positive constants λ and $\varkappa_\lambda$.

One natural question that arises when Theorems 1 and 2 are concerned is what happens if the model assumption is misspecified, that is, if the regression function $f(x)$ does not possess a single-index structure. It is known that the average derivative method gives (under rather restrictive assumptions) a $\sqrt{n}$-consistent estimator of the vector $\int \nabla f(x) w(x) dx$ with some weight function w which depends on the design density [cf. Stoker (1986) and Powell, Stock and Stoker (1989)]. A similar result holds for our first step estimator $\hat{\theta}_1$. However, now the rate of convergence is $n^{-1/4}$ for $d \le 4$ and $n^{-1/d}$ for $d > 4$. Unfortunately, the results of Theorems 1 and 2 cannot be extended without additional assumptions to the situation when the model structure (1.1) is not valid. This issue is confirmed by our simulated results in the next section. We refer to our forthcoming paper for an extension of the above-presented procedure which allows for a multiindex structure and which can be used for testing the single-index assumption.

4. Implementation and simulation results. This section illustrates the performance of the proposed procedure for some simulated data sets. First we present a slightly modified procedure which allows us to deal with an irregular design. This issue turned out to be important for the performance of the method with small and moderate sample sizes.

We start by some informal discussion. In the algorithm described above in Section 2 the estimator $\hat{\beta}_k$ is defined at each step as a linear combination of the estimated gradient vectors $\widehat{\nabla f}(X_i)$. To ensure good asymptotic properties of the procedure, the estimators $\widehat{\nabla f}(X_i)$ should be well defined. This requires some local regularity of the design in the corresponding neighborhoods of the points X_i (cf. Assumption 4). If the condition of design regularity does not hold even in a few points, the variance of the gradient estimators at these points can be very large, which may destroy the quality of the index estimators $\hat{\beta}$. This problem can be efficiently dealt with by using the following weighted scheme. The idea is to multiply each summand in the expression for $\hat{\beta}_k$ by some weight which expresses the degree of local regularity of the design. This leads to the following *modified procedure*.

1. Initialization: specify parameters $\rho_1, \rho_{\min}, a_\rho, h_1, h_{\max}, a_h, C_w$. Define $\bar{w}$ as the square root of the minimal eigenvalue of the matrix $\overline{\mathscr{V}}$ with

$$\overline{\mathscr{V}} = \frac{1}{\mathbf{E}K(\zeta^\top\zeta)}\mathbf{E}\binom{1}{\zeta}\binom{1}{\zeta}^\top K(\zeta^T\zeta),$$

where ζ is random and uniformly distributed over the ball $B_1 = \{x \in \mathbb{R}^d : |x| \le 1\}$: $\bar{w}^2 = \lambda_{\min}(\overline{\mathscr{V}})$; set $k = 1$, $\hat{\beta}_0 = 0$.

2. Compute $S_k = (I + \rho_k^{-2}\hat{\beta}_{k-1}\hat{\beta}_{k-1}^\top)^{1/2}$.
3. For every $i = 1, \dots, n$, compute the matrix $\widehat{\mathscr{V}}_k(X_i)$ with

$$\widehat{\mathscr{V}}_k(X_i) = \frac{1}{\sum_{j=1}^n K(|W_{ij,k}|^2)} \sum_{j=1}^n \binom{1}{W_{ij,k}}\binom{1}{W_{ij,k}}^\top K(|W_{ij,k}|^2),$$

where $W_{ij,k} = h_k^{-1}S_k(X_j - X_i)$ and define w_i as the square root of the minimal eigenvalue of $\widehat{\mathscr{V}}_k(X_i)$: $w_i^2 = \lambda_{\min}(\widehat{\mathscr{V}}_k(X_i))$.

4. If the condition

$$w_1 + \cdots + w_n \ge nC_w\bar{w}$$

is not fulfilled then increase h_k by the factor a_h, that is, $h_k := a_h h_k$. If $h_k > h_{\max}$, terminate, otherwise repeat from Step 3.

5. For every $i = 1, \dots, n$, compute $\widehat{\nabla f}_k(X_i)$:

$$\begin{pmatrix} \hat{f}_k(X_i) \\ \widehat{\nabla f}_k(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K\left(\frac{|S_k X_{ij}|^2}{h_k^2} \right) \right\}^{-1} \times \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\left(\frac{|S_k X_{ij}|^2}{h_k^2} \right).$$

6. Compute the vector $\hat{\beta}_k = (\sum_{i=1}^n w_i)^{-1} \sum_{i=1}^n \widehat{\nabla f}_k(X_i) w_i$ with the previously obtained w_i's. If $|\hat{\beta}_k| > 1$, then normalize $\hat{\beta}_k := \hat{\beta}_k / |\hat{\beta}_k|$.
7. Set $\rho_{k+1} = a_\rho \rho_k$ and $h_{k+1} = \min\{a_h h_k, h_{\max}\}$. If $\rho_{k+1} > \rho_{\min}$, then set $k = k + 1$ and continue with step 2; otherwise terminate.

The estimator $\hat{\beta}_k$ from the last iteration is applied for estimating θ^*: $\hat{\theta}_{\text{mod}} = \hat{\beta}_k / |\hat{\beta}_k|$.

In our simulation study we apply the modified procedure with the following parameter setting:

$$\begin{aligned} &h_1 = n^{-1/4 \vee d}, && h_{\max} = 2\sqrt{d}, && a_h = e^{1/2(4 \vee d)}, \\ &\rho_1 = 1, && \rho_{\min} = n^{-1/3} / h_{\max}, && a_\rho = e^{-1/6}. \end{aligned} \tag{4.1}$$

We also set $C_w = 0.5$ and apply the quartic kernel $K(t) = (1 - t^2)_+^2$.

The objective of our simulation study is to illustrate the following features of the procedure:

1. How the quality of estimation improves during iteration.
2. Dependence on the sample size n, dimensionality d and the noise level σ.
3. Relative performance to the one-step estimator with the "ideal" bandwidth.
4. Behavior of the estimator when the single-index assumption does not hold.

Note that the first-step estimator of the algorithm can be viewed as a version of the *average derivative estimator* (*ADE*) which is a natural competitor to the iterative estimator. In our study we calculate this estimator selecting the parameter h_1 by optimizing the corresponding mean losses (the "ideal" bandwidth).

The performance of the method is illustrated by means of the following examples. We consider the model described by (1.1) and (1.2) with $\sigma = 0.1, 0.2, 0.4$ and 0.8. The design $X_1, \dots, X_n$ is modeled randomly in the cube $[-1, 1]^d$ with independent components so that every component of $(X_i + 1)/2$ follows $B(\tau, 1)$-distribution. The parameter τ controls the skewness of the beta-distribution with $\tau = 1$ corresponding to the uniform design. We also set $g(u) = u^2 e^u$ and $\theta = (1, 2, 0, \dots, 0)^\top / \sqrt{5}$. The estimation error is measured in the l_1-norm in $\mathbb{R}^d$: $\|\hat{\theta} - \theta^*\|_1 = \sum_{j=1}^d |\hat{\theta}_j - \theta_j|$ which allows to easily evaluate the "error per parameter."

The empirical results for the mean absolute error $E\|\hat{\theta} - \theta^*\|_1$ based on 250 Monte Carlo replicates are collected in Table 1. The last column displays the

TABLE 1

MAE $\mathbf{E}\|\hat{\theta} - \theta^*\|_1$ *for ADE with optimal bandwidth and first, fifth, tenth and last iteration*[1]

n	d	σ	τ	ADE	First step	Fifth step	Tenth step	Last step	ADE/last
200	4	0.1	1	0.1165	0.1228	0.0868	0.0762	0.0425	2.74
200	4	0.2	1	0.1584	0.1584	0.1146	0.1063	0.0838	1.89
400	4	0.1	1	0.0770	0.0940	0.0606	0.0488	0.0294	2.62
400	4	0.2	1	0.1116	0.1157	0.0792	0.0709	0.0584	1.91
400	4	0.4	1	0.1741	0.1751	0.1277	0.1253	0.1165	1.49
800	4	0.1	1	0.0537	0.0782	0.0461	0.0344	0.0205	2.62
800	4	0.2	1	0.0809	0.0927	0.0585	0.0497	0.0409	1.98
800	4	0.4	1	0.1302	0.1322	0.0910	0.0870	0.0817	1.59
800	4	0.8	1	0.2188	0.2247	0.1663	0.1675	0.1634	1.34
200	6	0.1	1	0.2614	0.2614	0.1490	0.1023	0.0702	3.72
200	6	0.2	1	0.3054	0.3054	0.1954	0.1542	0.1391	2.20
400	6	0.1	1	0.1688	0.1749	0.0927	0.0609	0.0468	3.61
400	6	0.2	1	0.2092	0.2094	0.1267	0.1000	0.0932	2.25
400	6	0.4	1	0.2974	0.3084	0.2101	0.1883	0.1856	1.60
800	6	0.1	1	0.1120	0.1284	0.0635	0.0405	0.0314	3.57
800	6	0.2	1	0.1466	0.1511	0.0858	0.0671	0.0629	2.33
800	6	0.4	1	0.2131	0.2131	0.1402	0.1265	0.1257	1.70
800	6	0.8	1	0.3435	0.3608	0.2626	0.2486	0.2479	1.39
200	10	0.1	1	0.6094	0.6094	0.3597	0.2048	0.1397	4.36
200	10	0.2	1	0.6729	0.6752	0.4410	0.3027	0.2773	2.43
400	10	0.1	0.75	0.7670	0.8799	0.6841	0.5528	0.1447	5.30
400	10	0.1	1	0.4186	0.4196	0.2163	0.1105	0.0822	5.09
400	10	0.1	1.5	0.2482	0.2617	0.1958	0.1378	0.0412	6.02
400	10	0.2	1	0.4665	0.4665	0.2726	0.1763	0.1659	2.81
400	10	0.4	1	0.5916	0.6082	0.4210	0.3247	0.3287	1.80
800	10	0.1	1	0.2939	0.2939	0.1351	0.0678	0.0536	5.48
800	10	0.2	1	0.3272	0.3273	0.1758	0.1142	0.1070	3.06
800	10	0.4	1	0.4190	0.4262	0.2760	0.2151	0.2124	1.97
800	10	0.8	1	0.6302	0.6778	0.5018	0.4244	0.4112	1.53

[1] All values are estimated from 250 simulations.

relative improvement given by the iterative procedure compared with the ADE with the optimal (risk minimizing) bandwidth. This improvement is defined as the ratio of the corresponding risks. The results clearly illustrate how the estimation quality is improved during iterations and they are in agreement with the asymptotic root-n consistency of the estimator. Compared to the average derivative estimator with the "ideal" bandwidth choice, the proposed procedure provides superior results in all situations although the improvement is design dependent. For instance, it decreases as σ increases which has a simple explanation: the iterative procedure allows reducing the bias of the estimator but the stochastic error is still there and it increases with σ while the relative improvement decreases.

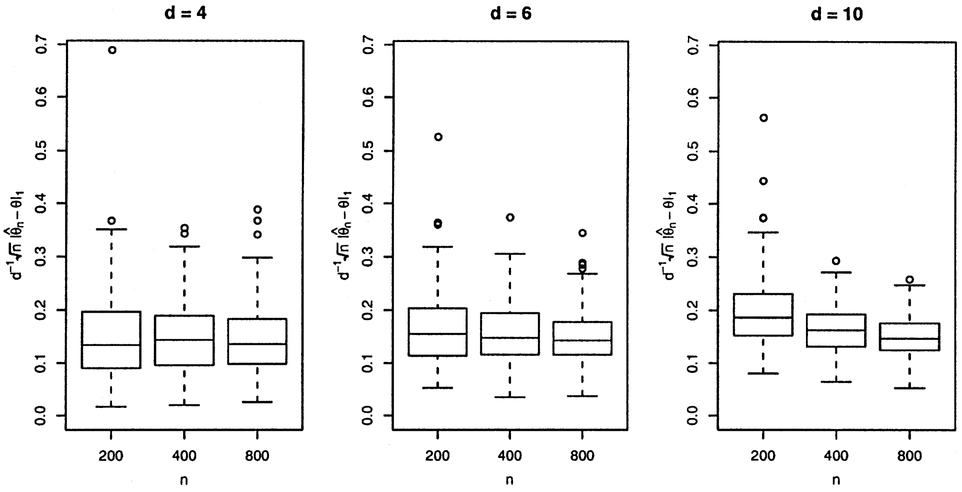

FIG. 1. *Box plots for* $d^{-1}\sqrt{n}\|\hat{\theta}_n - \theta^*\|_1$ *with* $\sigma = 0.1$ *and* $n = 200, 400, 800$.

It is also worth mentioning that the improvement attained by the iterative procedure increases with dimension d. For instance, with $n = 800$ and $\sigma = 0.1$, it varies from the factor 2.62 for $d = 4$ to 5.48 for $d = 10$. This fact is fully in agreement with our heuristic discussion and theoretical results. Indeed, the bias component in the risk of the ADE increases with dimension d but the iterative procedure allows one to eliminate it.

Another interesting observation is that the improvement increases for assymetric design.

Figure 1 displays Box-and-Whisker plots based on 250 replicates for $d^{-1}\sqrt{n}\|\hat{\theta} - \theta^*\|_1$ for different values of dimension d and sample size n with fixed $\sigma = 0.1$. Box plots are produced by function *boxplot* from *R*; see *The R Reference index* on http://www.ci.tuwien.ac.at/R/ for details. Similar box plots for $\sigma^{-1}d^{-1}\sqrt{n}\|\hat{\theta}-\theta^*\|_1$ with $n = 800$ but different values of σ are given in Figure 2. The figures clearly indicate that the losses of the iterative estimator are essentially proportional to the noise level σ and to $n^{-1/2}$. One can also see that the estimation "error per parameter" does not increase with the dimension d and that the variance and the interquartile range of the relative error even decrease.

Table 2 shows how the estimator works in the case where the single-index model assumption does not hold. We consider the ten-dimensional situation ($d = 10$) for the sample size $n = 400$ and $\sigma = 0.1$, the covariates are uniformly distributed over the cube $[-1, 1]^d$ and

$$f(x) = \sqrt{1 - \eta^2 x_1^2 e^{x_1}} + \eta x_2^2 e^{x_2}. \tag{4.2}$$

Here the parameter η controls the deviation from the single-index model (corresponding to $\eta = 0$) with $\eta = 1/\sqrt{2}$ leading to an essentially double-index model. For the model (4.2) the mean gradient β^* (averaged w.r.t. the design

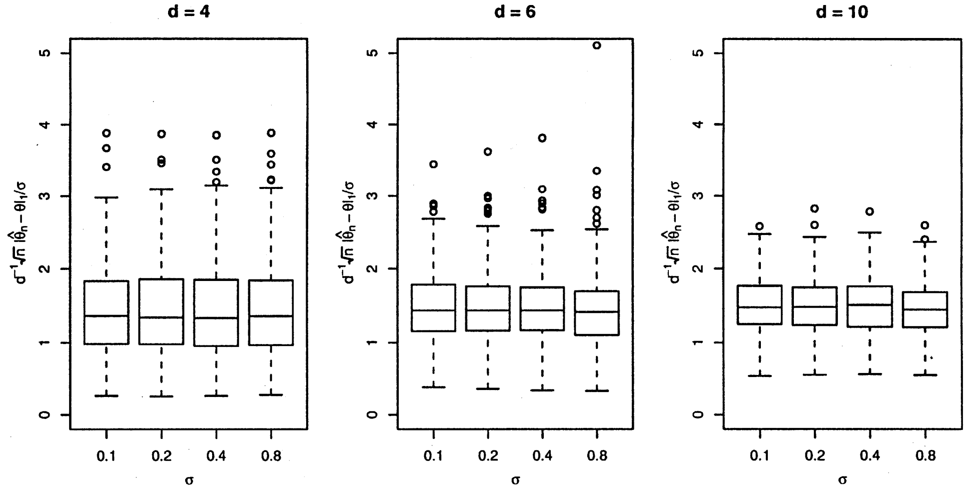

FIG. 2. *Box plots for* $\sigma^{-1}d^{-1}\sqrt{n}\|\hat{\theta}-\theta^*\|_1$ *with* $n = 800$ *and* $\sigma = 0.1, 0.2, 0.4, 0.8$.

distribution) is clearly in the direction $\theta^* = (\sqrt{1-\eta^2}, \eta, 0, \dots, 0)^\top$ with the same value of $|\beta^*| = 1$ for all η.

One can see that the iterative procedure still works for small values of η, that is, if the single-index assumption is not significantly violated. For η approaching $1/\sqrt{2}$, the starting iterations still provide some improvement while the further steps lead to a moderate loss of accuracy compared to the first step estimator.

The general conclusion of this simulation study is that the proposed procedure works well even for reasonably small sample sizes under a variety of design assumptions. It outperforms the average derivative estimator in all situations we considered especially in high dimensions, provided that the underlying single-index assumption is not significantly violated.

TABLE 2

MAE $\boldsymbol{E}\|\hat{\theta}-\theta^*\|_1$ *for ADE with optimal bandwidth, and first, fifth, tenth and last iteration for the model* $f(x) = \sqrt{1-\eta^2}x_1^2e^{x_1} + \eta x_2^2e^{x_2}$[1]

n	d	σ	η	ADE	First step	Fifth step	Tenth step	Last step	ADE/last
400	10	0.1	0.707	0.3817	0.4098	0.3768	0.4119	0.4791	0.80
400	10	0.1	0.500	0.3846	0.3937	0.3328	0.3495	0.3863	1.00
400	10	0.1	0.250	0.3544	0.3555	0.2362	0.1980	0.2020	1.75
400	10	0.1	0.125	0.3401	0.3402	0.1940	0.1253	0.1193	2.85
400	10	0.1	0.0	0.3341	0.3341	0.1766	0.0827	0.0637	5.24

[1] All values are estimated from 250 simulations.

Another conclusion of our study is that the iterative procedure needs some minimal number of observations to start working. This value increases with the noise level.

If the underlying single-index assumption does not hold, then the iterative procedure does not provide an improvement but delivers approximately the same quality as the ADE with the optimal bandwidth.

5. Proofs. Proofs of Theorems 1 and 2 are based on the following technical statement which qualifies the improvement of the estimator $\hat{\beta}_k$ at each iteration step. Suppose that we are given some fixed values h and ρ (which mean the current values h_k and ρ_k) and a fixed vector $b \in \mathbb{R}^d$ which can be viewed as a pilot estimator $\hat{\beta}_{k-1}$ of β^* obtained at the previous step. Define $S_b = \left(I + \rho^{-2} b b^\top\right)^{1/2}$ and set

$$\begin{pmatrix} \hat{f}_b(X_i) \\ \widehat{\nabla f}_b(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top K\left(\frac{|S_b X_{ij}|^2}{h^2} \right) \right\}^{-1} \times \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\left(\frac{|S_b X_{ij}|^2}{h^2} \right),$$

$$\hat{\beta}_b = \frac{1}{n} \sum_{i=1}^{n} \widehat{\nabla f}_b(X_i),$$

where, recall, $X_{ij} = X_j - X_i$. We aim to evaluate the estimation errors $\hat{\beta}_b - \beta^*$.

PROPOSITION 2. *Let Assumptions* 1 *through* 4 *hold. With* $P_\rho^* = (I + \rho^{-2} \times \beta^*(\beta^*)^\top)^{-1/2}$ *and some positive* $\delta < \rho/4$, *define the set* $\mathfrak{B}_{\delta,\rho} = \{b \in \mathbb{R}^d : |P_\rho^* \times (b - \beta^*)| \le \delta\}$. *There exists a Gaussian vector* η^* *in* $\mathbb{R}^d$ *such that* $\mathbf{E}|\eta^*|^2 \le 2\sigma^2 C_V^2 C_K^2$ *and it holds*

$$\mathbf{P}\left(\sup_{b \in \mathfrak{B}_{\delta,\rho}} \left| P_\rho^*(\hat{\beta}_b - \beta^*) - \frac{\eta^*}{h\sqrt{n}} \right| > \frac{\sqrt{2} C_g C_V h \rho^2}{(1-\alpha)^{3/2}} + \frac{\sigma C_{\alpha,n} \alpha}{h\sqrt{n}} \right) \le 2/n$$

with $\alpha = 2\delta/\rho + \delta^2/\rho^2$ *and*

$$C_{\alpha,n} = \frac{1}{2} \left(\frac{\sqrt{2} C_V C_{K'}}{(1-\alpha)^2} + \frac{2\sqrt{2} C_V^2 C_K C_{K'}}{(1-\alpha)^3} \right) \left(2 + \sqrt{(3+d)\log(4n)} \right). \tag{5.1}$$

Before proving this statement, we present one straightforward corollary.

COROLLARY 2. *It holds under Assumptions* 1 *through* 4 *for every* $z \ge 1$,

$$\mathbf{P}\left(\sup_{b \in \mathfrak{B}_{\delta,\rho}} \left| P_\rho^*(\hat{\beta}_b - \beta^*) \right| > \frac{\sqrt{2} C_g C_V\, h \rho^2}{(1-\alpha)^{3/2}} + \frac{z\sqrt{2}\sigma C_V C_K}{h\sqrt{n}} + \frac{\sigma C_{\alpha,n} \alpha}{h\sqrt{n}} \right) \le z e^{-(z^2-1)/2} + \frac{2}{n}.$$

Indeed, the statement of the corollary follows directly from Proposition 2 and Lemma 7 from the Appendix.

PROOF OF PROPOSITION 2. We begin with the following simple lemma.

LEMMA 1. *If β is a vector in $\mathbb{R}^d$ such that $|P^*_\rho(\beta-\beta^*)| \le \delta$ for some $\delta > 0$, then*

$$\left\| P^*_\rho\big(\beta\beta^\top - \beta^*(\beta^*)^\top\big)P_\rho\right\| \le 2\rho\delta + \delta^2.$$

PROOF. Since

$$\left|P^*_\rho\beta^*\right|^2 = \left\|P^*_\rho\beta^*(\beta^*)^\top P^*_\rho\right\| = \left\|\big(I+\rho^{-2}\beta^*(\beta^*)^\top\big)^{-1}\beta^*(\beta^*)^\top\right\| \le \rho^2,$$

the condition of the lemma yields

$$\begin{aligned}\left\|P^*_\rho\big(\beta\beta^\top - \beta^*(\beta^*)^\top\big)P^*_\rho\right\| &\le 2\|P^*_\rho(\beta-\beta^*)(\beta^*)^\top P^*_\rho\| + \|P^*_\rho(\beta-\beta^*)\big(\beta-\beta^*\big)^\top P^*_\rho\| \\ &\le 2|P^*_\rho(\beta-\beta^*)|\,|P^*_\rho\beta^*| + |P^*_\rho(\beta-\beta^*)|^2 \le 2\delta\rho + \delta^2\end{aligned}$$

as required. □

It is useful to introduce the notation

$$u = \rho^{-1}P^*_\rho b, \qquad U = P^*_\rho\big(I+\rho^{-2}bb^\top\big)P^*_\rho = (P^*_\rho)^2 + uu^\top$$

and similarly,

$$u^* = \rho^{-1}P^*_\rho\beta^*, \qquad U^* = P^*_\rho\Big(I+\rho^{-2}\beta^*(\beta^*)^\top\Big)P^*_\rho = I.$$

Obviously the condition $|P^*_\rho(b-\beta^*)| \le \delta$ implies $|u-u^*| \le \delta/\rho$; that is, the inclusion $b \in \mathfrak{B}_{\delta,\rho}$ is equivalent to $u \in \{u : |u-u^*| \le \delta/\rho\}$ and by Lemma 1 also $\|U-U^*\| = \|uu^\top - u^*(u^*)^\top\| \le 2\delta/\rho + \delta^2/\rho^2 = \alpha$.

Next, for every $i, j \le n$, define $Z_{ij} = h^{-1}(P^*_\rho)^{-1}(X_j - X_i)$ and set

$$\begin{aligned}\mathscr{V}_i(U) &= \sum_{j=1}^n \begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top K(Z_{ij}^\top U Z_{ij}),\\ \hat{s}_i(U) &= h^{-1}\mathscr{V}_i(U)^{-1}\sum_{j=1}^n \begin{pmatrix}1\\ Z_{ij}\end{pmatrix} Y_j K(Z_{ij}^\top U Z_{ij}).\end{aligned}$$

It is easy to check that $\hat{s}_i(U) = \big(h^{-1}\hat{f}_b(X_i), P^*_\rho\widehat{\nabla f}_b(X_i)\big)^\top$ and hence,

$$P^*_\rho\hat{\beta}_b = \mathscr{E}_d n^{-1}\sum_{i=1}^n \hat{s}_i(U), \tag{5.2}$$

where $\mathscr{E}_d$ is the mapping from $\mathbb{R}^{d+1}$ onto $\mathbb{R}^d$ keeping the last d coordinates and leaving out the first one. The model equations (1.1) and (3.1) imply

$$\hat{s}_i(U) = s_i(U) + \zeta_i(U) \tag{5.3}$$

with

$$s_i(U) = h^{-1}\mathscr{V}_i(U)^{-1}\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix} f(X_j)K(Z_{ij}^{\top}UZ_{ij})$$

$$= h^{-1}\mathscr{V}_i(U)^{-1}\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix} g(X_j^{\top}\beta^*)K(Z_{ij}^{\top}UZ_{ij}),$$

$$\zeta_i(U) = h^{-1}\mathscr{V}_i(U)^{-1}\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix} \varepsilon_j K(Z_{ij}^{\top}UZ_{ij})$$

so that

$$P_\rho^*(\hat\beta_b - \beta^*) = \mathscr{E}_d n^{-1}\biggl(\sum_{i=1}^{n} s_i(U) + \sum_{i=1}^{n}\zeta_i(U)\biggr) - n^{-1}\sum_{i=1}^{n} P_\rho^*\nabla f(X_i). \tag{5.4}$$

We define $\eta^* = h\sqrt{n}\mathscr{E}_d\zeta(U^*)$ where $\zeta(U) = n^{-1}\sum_{i=1}^n \zeta_i(U)$. The assertion of the proposition easily follows from (5.4) if we show that

$$\sup_{u:\,|u-u^*|\le\delta/\rho} |\mathscr{E}_d s_i(U) - P_\rho^*\nabla f(X_i)| \le \frac{\sqrt{2}C_g C_V}{(1-\alpha)^{3/2}} h\rho^2, \qquad i = 1,\dots,n, \tag{5.5}$$

$$\mathbf{P}\biggl(\sup_{u\,:|u-u^*|\le\delta/\rho} |\zeta(U) - \zeta(U^*)| > \frac{\sigma C_{\alpha,n}\alpha}{h\sqrt{n}}\biggr) \le 2/n \tag{5.6}$$

with $U = (P_\rho^*)^2 + uu^\top$ and $U^* = I$, and that

$$\mathbf{E}|\zeta(U^*)|^2 \le \frac{2\sigma^2 C_V^2 C_K^2}{h^2 n}. \tag{5.7}$$

Recall that $|u - u^*| \le \delta/\rho$ implies $\|U - I\| \le \alpha$ for $U = (P_\rho^*)^2 - uu^\top$. The following statement will be useful in the sequel.

LEMMA 2. *Let $\|U - I\| \le \alpha < 1$. Then for all i, j with $Z_{ij}^\top UZ_{ij} \le 1$, it holds $|Z_{ij}|^2 \le (1-\alpha)^{-1}$.*

PROOF. Note that the inequalities $Z_{ij}^\top UZ_{ij} \le 1$ and $\|U - I\| \le \alpha$ imply

$$\bigl|Z_{ij}^\top UZ_{ij} - |Z_{ij}|^2\bigr| = \bigl|Z_{ij}^\top (U-I)Z_{ij}\bigr| \le \alpha|Z_{ij}|^2$$

and hence $|Z_{ij}|^2 \le (1-\alpha)^{-1}Z_{ij}^\top UZ_{ij}$. □

Next we check (5.5). Since

$$\begin{pmatrix}h^{-1}f(X_i)\\ P_\rho^*\nabla f(X_i)\end{pmatrix} = \mathscr{V}_i(U)^{-1}\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^{\top}\begin{pmatrix}h^{-1}f(X_i)\\ P_\rho^*\nabla f(X_i)\end{pmatrix} K(Z_{ij}^\top UZ_{ij})$$

$$= h^{-1}\mathscr{V}_i(U)^{-1}\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\bigl\{f(X_i) + (X_j - X_i)^\top\nabla f(X_i)\bigr\}K(Z_{ij}^\top UZ_{ij})$$

it holds

$$\begin{aligned}
s_i(U) - \begin{pmatrix} h^{-1} f(X_i) \\ P_\rho^* \nabla f(X_i) \end{pmatrix}
&= h^{-1} \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \Big\{ f(X_j) - f(X_i) \\
&\qquad\qquad - (X_j - X_i)^\top \nabla f(X_i) \Big\} K(Z_{ij}^\top U Z_{ij}) \\
&= h^{-1} \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} r_{ij} K(Z_{ij}^\top U Z_{ij}),
\end{aligned}$$

where, in view of (3.1),

$$r_{ij} = g(X_j^\top \beta^*) - g(X_i^\top \beta^*) - (X_j - X_i)^\top \beta^* g'(X_i^\top \beta^*).$$

Since the kernel K vanishes outside $[-1,1]$, it suffices to consider only summands with $|Z_{ij}^\top U Z_{ij}| \le 1$. It is clear that

$$\begin{aligned}
|X_j^\top \beta^* - X_i^\top \beta^*|^2 &= (X_j - X_i)^\top \beta^* (\beta^*)^\top (X_j - X_i) \\
&\le \rho^2 (X_j - X_i)^\top \Big(I + \rho^{-2} \beta^* (\beta^*)^\top \Big) (X_j - X_i) = h^2 \rho^2 |Z_{ij}|^2,
\end{aligned}$$

which implies by Lemma 2 and Assumption 3 for every pair (i,j) with $Z_{ij}^\top U \times Z_{ij} \le 1$:

$$|r_{ij}| \le \frac{C_g h^2 \rho^2}{1-\alpha}, \qquad 1 + \left|Z_{ij}\right|^2 \le 1 + \frac{1}{1-\alpha} \le \frac{2}{1-\alpha}.$$

Using Assumption 4 we bound

$$\begin{aligned}
|\mathscr{E}_d s_i(U) - P_\rho^* \nabla f(X_i)| &\le h^{-1} \left| \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} r_{ij} K(Z_{ij}^\top U Z_{ij}) \right| \\
&\le \frac{C_g h \rho^2}{1-\alpha} \| \mathscr{V}_i(U) \|^{-1} \left| \sum_{j=1}^{n} \Big(1 + |Z_{ij}|^2 \Big)^{1/2} K(Z_{ij}^\top U Z_{ij}) \right| \\
&\le \sqrt{2} (1-\alpha)^{-3/2} C_g C_V h \rho^2
\end{aligned}$$

and (5.5) follows.

Further, we study the stochastic component $\zeta(U) = (1/n) \sum_{i=1}^{n} \zeta_i(U)$. It follows directly from the definition that there are vector coefficients $c_i(U)$ such that

$$\zeta(U) = \sum_{i=1}^{n} c_i(U) \varepsilon_i.$$

We show that these coefficients satisfy the following conditions.

LEMMA 3. *It holds that*:

(i)

$$\sum_{i=1}^{n} |c_i(U^*)|^2 \le \frac{2C_V^2 C_K^2}{h^2 n}.$$

(ii)

$$\sup_{U:\|U-I\|\le\alpha} \sum_{i=1}^{n} |c_i(U)|^2 \le \frac{2C_V^2 C_K^2}{(1-\alpha)h^2 n}.$$

(iii) *For every unit vector* $e \in \mathbb{R}^d$,

$$\sup_{U:\|U-I\|\le\alpha} \left\| \frac{d}{dU} e^\top c_i(U) \right\| \le \frac{\varkappa_\alpha}{nh}$$

with

$$\varkappa_\alpha = \sqrt{2}(1-\alpha)^{-3/2} C_V C_{K'} + 2\sqrt{2}(1-\alpha)^{-5/2} C_V^2 C_K C_{K'}. \tag{5.8}$$

PROOF. Define for $i, j = 1, \dots, n$,

$$N_i(U) = \sum_{j=1}^{n} K(Z_{ij}^\top U Z_{ij}), \qquad v_{ij}(U) = \mathscr{V}_i(U)^{-1} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix}.$$

Then

$$\begin{aligned} \zeta(U) &= \frac{1}{nh} \sum_{i=1}^{n} \mathscr{V}_i(U)^{-1} \sum_{j=1}^{n} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \varepsilon_j K(Z_{ij}^\top U Z_{ij}) \\ &= \sum_{j=1}^{n} \left(\frac{1}{nh} \sum_{i=1}^{n} v_{ij}(U) K(Z_{ij}^\top U Z_{ij}) \right) \varepsilon_j \\ &= \sum_{j=1}^{n} c_j(U) \varepsilon_j. \end{aligned}$$

It follows from Lemma 2 and Assumption 4(i) that, if $\|U - I\| \le \alpha$, then for every i, j with $Z_{ij}^\top U Z_{ij} \le 1$, it holds

$$\left| N_i(U) v_{ij}(U) \right| \le C_V (1 + |Z_{ij}|^2)^{1/2} \le C_V \sqrt{2} (1-\alpha)^{-1/2} \tag{5.9}$$

and by Assumption 4(ii),

$$|c_j(U)| \le \frac{\sqrt{2} C_V}{nh(1-\alpha)^{1/2}} \sum_{i=1}^{n} \frac{K(Z_{ij}^\top U Z_{ij})}{N_i(U)} \le \frac{\sqrt{2} C_V C_K}{nh(1-\alpha)^{1/2}}. \tag{5.10}$$

As a particular case, with $U=U^*=I$ and $\alpha=0$, this yields

$$|c_i(U^*)|\le\frac{\sqrt{2}C_VC_K}{nh}$$

and the first two assertions of the lemma follows.

Next we bound the derivative of each coefficient $c_i(U)$ w.r.t. the matrix U. Let e_1 and e_2 be two unit vectors in $\mathbb{R}^d$. Clearly,

$$\frac{d}{dU}e_1^\top\mathcal{V}_i(U)e_2=\sum_{j=1}^n e_1^\top\begin{pmatrix}1\\Z_{ij}\end{pmatrix}\begin{pmatrix}1\\Z_{ij}\end{pmatrix}^\top e_2K'(Z_{ij}^\top UZ_{ij})Z_{ij}Z_{ij}^\top$$

and Lemma 2 and Assumption 4 yield

$$\left\|\frac{d}{dU}e_1^\top\mathcal{V}_i(U)e_2\right\|\le 2(1-\alpha)^{-2}\sum_{j=1}^n|K'(Z_{ij}^\top UZ_{ij})|\le 2(1-\alpha)^{-2}C_{K'}N_i(U).$$

Next, (i) and Assumption 4 provide

$$\begin{aligned}\left\|\frac{d}{dU}e_1^\top\mathcal{V}_i(U)^{-1}e_2\right\| &= \left\|e_1^\top\mathcal{V}_i(U)^{-1}\frac{d}{dU}\mathcal{V}_i(U)e_2\mathcal{V}_i(U)^{-1}\right\|\\ &\le \|\mathcal{V}_i(U)^{-1}\|^2\left\|\frac{d}{dU}e_1^\top\mathcal{V}_i(U)e_2\right\|\\ &\le \frac{C_V^2}{N_i(U)}2(1-\alpha)^{-2}C_{K'}.\end{aligned}$$

These results combined with Lemma 2 and (5.9) imply for every U with $\|U-I\|\le\alpha$ and for every unit vector e,

$$\begin{aligned}&\left\|\sum_{i=1}^n\frac{d}{dU}e^\top v_{ij}(U)K(Z_{ij}^\top UZ_{ij})\right\|\\ &\quad\le\sum_{i=1}^n|Z_{ij}|^2|v_{ij}(U)||K'(Z_{ij}^\top UZ_{ij})|+\sum_{i=1}^n\left\|\frac{d}{dU}e^\top v_{ij}(U)\right\|K(Z_{ij}^\top UZ_{ij})\\ &\quad\le\frac{\sqrt{2}C_V}{(1-\alpha)^{3/2}}\sum_{i=1}^n\frac{|K'(Z_{ij}^\top UZ_{ij})|}{N_i(U)}+\frac{2\sqrt{2}C_V^2C_{K'}}{(1-\alpha)^{5/2}}\sum_{i=1}^n\frac{K(Z_{ij}^\top UZ_{ij})}{N_i(U)}\\ &\quad\le\sqrt{2}(1-\alpha)^{-3/2}C_VC_{K'}+2\sqrt{2}(1-\alpha)^{-5/2}C_V^2C_KC_{K'}=\varkappa_\alpha.\end{aligned}$$

This implies for every $i\le n$,

$$\left\|\frac{d}{dU}e^\top c_i(U)\right\|\le\frac{\varkappa_\alpha}{nh}$$

with $\varkappa_\alpha$ from (5.8) as required. □

Let u fulfill $|u-u^*|\le\delta/\rho$ and $U=(P^*_\rho)^2+uu^\top$. Then

$$|u|\le 1+\delta/\rho=\sqrt{1+\alpha}\le(1-\alpha)^{-1/2}$$

and hence, $\|dU/du\|=|u|\le(1-\alpha)^{-1/2}$. Now Lemma 3(iii) ensures for every unit vector $e\in\mathbb{R}^d$ and all $i=1,\dots,n$,

$$\left|\frac{d}{du}e^\top c_i(U)\right|\le\left\|\frac{d}{dU}e^\top c_i(U)\right\|\left\|\frac{dU}{du}\right\|\le\frac{\varkappa_\alpha|u|}{nh}\le\frac{\varkappa'_\alpha}{nh}\tag{5.11}$$

with

$$\varkappa'_\alpha=\frac{\varkappa_\alpha}{\sqrt{1-\alpha}}=\frac{\sqrt{2}C_VC_{K'}}{(1-\alpha)^2}+\frac{2\sqrt{2}C_V^2C_KC_{K'}}{(1-\alpha)^3}.$$

Now we are ready to show (5.7) and (5.6). By Lemma 3(i),

$$\mathbf{E}|\zeta(U^*)|^2=\sigma^2\sum_{i=1}^n|c_i(U^*)|^2\le\sigma^2\frac{2C_V^2C_K^2}{h^2n},$$

which implies (5.7). The assertion (5.6) follows from (5.11) and Lemma 8; see Appendix, applied with $a_i(u)=c_i(U)\sqrt{n}$ for $U=U_u=(P^*_\rho)^2+uu^\top$, $r=\alpha/2$ and $\varkappa=\varkappa'_\alpha/h\sqrt{n}$.

5.1. *Proof of Proposition 1.* This statement can be proved in the same way as Proposition 2. It suffices to follow the proof of Proposition 2 and to replace P^*_ρ formally by the unit operator and b by zero vector. We omit the details.

5.2. *Proof of Theorem 2.* To be able to apply Proposition 2 to the estimators $\hat\beta_k$ at step k, we need that the vector $b=\hat\beta_{k-1}$ coming as the result of the preceding iteration belongs to the set $\mathfrak{B}_{\rho,\delta}=\{b:|P^*_\rho b|\le\delta\}$ with $\rho=\rho_k$ and some $\delta_k<\rho_k/4$. Since the vector $\hat\beta_{k-1}$ is random, we have to ensure that the probability of the event $\{\hat\beta_{k-1}\in\mathfrak{B}_{\rho_k,\delta_k}\}$ is sufficiently large. We now aim to show that this condition is satisfied when n is large enough.

Let the numbers h_k and ρ_k be as in the algorithm description, $k=1,\dots,k(n)$. Define successively the values δ_k and α_k, $k=1,\dots,k(n)$ by $\alpha_1=0$ and

$$\begin{aligned}\delta_k&=\frac{\sqrt{2}C_gC_V}{(1-\alpha_k)^{3/2}}h_k\rho_k^2+\frac{\sqrt{2}\sigma C_VC_Kz_n}{h_k\sqrt{n}}+\frac{\sigma C_{\alpha_k,n}\alpha_k}{h_k\sqrt{n}},\\ \alpha_{k+1}&=\rho_{k+1}^{-2}(2\delta_k\rho_k+\delta_k^2)\end{aligned}\tag{5.12}$$

with $z_n=(1+2\log n+2\log\log n)^{1/2}$.

LEMMA 4. *For n sufficiently large, the α_k's fulfill* $\max_{k\le k(n)}\alpha_k<1/4$. *In addition, for the last iteration $k(n)$, it holds*

$$\mu_n:=\frac{\sqrt{2}C_gC_V}{(1-\alpha_{k(n)})^{3/2}}h_{k(n)}\rho_{k(n)}^2+\frac{\sigma C_{\alpha_{k(n)},n}\alpha_{k(n)}}{h_{k(n)}\sqrt{n}}\le C_1z_n^2n^{-2/3}.$$

PROOF. The rule (2.6) implies $\alpha_{k+1} = 2a^c \delta_k/\rho_k + a^{2c}\delta_k^2/\rho_k^2$. Next,

$$\frac{\delta_k}{\rho_k} = \frac{\sqrt{2} C_g C_V}{(1-\alpha_k)^{3/2}} h_k \rho_k + \frac{\sqrt{2}\sigma C_V C_K z_n + \sigma C_{\alpha_k,n}\alpha_k}{h_k \rho_k \sqrt{n}}.$$

By (2.6), $h_k\rho_k = h_{k-1}\rho_{k-1}a^{1-c} < h_{k-1}\rho_{k-1}$, and hence, $h_k\rho_k \le h_1\rho_1 = C_0 n^{-1/(4\vee d)}$ and $\sqrt{n} h_k \rho_k \ge \sqrt{n} h_{k(n)}\rho_{k(n)} \ge C_0 n^{1/6}$ which ensure the first assertion of the lemma. Next, it is easy to see that $\delta_{k(n)-1} \le C z_n n^{-1/2}$ implies $\alpha_{k(n)} \le C_1 z_n n^{-1/6}$ and, since $C_{\alpha_k,n} \le C z_n$, also $\mu_n \le C_1 z_n^2 n^{-2/3}$. □

Next, successive application of the results of Proposition 2 and Corollary 2 leads to the following.

LEMMA 5. *Let n be sufficiently large. There exist random sets $\mathscr{A}_1 \supseteq \cdots \supseteq \mathscr{A}_{k(n)}$ such that $\mathbf{P}(\mathscr{A}_k) \ge 1 - 3k/n$ and it holds on $\mathscr{A}_k$,*

$$\left|P^*_{\rho_{k+1}}(\hat\beta_k - \beta^*)\right| \le \delta_k, \qquad k = 1, \ldots, k(n) - 1.$$

PROOF. We proceed by induction in k. First we apply Proposition 1 with $z = z_n$ to the first-step estimator $\hat\beta_1$ and use that $z_n e^{-(z_n^2-1)/2} < 1/n$. We then obtain

$$\mathbf{P}(|\hat\beta_1 - \beta^*| \ge \delta_1) \le 1/n;$$

that is, there exists a random set $\mathscr{A}_1$ with $\mathbf{P}(\mathscr{A}_1) \ge 1 - 1/n$ such that $|\hat\beta_1 - \beta^*| \le \delta_1$ on $\mathscr{A}_1$. This obviously implies

$$|P^*_{\rho_2}(\hat\beta_1 - \beta^*)| \le \delta_1.$$

Suppose now that there is $\mathscr{A}_{k-1}$ such that $\mathbf{P}(\mathscr{A}_{k-1}) \ge 1 - 3(k-1)/n$ and it holds on $\mathscr{A}_{k-1}$

$$\left|P^*_{\rho_k}(\hat\beta_{k-1} - \beta^*)\right| \le \delta_{k-1},$$

that is, $\hat\beta_{k-1} \in \mathfrak{B}_{\delta_{k-1},\rho_k}$. By Corollary 1 applied again with $z = z_n$, there exists another random set A_k with $\mathbf{P}(A_k) \ge 1 - 3/n$ such that it holds for every $b \in \mathfrak{B}_{\delta_{k-1},\rho_k}$,

$$|P^*_{\rho_k}(\hat\beta_b - \beta^*)| \le \delta_k$$

so that, with $\mathscr{A}_k = \mathscr{A}_{k-1} \cap A_k$, we obtain $\mathbf{P}(\mathscr{A}_k) \ge 1 - 3k/n$ and it holds on $\mathscr{A}_k$,

$$|P^*_{\rho_k}(\hat\beta_k - \beta^*)| \le \delta_k.$$

and, since for every $\rho' < \rho$,

$$\left\|P^*_{\rho'}(P^*_\rho)^{-1}\right\|^2 = \left\|\left(I + (\rho')^{-2}\beta^*(\beta^*)^\top\right)^{-1}\left(I + \rho^{-2}\beta^*(\beta^*)^\top\right)\right\| \le 1,$$

the assertion follows. □

Let now $\mathscr{A}_{k(n)-1}$ be the random set with $\mathbf{P}(\mathscr{A}_{k(n)-1}) \geq 1-(3k(n)-3)/n$ shown in Lemma 5 so that on this set the estimator $\hat{\beta}_{k(n)-1}$ belongs to $\mathfrak{B}_{\delta,\rho}$ with $\delta = \delta_{k(n)-1}$ and $\rho = \rho_{k(n)}$. Let then η^* be the Gaussian vector from Proposition 2 applied with $h = h_{k(n)}$ and the above δ and ρ. Due to this proposition, there exists a random set $A_{k(n)}$ with $\mathbf{P}(A_{k(n)}) \geq 1-2/n$, so that on $A_{k(n)}$ it holds for all $b \in \mathfrak{B}_{\delta,\rho}$,

$$\left| P_\rho^*(\hat{\beta}_b - \beta^*) - \frac{\eta^*}{h\sqrt{n}} \right| \leq \mu_n,$$

where μ_n is defined in Lemma 4. This yields for the set $\mathscr{A}_{k(n)} = \mathscr{A}_{k(n)-1} \cap A_n$ that $\mathbf{P}(\mathscr{A}_{k(n)}) \geq 1-(3k(n)-1)/n$ and the final estimator $\hat{\beta} = \hat{\beta}_{k(n)}$ satisfies on $\mathscr{A}_{k(n)}$,

$$\left| P_\rho^*(\hat{\beta} - \beta^*) - n^{-1/2}\xi^* \right| \leq \mu_n,$$

where $\xi^* = h^{-1}\eta^*$. In view of $h = h_{k(n)} \geq 1$,

$$\mathbf{E}|\xi^*|^2 = h^{-2}\mathbf{E}|\eta^*|^2 \leq 2\sigma^2 C_V^2 C_K^2$$

and Theorem 2 is completely proved. □

5.3. *Proof of Theorem* 1. We use the result of Theorem 2 and the following technical statement.

LEMMA 6. *Let $\beta \in \mathbb{R}^d$ be such that $|P_\rho^*(\beta - \beta^*)| \leq \delta$ for some $\delta < \rho/4$ and $|\beta^*| \geq 4\delta$. Then it holds for $\theta = \beta/|\beta|$ and $\theta^* = \beta^*/|\beta^*|$,*

$$\left| \theta - \theta^* - \frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta^*|} \right| \leq \frac{2\delta^2(1+|\beta^*|/\rho)}{|\beta^*|^2}.$$

Here Π^* denotes the projector in $\mathbb{R}^d$ on the vector β^*, that is, $\Pi^* = \theta^*(\theta^*)^\top$.

PROOF. Let (β, β^*) denote the angle between two vectors β and β^*. Then

$$\sin(\beta,\beta^*) = \frac{|(I-\Pi^*)\beta|}{|\beta|}, \quad \cos(\beta,\beta^*) = \frac{|\Pi^*\beta|}{|\beta|}, \quad \tan(\beta,\beta^*) = \frac{|(I-\Pi^*)\beta|}{|\Pi^*\beta|}.$$

The simple geometry gives

$$|\theta - \theta^*| \leq \tan(\beta,\beta^*),$$

$$|\Pi^*(\theta-\theta^*)| = |\theta-\theta^*|\sin\frac{(\beta,\beta^*)}{2} \leq |\theta-\theta^*|\frac{\tan(\beta,\beta^*)}{2} \leq \frac{\tan^2(\beta,\beta^*)}{2}$$

so that

$$|\theta - \theta^* - (I-\Pi^*)\theta| = |\Pi^*(\theta-\theta^*)| \leq \frac{\tan^2(\beta,\beta^*)}{2} = \frac{|(I-\Pi^*)\beta|^2}{2|\Pi^*\beta|^2}. \tag{5.13}$$

Note that $I-\Pi^*$ is the projector onto the hyperplane $\mathscr{L}=\{v:\Pi^*v=0\}$. By definition, the operator $P_\rho^*=\big(I+\rho^{-2}\beta^*(\beta^*)^\top\big)^{-1/2}$ coincides with the unity operator within the hyperplane $\mathscr{L}$; that is, $P_\rho^*(I-\Pi^*)=(I-\Pi^*)P_\rho^*=I-\Pi^*$. Hence,

$$(I-\Pi^*)\beta=(I-\Pi^*)(\beta-\beta^*)=(I-\Pi^*)P_\rho^*(\beta-\beta^*),$$

$$(I-\Pi^*)\theta=\frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta|}.$$

This, (5.13) and the inequality $|P_\rho^*(\beta-\beta^*)|\le\delta$ imply

$$\left|\theta-\theta^*-\frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta|}\right|\le\frac{\delta^2}{2|\Pi^*\beta|^2}. \tag{5.14}$$

It is obvious that

$$\|(P_\rho^*)^{-1}\|=\big\|\big(I+\rho^{-2}\beta^*(\beta^*)^\top\big)^{1/2}\big\|=\sqrt{1+|\beta^*|^2/\rho^2}\le1+|\beta^*|/\rho,$$

so that

$$|\beta-\beta^*|=\big|(P_\rho^*)^{-1}P_\rho^*(\beta-\beta^*)\big|\le\big\|(P_\rho^*)^{-1}\big\|\,|P_\rho^*(\beta-\beta^*)|\le(1+|\beta^*|/\rho)\delta$$

and hence,

$$|\beta^*|(1-\delta/\rho)-\delta\le|\beta|\le|\beta^*|(1+\delta/\rho)+\delta. \tag{5.15}$$

Since $\Pi^*\beta^*=\beta^*$, it holds

$$\big|\Pi^*\beta-\beta^*\big|=\big|\Pi^*(\beta-\beta^*)\big|\le\big|\beta-\beta^*\big|\le\big(1+\big|\beta^*\big|/\rho\big)\delta,$$

which along with the conditions $\delta/\rho\le1/4$, $|\beta^*|\ge4\delta$ provides

$$|\Pi^*\beta|\ge|\beta^*(1-\delta/\rho)-\delta|\ge|\beta^*|/2.$$

This and (5.14) yield

$$\left|\theta-\theta^*-\frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta|}\right|\le\frac{2\delta^2}{|\beta^*|^2}.$$

Further, by (5.15),

$$\begin{aligned}\left|\theta-\theta^*-\frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta^*|}\right| &\le \left|\theta-\theta^*-\frac{(I-\Pi^*)P_\rho^*(\beta-\beta^*)}{|\beta|}\right|+\delta\left|\frac{1}{|\beta|}-\frac{1}{|\beta^*|}\right|\\ &\le\frac{2\delta^2}{|\beta^*|^2}+\delta\frac{(1+|\beta^*|/\rho)\delta}{|\beta^*|\{|\beta^*|(1-\delta/\rho)-\delta\}}\le\frac{2\delta^2(1+|\beta^*|/\rho)}{|\beta^*|^2}\end{aligned}$$

as required. □

Due to Theorem 2 and Corollary 1 applied with $z=z_n=(1+2\log n+2\log\log n)^{1/2}$, there exists a random set $\mathscr{A}_{k(n)}$ with $\mathbf{P}(\mathscr{A}_{k(n)})\ge1-3k(n)/n$ such that it holds on this set,

$$|P_\rho^*(\hat\beta-\beta^*)|\le\delta_{k(n)},\qquad\left|P_\rho^*(\hat\beta-\beta^*)-\frac{\xi^*}{\sqrt{n}}\right|\le\mu_n$$

with $\mu_n \le C_1 z_n^2 n^{-2/3}$, $\delta_{k(n)} \le \sqrt{2}\sigma C_V C_K z_n n^{-1/2} + \mu_n$ (see Lemma 4) and with a Gaussian vector ξ^* satisfying $\mathbf{E}|\xi^*|^2 \le 2\sigma^2 C_V^2 C_K^2$. Now Lemma 6 provides for $\gamma^* = (I - \Pi^*)\xi^*/|\beta^*|$ on $\mathscr{A}_{k(n)}$,

$$\left|(\hat{\theta} - \theta) - \frac{\gamma^*}{\sqrt{n}}\right| \le \frac{\mu_n}{|\beta^*|} + \frac{2\delta_{k(n)}^2(1 + |\beta^*|/\rho)}{|\beta^*|^2}.$$

The use of $\delta_{k(n)} \le C z_n n^{-1/2}$, $\rho \approx n^{-1/3}$ and $|\beta^*| \ge C z_n n^{-1/2}$ completes the proof of Theorem 1.

APPENDIX

Here we present two general assertions about Gaussian random vectors.

LEMMA 7. *Let ξ be a Gaussian vector in $\mathbb{R}^d$. Then for every $z \ge 1$,*

$$\mathbf{P}\left(|\xi| > z\sqrt{\mathbf{E}|\xi|^2}\right) \le z e^{-(z^2-1)/2}.$$

PROOF. For every orthonormal $d \times d$-matrix U, the vector $U\xi$ is also Gaussian and $|\xi| = |U\xi|$. Therefore, selecting a proper transform U, we can reduce the general case to the situation when the components ξ_i of the vector ξ are independent. Denote $v_i^2 = \mathbf{E}\xi_i^2$ and $V^2 = \mathbf{E}|\xi|^2$. Obviously $V^2 = \sum_{i=1}^d v_i^2$ and by the Chebyshev inequality, it holds for every $\mu > 0$ with $2\mu v_i^2 < 1$ for all $i \le d$,

$$\mathbf{P}(|\xi| > zV) = \mathbf{P}(|\xi|^2 > z^2\mathbf{E}|\xi|^2) \le \exp(-\mu z^2 V^2)\mathbf{E}\exp(\mu|\xi|^2).$$

Since the components ξ_i of ξ are independent,

$$\begin{aligned}\mathbf{E}\exp(\mu|\xi|^2) &= \mathbf{E}\exp\left(\sum_{i=1}^d \mu\xi_i^2\right) = \prod_{i=1}^d \mathbf{E}\exp(\mu\xi_i^2)\\ &= \prod_{i=1}^d \frac{1}{\sqrt{1 - 2\mu v_i^2}} = \exp\left(-\frac{1}{2}\sum_{i=1}^d \log(1 - 2\mu v_i^2)\right)\end{aligned}$$

so that

$$\mathbf{P}(|\xi| > zV) \le \exp\left(-\mu z^2 V^2 - \frac{1}{2}\sum_{i=1}^d \log(1 - 2\mu v_i^2)\right).$$

We now apply this inequality with $\mu = (1 - z^{-2})V^{-2}/2$ and use that $-\log(1 - x) - \log(1 - y) \le -\log(1 - x - y)$ for all positive x, y with $x + y < 1$. This yields

$$\begin{aligned}\mathbf{P}(|\xi| > zV) &\le \exp\left(-\mu z^2 V^2 - \frac{1}{2}\log(1 - 2\mu V^2)\right)\\ &= \exp\left(-\frac{z^2 - 1}{2} - \frac{1}{2}\log(z^{-2})\right) = z\exp\left(-\frac{z^2 - 1}{2}\right)\end{aligned}$$

as required. □

LEMMA 8. *Let $r > 0$ and let vector functions $a_i(u)$ obey the conditions*

$$\sup_{|u-u^*|\le r}\left|\frac{d}{du}a_i(u)\right| \le \varkappa, \qquad i=1,\dots,n. \tag{A.1}$$

If ε_i are independent $\mathscr{N}(0,\sigma^2)$-distributed random variables, then

$$\mathbf{P}\left(\sup_{|u-u^*|\le r}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\{a_i(u)-a_i(u^*)\}\varepsilon_i\right| > \sigma\varkappa r\left(2+\sqrt{(3+d)\log(4n)}\right)\right) \le \frac{2}{n}.$$

PROOF. Let B_r be the ball $\{u : |u-u^*| \le r\}$ and Σ_r be the ϵ-net on B_r such that for any $u \in B_r$ there is an element u_l of Σ_r such that $|u-u_l| \le r/\sqrt{n}$. It is easy to see that such a net with cardinality $N_r < (4n)^{d/2}$ can be constructed. For a $u_l \in \Sigma_r$ we denote

$$\eta(u_l) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{a_i(u_l)-a_i(u^*)\}\varepsilon_i.$$

Then by (A.1),

$$\mathbf{E}|\eta(u_l)|^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n}\left|a_j(u_l)-a_j(u^*)\right|^2 \le \sigma^2\varkappa^2 r^2,$$

and for any $t \ge 1$ by Lemma 7,

$$\mathbf{P}(|\eta(u_l)| > t\sqrt{\mathbf{E}|\eta(u_l)|^2}) \le te^{-(t^2-1)/2}.$$

Hence, if $t = \sqrt{2\log(N_r n^{3/2})}$, then

$$\begin{aligned}\mathbf{P}\left(\sup_{u_l\in\Sigma_r}|\eta(u_l)| > t\sigma\varkappa r\right) &\le \sum_{l=1}^{N_r}\mathbf{P}(|\eta(u_l)| > t\sigma\varkappa r)\\ &\le tN_r\exp\left(-\log\left(N_r n^{3/2}\right)+1/2\right) < 1/n.\end{aligned} \tag{A.2}$$

Meanwhile, by construction of the net Σ_r, for any $u \in B_r$ there is $u_l(u) \in \Sigma_r$ such that $|u-u_l(u)| \le r/\sqrt{n}$. Then we have by the Cauchy–Schwarz inequality and (A.1),

$$|\eta(u)-\eta(u_l(u))|^2 \le \frac{1}{n}\sum_{i=1}^{n}|a_i(u_l(u))-a_i(u)|^2\sum_{i=1}^{n}\varepsilon_i^2 \le \frac{\varkappa^2 r^2}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

Since the probability $\mathbf{P}((1/n)\sum_{i=1}^{n}\varepsilon_i^2 > 4\sigma^2)$ is certainly less than n^{-1}, it follows that

$$\mathbf{P}\left(\sup_{u\in B_r}|\eta(u)-\eta(u_l(u))| > 2\varkappa\sigma r\right) \le \frac{1}{n}. \tag{A.3}$$

Now (A.2) and (A.3) and the bound $n^{3/2}N_r \le (4n)^{(3+d)/2}$ imply, in an obvious way,

$$\begin{aligned}\mathbf{P}\Bigl(\sup_{u\in B_r}|\eta(u)| > \varkappa\sigma r\Bigl(2+\sqrt{(3+d)\log(4n)}\Bigr)\Bigr)\\ \le \mathbf{P}\Bigl(\sup_{u_l\in\Sigma_r}|\eta(u_l)| > \varkappa\sigma r\sqrt{(3+d)\log(4n)}\Bigr)\\ +\mathbf{P}\Bigl(\sup_{u\in B_r}|\eta(u)-\eta(u_l(u))| > 2\varkappa\sigma r\Bigr) \le \frac{2}{n}\end{aligned}$$

and the lemma follows. □

Acknowledgments. The authors thank two anonymous referees, the editors and J. Horowitz for the remarks and criticisms which helped very much to improve the paper. The assistance of J. Polzehl in carrying out the simulation study is gratefully acknowledged.

REFERENCES

BONNEU, M., DELECROIX, M. and HRISTACHE, M. (1997). Semiparametric estimation of generalized linear models. Unpublished manuscript.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

DELECROIX, M., HÄRDLE, W. and HRISTACHE, M. (1997). Efficient estimation in single-index regression. Discussion paper 37, SFB 373, Humboldt Univ. Berlin.

DELECROIX, M. and HRISTACHE, M. (1999). *M*-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc.* **6** 161–185.

DELECROIX, M., HRISTACHE, M. and PATILEA, V. (1999). Optimal smoothing in semiparametric index approximation of regression functions. Working paper 9952, CREST, Paris.

DUAN, N. and LI, K.-C. (1991). Slicing regression: a link-free regression method. *Ann. Statist.* **19** 505–530.

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

FRIEDMAN, F. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.

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

HÄRDLE, W. and TSYBAKOV, A. B. (1993). How sensitive are average derivatives *J. Econometrics* **58** 31–48.

HOROWITZ, J. L. and HÄRDLE, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91** 1632–1640.

IBRAGIMOV, I. and KHASMINSKI, R. (1987). Estimation of linear functionals in Gaussian noise. *Theory Probab. Appl.* **32** 30–39.

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58** 71–120.

KLEIN, R. L. and SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86** 316–342.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Powell, J. L., Stock, J. M. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430.

Samarov, A. (1991). On asymptotic efficiency of average derivative estimates. *Nonparametric functional estimation and related topics. NATO Adv. Sci. Inst. Ser. C* **335** 167–172.

Samarov, A. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88** 836–847.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481.

M. Hristache
ENSAI and CREST
Campus de Ker Lann
rue B. Pascal
35170 Bruz
France
E-mail: hristach@ensai.fr

A. Juditsky
Université Joseph Fourier
38041 Grenoble Cedex 9
France
E-mail: anatoli.iouditski@inrailps.fr

V. Spokoiny
Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-mail: spokoiny@wias-berlin.de